Unreliable evidence:
Two sources of uncertainty during perceptual choice

Elizabeth Michael*[1], Vincent de Gardelle*[2], Alejo Nevado-Holgado[1], and Christopher Summerfield[1]

[1] Department of Experimental Psychology, University of Oxford, Oxford OX1 3UD, UK.
[2] CNRS UMR 8158, Laboratoire Psychologie de la Perception, 75006 Paris, France.
* denotes equal author contribution

abstract word count: 150
main text word count: 4337
figures: 4

ABSTRACT

Perceptual decisions often involve integrating evidence from multiple concurrently-available sources. Uncertainty arises when the integrated (mean) evidence fails to support one alternative over another. However, evidence heterogeneity (variability) also provokes uncertainty. Here, we asked whether these two sources of uncertainty have independent behavioural and neural effects during choice. Human observers undergoing functional neuroimaging judged the average colour or shape of multi-element array. The mean and variance of the feature values exterted independent influences on behaviour and brain activity. Surprisingly, BOLD signals in the dorsomedial prefrontal cortex (dmPFC) showed polar opposite responses to the two sources of uncertainty, with the strongest response to ambiguous tallies of evidence (high mean uncertainty) and to homogenous arrays (low variance uncertainty). These findings challenge models that emphasise the role of the dmPFC in detecting conflict, errors or surprise. We suggest an alternative explanation, whereby the dmPFC signals distance to a decision boundary during categorical choice.

INTRODUCTION

Perceptual classification judgments involve quantifying sensory evidence and comparing it to a criterion or *category boundary*[1-4] . For example, a particular colour might separate 'ripe' from 'unripe' fruit for a foraging animal. Items with feature values close to this boundary elicit prolonged decision latencies, a delay that computational models attribute to the need resolve uncertainty – or conflict – between closely-matched rival responses[5-8]. In the primate brain, this process has been attributed to a network including the medial prefrontal cortex, anterior insular cortex, dorsomedial thalamus, and lateral parietal cortex[9-15].

In the laboratory, categorisation is often studied by asking observers to classify an isolated visual stimulus with respect to a boundary. However, real-world decisions often require observers to integrate information from multiple sources (e.g. by averaging). For example, a hungry animal might decide where to forage by averaging the size or colour of all the fruit in a tree, or a punter betting on a soccer match might consider the average skill level of all the players in a team. Importantly, judgments about the average information in a multi-element array are made in the face of two potentially orthogonal sources of decision-level uncertainty – those owing to the *mean* and *variability* of the feature values respectively. In a soccer team, mean player ability might closely match that of the rival team, making it hard to predict which will win – but equally, skill levels in a team might also be variable (e.g. one team might have excellent strikers but a weak goalkeeper), making it hard to estimate their potential. Previous work has shown that summary information about a multi-element array can prime behaviour independently of the values of individual items[16-18], and the mean (or strength) and variance (or reliability) of evidence can both increase error rates and delay decisions[19]. These findings suggest unique computational mechanisms for estimating summary statistics such as the mean and variance of perceptual information, but the nature of these mechanisms and their neural implementation both remain unknown.

In the current study, thus, we asked human observers to classify an array of eight elements according to either its average colour or shape, while recording functional magnetic resonance imaging (fMRI) data. Critically, across trials we manipulated independently the mean and variance of the Gaussian distributions from which feature information was drawn, allowing us to assess the orthogonal influences of uncertainty due to the mean (U_M) and variance (U_V) of the evidence on both behaviour and brain activity. Replicating previous findings[10,12,20,21], we observed that U_M was associated with increased BOLD activity in the dorsomedial prefrontal cortex (dmPFC) and anterior insular cortices (AIC), brain regions that are known to vary with decision uncertainty[22,23]. Remarkably however, and contrary to our predictions, U_V had the opposite effect, with dmPFC and AIC showing relatively *decreased* BOLD responses when evidence variability was increased, i.e. when the feature values were more heterogeneous (and performance declined). These findings call into question current models in which the function of the dmPFC is to monitor for uncertainty[5] or surprise[6], predict errors[24], determine time-on-task[25],

or learn the value of actions[26]. Rather, using computational simulations, we show that our findings can be accounted for if these regions are signalling distance to a choice point that determines behaviour, such as the decision criterion or category boundary in classification tasks.

RESULTS

On each trial, observers classified a circular array of eight elements ('squircles') according to their average colour or shape, receiving auditory feedback after each response (Fig. 1a). Each of the eight elements took on a colour value (red to blue) and shape value (square to circle), both parameterised in the range -1 to 1. The eight feature values for colour and shape were drawn pseudorandomly from normal distributions whose mean μ was varied at four symmetric levels around zero, tailored to ensure overall accuracy of ~80%. Array variance σ^2 was varied orthogonally at three levels. Colour and shape were deemed decision-relevant in alternating blocks and the optimal policy was thus to respond circle/square or red/blue according to whether the average feature value on the *relevant* dimension was greater or less than zero (ignoring the irrelevant dimension). We thus conceived of our experiment as a 2 (task-relevant $|\mu|$) x 3 (task-relevant σ^2) x 2 (task-irrelevant $|\mu|$) x 3 (task-irrelevant σ^2) factorial design.

Behavioural data. Pre-experimental calibration ensured comparable performance in shape and colour tasks for the fMRI experiment (see methods and figure S1). ANOVAs revealed main effects of both mean (i.e. proximity to category boundary) [$F_{(1,20)}=92.7$, $p<0.001$] and variance (i.e. element heterogeneity) [$F_{(2,40)}=18.0$, $p<0.001$] on correct response latencies (fig. 1b, top left panel), but no significant interaction [$F_{(2,40)}=0.46$, $p<0.58$]. Comparable effects were found for accuracy (fig. 1b, bottom left panel), with more errors occurring on trials with low mean than high mean [$F_{(1,20)}=104$, $p<0.001$], or trials with high variance vs. low variance [$F_{(2,40)}=7.80$, $p<0.001$], and no interaction [$F_{(2,40)}=2.70$, $p<0.15$]. The irrelevant mean and variance had no significant effect on RT or errors (all p-values > 0.05; fig. 1b, right panels). We also monitored eye movements to ensure that they did not differ between conditions (fig. S4). Together, these data confirm that array mean and variance have independent effects on behaviour.

Decision weighting. When decisions involve integration of multiple sources of evidence, the question arises of how much each source contributes to the final choice. In a previous report[19], demonstrated that elements that were outlying in value space (e.g. extremes of red or blue) elements carried less weight in the choice than inlying elements ('robust averaging'). These findings were replicated here (fig. 1d, left), with significantly higher weights for inlying vs. outlying items [$t_{(20)}=2.72$, $p < 0.013$]. This effect, and the prolongation of decision latencies on high-variance trials, can both be explained if decision values undergo a sigmoidal 'squashing' transformation, for example with each element's adjusted decision value indexing the log of the relative likelihood of each response (log probability ratio or LPR), conditioned on the occurrence of that element (see online methods). Here we used an identical modelling

approach, finding that once again we could predict the pattern of RTs and errors (fig. 1c, lines) and outlier downweighting (fig. 1d, right panel) by transforming the raw or ('native space') colour or shape values (μ and σ) according to the extent that they predicted each response to give new quantities corresponding to the mean and standard deviation of LPR values (we call these 'decision space' values) for both the relevant (U_{Mr} and U_{Vr}) and irrelevant (U_{Mi} and U_{Vi}) dimensions. We used these values, which offered a good fit to behaviour, to predict brain activity in the fMRI scanner at the single trial level. Model fitting is described in more detail in the online methods.

fMRI experiment. Our factorial design allowed us to identify voxels that responded with enhanced BOLD signal to increasing uncertainty associated with the mean (i.e. when average evidence fell near the category boundary) and variance (i.e. when evidence was more heterogeneous). Importantly, this could be done independently both for task-relevant information and task-irrelevant information, allowing us to control for low-level perceptual aspects of the 8-element array. We thus constructed a design matrix by parametrically modulating the height of the predicted BOLD response on each trial by the mean and standard deviation of the decision-space values for both the relevant (U_{Mr} and U_{Vr}) and irrelevant (U_{Mi} and U_{Vi}) dimensions and included nuisance variables to account for the effect of positive and negative feedback, movement parameters and other parameters of no interest (see online methods).

Analyses in decision space. Guided by our modelling analyses, we began by analysing the LPR-transformed decision values, which provided a good fit to behaviour. We first identified a region or regions of interest (ROI) that responded to uncertainty due to the mean (U_{Mr}) of the array at a threshold of $p < 0.0001$ uncorrected (except where noted). Subsequently, we tested its correlation with uncertainty due to the variance (U_{Vr}), and vice versa. These tests are independent, because mean and variance were manipulated orthogonally in our design. ROIs were defined as a 5mm radius sphere centered on the peak activated voxel in each cluster.

Visual cortices. Regions of the visual cortex tended to prefer uncertainty due to the variance rather than the mean. A dorsal visual region focussed on the middle occipital gyrus (sVis; visual area 3) was activated by U_{Vr} [left peak: -30, -84, 14 $t_{(20)}$ =8.15 $p<0.000001$; right peak: 34, -76, 14, $t_{(20)}$ = 7.37, $p<0.00001$] but not U_{Mr} [$t_{(20)}$= 0.99102, $p<0.17$]. Alone among the regions reported here, this region also exhibited a sensitivity to the irrelevant variance U_{Vi} [$t_{(20)}$=2.59, $p<0.009$] (fig. 2d). A more ventral visual region overlapping with the inferior occipital gyrus (iVis) was activated by both U_{Vr} [$t_{(20)}$=4.02, $p<0.0004$] and, more weakly, by U_{Mr} [$t_{(20)}$=1.80, $p<0.04$] (fig. 2e).

Dorsolateral prefrontal and parietal cortices. Distinct regions of the parietal and prefrontal cortices responded preferentially to uncertainty owing to the mean (red rendering) and variance (yellow rendering) respectively (fig. 2) The inferior parietal cortex (IPL) responded strongly to U_{Mr} [left peak: -34, -48, 50; $t_{(20)}$=6.22, $p<0.00001$; right peak: 38, -44, 54; $t_{(20)}$=4.40 $p <0.001$], but not

to U_{Vr} [all p-values > 0.09] (fig. 2a). Similarly, the dorsolateral prefrontal cortex [left peak: -46, 4, 26 $t_{(20)}$=5.77, p<0.0001; right peak: 50, 8, 26; $t_{(20)}$=4.90, p<0.0001] was activated by U_{Mr}, but not U_{Vr} [all p-values > 0.05] (fig. 2b). By contrast, the superior parietal cortex identified by its positive correlation with U_{Vr} [left peak: -18, -60, 54, $t_{(20)}$=5.56, p<0.0001; right peak: 26, -56, 54, $t_{(20)}$ = 5.90, p<0.00001], showed a tendency to respond to U_{Mr} as well, at least in the right hemisphere [left peak, $t_{(20)}$=2.334, p=0.06; right peak, $t_{(20)}$ = 2.732, p=0.006] (fig. 2c). No frontal or parietal regions showed any responsivity to the irrelevant mean (U_{Mi}) or variance (U_{Vi}). We additionally observed voxels that correlated negatively with uncertainty due to the mean (e.g. with the probability of a correct response) in the ventromedial prefrontal cortex and posterior cingulate (fig. S2).

Dorsomedial prefrontal and anterior insular cortices. In the dorsomedial prefrontal cortex (dmPFC) and anterior insula cortex (AIC), we observed a quite different pattern of data. In fig. 3a, we plot the voxels activated by uncertainty due to the mean in red, and those that respond *negatively* to uncertainty due to the variance in green. Consistent with previous studies, we observed positive-going activations for U_{Mr} in the dmPFC [peak: 0, 22, 48; $t_{(20)}$=5.22, p<0.0001] (fig. 3a) and anterior insula [left peak -30, 20, 2; $t_{(20)}$ = 6.07, p<0.00001; right peak: 34, 24, 2; $t_{(20)}$ = 6.14, p<0.00001] (fig. 3b). However, extracting a sphere centered on the peak of each cluster, we also observed a *negative-going* correlation with U_{Vr}, both for the dmPFC ($t_{(20)}$=5.16, p<0.00003) but also reliable in the right ($t_{(20)}$=2.44, p<0.012) and left ($t_{(20)}$=2.88, p<0.005) anterior insular cortices. Similarly, when voxels identified as negatively responsive to U_{Vr} were identified in whole-brain analyses, they were found to respond positively to U_{Mr} in ROI analyses [dmPFC: $t_{(20)}$ = 4.60, p<0.0001, combined AIC: $t_{(20)}$ = 5.58, p<0.00001] (fig. 3a and 3b, green rendering). No correlations with task-irrelevant mean or variance were found. This pattern of data persisted when RT was included as an additional nuisance regressor (fig. S5).

Analyses in native space. Could the negative correlation between dmPFC and AIC BOLD signal and U_{Vr} be due to some artefact of our log-probability transform of decision values? To rule out this possibility, we conducted the same analyses as described above but using the raw (native space) feature values rather than their LPR-transformed counterparts. Globally, the results were qualitatively similar, but statistically more modest. In figure 3c, we show the overlapping clusters responding positively to mean-related uncertainty (negative correlation with $|\mu|$) and negatively to variance-related uncertainty (positive correlation with σ) in the medial PFC for decision-space and native-space analyses. Similar results were obtained for the AIC and visual, parietal and prefrontal regions (fig. S3).

Haemodynamic response functions. To additionally ensure that our unexpected findings were not due to misfitting of our basis function (canonical haemodynamic response) to the data, we reanalysed our whole-brain data using a finite impulse response (FIR) filter (which makes no assumptions about the shape of the BOLD response) and plotted the HRFs for low and high mean

|μ|(fig. 3d, left) , and low, medium and high standard deviation σ (fig. 3d, right) separately. An ANOVA on the peak BOLD response averaged across 4s and 6s post-stimulus onset confirmed the pattern of previous analyses, with larger responses to smaller values of |μ| ($t_{(20)}$ = 2.414, $p<0.0127$) and larger responses on trials with low values of σ ($t_{(20)}$=2.390, $p<0.014$).

Adjusting the decision criterion: a computational model. What variable might be encoded by a brain region that responds positively when the evidence approaches the category boundary, but also responds when the evidence is relatively homogeneous? One quantity that varies in this fashion is the total *distance to category boundary* of the information in the array. Consider the sum of the total absolute distance between each element and the categorical boundary k=0 for a range of values of μ {-0.2…0.2} and σ {0…0.3}. As can be seen in fig. 4b, this absolute distance is shortest when both μ and σ are low, i.e. when uncertainty due to the mean is high but uncertainty due to the variance is low. Thus, one explanation for our findings is that medial PFC neurons compute the proximity of the decision-relevant evidence to the choice point that governs behaviour. The aggregate signal recorded as the BOLD response thus reflects the sum of this activity across all elements.

Intuitively, trials where the elements fall close to the category boundary offer the most unambiguous information about where the decision criterion is likely to lie. Relatively little can be learned about the choice point when the mean evidence lies far from it; similarly, when evidence is heterogeneous, the choice point is hard to identify. Thus, whilst computing distance to category boundary is likely to be relevant for signalling change points in behaviour, it is also important for learning about the precise location of the choice point itself (see discussion). Refining our initial model of how observers perform the task, we considered an ideal observer who learns the posterior probability of the decision criterion falling at each possible location in the task-relevant feature space. The observer updates his or her estimate of the bound (corrupted with some noise N_{bound}, a 'leak' parameter that reflects imperfect memory from the previous trial) in a manner that depends on the cumulative distribution of the elements most recently viewed, according to Bayes' rule (figure 4c and online methods).
Each element E with relevant feature value e (in decision space) thus contributes to the decision according to its choice probability CP_E

$$CP_E = \text{normcdf}(e+N_{decision}, mu_{bound}, sig_{bound})$$

where n is zero-mean Gaussian 'decision' noise and mu_{bound} and sig_{bound} are the mean and standard deviation of the posterior of the decision boundary estimated by the model. Decisions thus reflect the log likelihood ratio calculated with respect to the estimated boundary

$$LLR = \Sigma\log(CP_E./1\text{-}CP_E)$$

with different choices being made according to whether LLR > 0 or LLR < 0.

Estimates of the LLR on trials with differing levels of array mean and variance $|\mu|$ and σ are shown in figure 4d. As expected, the LLR varies in a fashion that is predicted by the accuracy and reaction time of the participants, but not by the dmPFC activity, with larger LLR for high mean and low variance trials. However, when we calculated the total absolute distance to bound across elements

$$DB = \Sigma|\log(p_e)|$$

We observed a pattern that mirrored that of the dmPFC BOLD signal (fig. 4e). Of note, the Kullback-Leibler (KL) divergence between prior and posterior estimates of the bound, a measure of the amount that his learned from the feedback on each trial, also varied with the dmPFC signal (fig. 4f).

Predicting behavioural adjustments. The model described above was devised post-hoc, to account for a pattern of data in the dmPFC that we had not anticipated. Nevertheless, it makes novel predictions about how participants should adjust their behaviour according to the mean and variance of the array on the previous trial, and whether they were correct or not. Firstly, the model predicts that overall mean accuracy should improve on trials on which the array had high mean or high variance, although more sharply on the former. This prediction occurs because although the estimate of the bound is more precise on low mean/low variance trials (fig. 4g), it is also more prone to deviate from the true boundary located at k = 0 (fig. 4h).

Captialising on an independent dataset collected in a separate experiment, we tested this prediction by sorting human accuracy rates according to the condition on the current and previous trial. Human participants were indeed more likely to make a error response after a correct low mean than correct high mean trial ($F_{(1,76)}= 7.34$, $p < 0.008$) and more likely to err after a correct low variance trial than a correct high variance trial ($F_{(2,152)}= 4.55$, $p < 0.02$), with no interaction between the two factors. Figure 5 shows model fits (lines) to performance (dots) binned by current trial mean and variance (fig. 5a) previous trial mean and variance (fig.5b), current trial mean x previous trial mean (fig. 5c) and current trial mean x curren trial variance (fig. 5d). Although the model has only two free parameters (the level of noise N_{bound} and $N_{decision}$), model fits very closely approximate human performance. Under the same parameterisation (see methods), the model also predicts qualitatively (but not quantitatively) two further features of human behaviour: the tendency to repeat the same response (‘stay’) after a correct trial, and the modulation of this effect by the previous trial mean. In other words, our model is able to capture a number of behavioural adjustments during performance of the multi-element classification task.

DISCUSSION

Most laboratory-based categorisation tasks require observers to classify a single, isolated element in to one of two categories. Where evidence is ambiguous, observers equivocate; formal models capture these prolonged

decision latencies with mutual inhibition between competing response nodes[5,8], or with overt mechanisms that put the brakes on responding in order to optimise performance[27,28] . Functional neuroimaging studies have attributed this function to a characteristic network of interconnected brain regions (including the dorsomedial prefrontal and anterior insular cortices) which are known to become active when competing options exhibit similar response values, and that predict the slow-down[25] and increased error probability[24] characteristic of these trials. However, these studies have tended to manipulate the degree of conflict between options, but not formally dissociated the twin influences of mean-related and variance-related uncertainty[7,10,12,20,21]. Other researchers have controlled the sensitivity of sensory discrimination judgments by adjusting the signal-to-noise ratio in a stimulus (e.g. the ratio of coherently to randomly moving dots in a random dot kinetogram[19]), which varies uncertainty but precludes an assessment of the independent influences evidence mean and evidence variability on choice.

Here, manipulating the mean and variability of evidence independently, we report a new finding that is hard to reconcile with current models emphasising the role of the dmPFC processing conflict[9], error likelihood[24], surprise[6], or time on task[25]. Irrespective of its mean value, when evidence is more heterogenous (variable), reaction times and errors increase, but the neural response in the dmPFC and AIC is the polar opposite: *less* activity for *more* variable evidence. This finding was observed both for raw (native-space) values of mean and standard deviation of the evidence, as well as following a log-probability scaling of these values, a transformation that allowed us to better account for the behavioural data. This finding places an important new constraint on models that have sought to link neural observations in these brain regions to computational-level descriptions of decisions made under uncertainty.

One alternative theory of the function of the dorsomedial PFC that has come to prominence recently is that it is responsible for learning the value of actions[26,29,30]. Under this framework, increased dmPFC responses to conflict can be attributed to the simultaneous activation of many rival action-outcome parings in that region. Our explanation for the pattern of dmPFC activity observed in the current task is closely related to this view. We argue that the dmPFC computes the total proximity of current evidence to the decision criterion, a quantity that in our experiment is greatest when the mean is low and when the variance is low. Interestingly, recent work has implicated the dmPFC and underlying anterior cingulate cortex (ACC) in signalling the proximity to a switch in behaviour, for example prior to a switch to a new patch from which to forage[31-33]. One emerging view thus is that the dmPFC/ACC signals the distance between the current decision-relevant evidence and a choice point that determines whether a current, or alternative, course of action should be taken[31,34].

Lesions of the dmPFC/ACC suggest that it also plays a key role in learning the weight that information should carry in a decision[34], a finding backed up by functional imaging work[35,36]. Our model makes a related proposal: that the distance to category boundary offers information about where the category

boundary should lie, even in a task such as ours where there was no explicit drift in the objective (i.e. experimenter-imposed) criterion over time, and that the dmPFC contributes to this learning. Many years of research in experimental psychology have demonstrated that even in a stationary decision task, minor adjustments to behaviour are constantly being made on the basis of the stimuli and feedback[37,38], and imaging studies have identified neural correlates of these adjustments in the dmPFC[39] as well as in subcortical regions[40]. This is likely to reflect the fact that in the wild, optimal choice boundaries are constantly varying: for example, the criterion at which a fruit is deemed worthy of picking will depend on the overall richness of the patch from which it is being harvested. Here, behavioural adjustments were observed in observers performing the multi-element decision task, with performance varying in a complex manner as a function of mean and variance of the array elements on the previous trial, and whether the response made was correct or incorrect. Building a simple computational model in which an uncertain decision criterion is updated under an ideal observer framework, we show that these adjustments can be seen as a natural consequence of adjusting the decision boundary that separates the two options. By introducing two sources of noise – one that reflects imperfect memory for where the bound lies, and that other imperfect integration of the evidence – we were able to reproduce the pattern of behavioural adjustments shown by observers performing the multi-element classification task. Thus, in addition to any role in signalling the distance to category boundary, the dmPFC may contribute to learning about the bound. However, in the current dataset we are unable to disambiguate those two hypotheses, because our design did not separate the decision and learning phases of the trial sufficiently to model them with separate predictors of the BOLD response. One avenue for future research, thus, is to use a neuroimaging technique with finer temporal resolution, such as electroencephalography (EEG), to disambiguate the response to mean and variance that follows the onset of the stimulus from that evoked by the feedback.

However, not all brain regions responded with the pattern of activation observed in the dmPFC and AIC. In the the parietal cortex, a variety of responses to evidence variability were observed, with more inferior regions showing a relative insensitivity to the variance, and more superior regions responding positively uncertainty due to both mean and variance, in a fashion that inversely tracked the log-likelihood ratio associated with the choice. Importantly, and unlike in the visual cortex, these correlations occurred exclusively for the decision-relevant dimension, precluding interpretations based on low-level perceptual aspects of the display. The finding that polar opposite responses to evidence mean and variability were not ubiquitous across the brain rules out spurious interpretations of our findings in the dmPFC, but also hints that the brain may treat evidence uncertainty in a variety of different ways, contrary to the claim that mean signal strength is the sole determinant of decision confidence[13]. For example, the first two statistical moments of the perceptual evidence in a natural scene provide estimates of the range and central tendency of the information, which might allow the observer to adjust the gain of neuronal responding to deal with currently

available information[41]. Models incorporating range adjustment have been applied to account for categorical judgments, and may provide an explanation for decoy effects, whereby the value of an outlying but irrelevant option corrupts choices between two items[42]. Evidence variability may also play a role in computing the gist of perceptual information, which can in turn guide saccadic exploration strategies[43] and facilitate rapid decision-making, for example by permitting divisive normalisation[44]. In our experiment, the fact that evidence variability seemed to be associated with changes in neural activity relatively early in the processing stream (e.g. in visual regions) points to an early role in the choice process, albeit confined to task-relevant information. Investigations of evidence variability and the extraction of gist-like information in complex visual arrays may prove a fruitful area of future research.

Finally, our data speak to a growing debate about how uncertainty is represented in the nervous system. Elegant mathematical models have shown how uncertainty encoded in the dispersion of firing rates in a neuronal population[45] or across time[46] can be harnessed to weight information optimally by its reliability. However, while compelling, these schemes have garnered relatively modest empirical support thus far[47]. An alternative account, based mainly on data from economic tasks, has emphasised the separability of neuronal classes[48] or brain regions[49] that encode outcome mean (value) and variance (risk). Our data suggest that at the very least, the brain treats the uncertainty due to the mean and variance in a dissociable manner. However, translating the proposed neural coding schemes into predictions about the BOLD signal is very challenging[50], and so we leave it to future electrophysiological studies to assess how perceptual uncertainty due to mean and variance is encoded at the single-cell level.

REFERENCES

1. Ashby, F.G. & Maddox, W.T. Human category learning. *Annu Rev Psychol* **56**, 149-178 (2005).
2. Freedman, D.J. & Miller, E.K. Neural mechanisms of visual categorization: insights from neurophysiology. *Neurosci Biobehav Rev* **32**, 311-329 (2008).
3. Green, D.M. & Swets, J.A. *Signal Detection Theory and Psychophysics*, (Wiley, New York, 1966).
4. Wald, A. & Wolfowitz, J. Bayes Solutions of Sequential Decision Problems. *Proc Natl Acad Sci U S A* **35**, 99-102 (1949).
5. Botvinick, M.M., Braver, T.S., Barch, D.M., Carter, C.S. & Cohen, J.D. Conflict monitoring and cognitive control. *Psychol Rev* **108**, 624-652 (2001).
6. Alexander, W.H. & Brown, J.W. Medial prefrontal cortex as an action-outcome predictor. *Nat Neurosci* **14**, 1338-1344 (2011).
7. Bach, D.R., Hulme, O., Penny, W.D. & Dolan, R.J. The known unknowns: neural representation of second-order uncertainty, and ambiguity. *J Neurosci* **31**, 4811-4820 (2011).
8. Bogacz, R., Brown, E., Moehlis, J., Holmes, P. & Cohen, J.D. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol Rev* **113**, 700-765 (2006).
9. Botvinick, M., Nystrom, L.E., Fissell, K., Carter, C.S. & Cohen, J.D. Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature* **402**, 179-181 (1999).
10. Grinband, J., Hirsch, J. & Ferrera, V.P. A neural representation of categorization uncertainty in the human brain. *Neuron* **49**, 757-763 (2006).
11. Hazeltine, E., Poldrack, R. & Gabrieli, J.D. Neural activation during response competition. *J Cogn Neurosci* **12 Suppl 2**, 118-129 (2000).
12. Huettel, S.A., Song, A.W. & McCarthy, G. Decisions under uncertainty: probabilistic context influences activation of prefrontal and parietal cortices. *J Neurosci* **25**, 3304-3311 (2005).
13. Kiani, R. & Shadlen, M.N. Representation of confidence associated with a decision by neurons in the parietal cortex. *Science* **324**, 759-764 (2009).
14. Ridderinkhof, K.R., Ullsperger, M., Crone, E.A. & Nieuwenhuis, S. The role of the medial frontal cortex in cognitive control. *Science* **306**, 443-447 (2004).
15. Singer, T., Critchley, H.D. & Preuschoff, K. A common role of insula in feelings, empathy and uncertainty. *Trends Cogn Sci* **13**, 334-340 (2009).
16. Chong, S.C. & Treisman, A. Representation of statistical properties. *Vision Res* **43**, 393-404 (2003).
17. de Fockert, J.W. & Marchant, A.P. Attention modulates set representation by statistical properties. *Percept Psychophys* **70**, 789-794 (2008).
18. Haberman, J. & Whitney, D. The visual system discounts emotional deviants when extracting average expression. *Atten Percept Psychophys* **72**, 1825-1838 (2010).
19. de Gardelle, V. & Summerfield, C. Robust averaging during perceptual judgment. *Proc Natl Acad Sci U S A* **108**, 13341-13346 (2011).

20. Freedman, D.J., Riesenhuber, M., Poggio, T. & Miller, E.K. Visual categorization and the primate prefrontal cortex: neurophysiology and behavior. *J Neurophysiol* **88**, 929-941 (2002).
21. Yang, T. & Shadlen, M.N. Probabilistic reasoning by neurons. *Nature* **447**, 1075-1080 (2007).
22. Banko, E.M., Gal, V., Kortvelyes, J., Kovacs, G. & Vidnyanszky, Z. Dissociating the effect of noise on sensory processing and overall decision difficulty. *J Neurosci* **31**, 2663-2674 (2011).
23. Ratcliff, R., Philiastides, M.G. & Sajda, P. Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG. *Proc Natl Acad Sci U S A* **106**, 6539-6544 (2009).
24. Brown, J.W. & Braver, T.S. Learned predictions of error likelihood in the anterior cingulate cortex. *Science* **307**, 1118-1121 (2005).
25. Grinband, J.*, et al.* The dorsal medial frontal cortex is sensitive to time on task, not response conflict or error likelihood. *Neuroimage* **57**, 303-311 (2011).
26. Rushworth, M.F. & Behrens, T.E. Choice, uncertainty and value in prefrontal and cingulate cortex. *Nat Neurosci* **11**, 389-397 (2008).
27. Cavanagh, J.F.*, et al.* Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nat Neurosci* **14**, 1462-1467 (2011).
28. Ratcliff, R. & Frank, M.J. Reinforcement-based decision making in corticostriatal circuits: mutual constraints by neurocomputational and diffusion models. *Neural Comput* **24**, 1186-1229 (2012).
29. Rushworth, M.F., Noonan, M.P., Boorman, E.D., Walton, M.E. & Behrens, T.E. Frontal cortex and reward-guided learning and decision-making. *Neuron* **70**, 1054-1069 (2011).
30. Rushworth, M.F., Kennerley, S.W. & Walton, M.E. Cognitive neuroscience: resolving conflict in and over the medial frontal cortex. *Curr Biol* **15**, R54-56 (2005).
31. Kolling, N., Behrens, T.E., Mars, R.B. & Rushworth, M.F. Neural mechanisms of foraging. *Science* **336**, 95-98 (2012).
32. Hayden, B.Y., Pearson, J.M. & Platt, M.L. Neuronal basis of sequential foraging decisions in a patchy environment. *Nat Neurosci* **14**, 933-939 (2011).
33. Quilodran, R., Rothe, M. & Procyk, E. Behavioral shifts and action valuation in the anterior cingulate cortex. *Neuron* **57**, 314-325 (2008).
34. Rushworth, M.F., Kolling, N., Sallet, J. & Mars, R.B. Valuation and decision-making in frontal cortex: one or many serial or parallel systems? *Curr Opin Neurobiol* (2012).
35. 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).
36. Summerfield, C., Behrens, T.E. & Koechlin, E. Perceptual classification in a rapidly changing environment. *Neuron* **71**, 725-736 (2011).
37. Gratton, G., Coles, M.G. & Donchin, E. Optimizing the use of information: strategic control of activation of responses. *J Exp Psychol Gen* **121**, 480-506 (1992).

38. Egner, T. Congruency sequence effects and cognitive control. *Cogn Affect Behav Neurosci* **7**, 380-390 (2007).
39. Kerns, J.G.*, et al.* Anterior cingulate conflict monitoring and adjustments in control. *Science* **303**, 1023-1026 (2004).
40. Wunderlich, K., Dayan, P. & Dolan, R.J. Mapping value based planning and extensively trained choice in the human brain. *Nat Neurosci* **15**, 786-791 (2012).
41. Parducci, A. Category judgment: a range-frequency model. *Psychol Rev* **72**, 407-418 (1965).
42. Soltani, A., De Martino, B. & Camerer, C. A range-normalization model of context-dependent choice: a new model and evidence. *PLoS Comput Biol* **8**, e1002607 (2012).
43. Torralba, A., Oliva, A., Castelhano, M.S. & Henderson, J.M. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychol Rev* **113**, 766-786 (2006).
44. Ohshiro, T., Angelaki, D.E. & DeAngelis, G.C. A normalization model of multisensory integration. *Nat Neurosci* **14**, 775-782 (2011).
45. Ma, W.J., Beck, J.M., Latham, P.E. & Pouget, A. Bayesian inference with probabilistic population codes. *Nat Neurosci* **9**, 1432-1438 (2006).
46. Fiser, J., Berkes, P., Orban, G. & Lengyel, M. Statistically optimal perception and learning: from behavior to neural representations. *Trends Cogn Sci* **14**, 119-130 (2010).
47. Bowers, J.S. & Davis, C.J. Bayesian just-so stories in psychology and neuroscience. *Psychol Bull* **138**, 389-414 (2012).
48. Schultz, W., O'Neill, M., Tobler, P.N. & Kobayashi, S. Neuronal signals for reward risk in frontal cortex. *Ann N Y Acad Sci* **1239**, 109-117 (2011).
49. Preuschoff, K., Bossaerts, P. & Quartz, S.R. Neural differentiation of expected reward and risk in human subcortical structures. *Neuron* **51**, 381-390 (2006).
50. O'Reilly, J.X., Jbabdi, S. & Behrens, T.E. How can a Bayesian approach inform neuroscience? *Eur J Neurosci* **35**, 1169-1179 (2012).

FIGURE 1

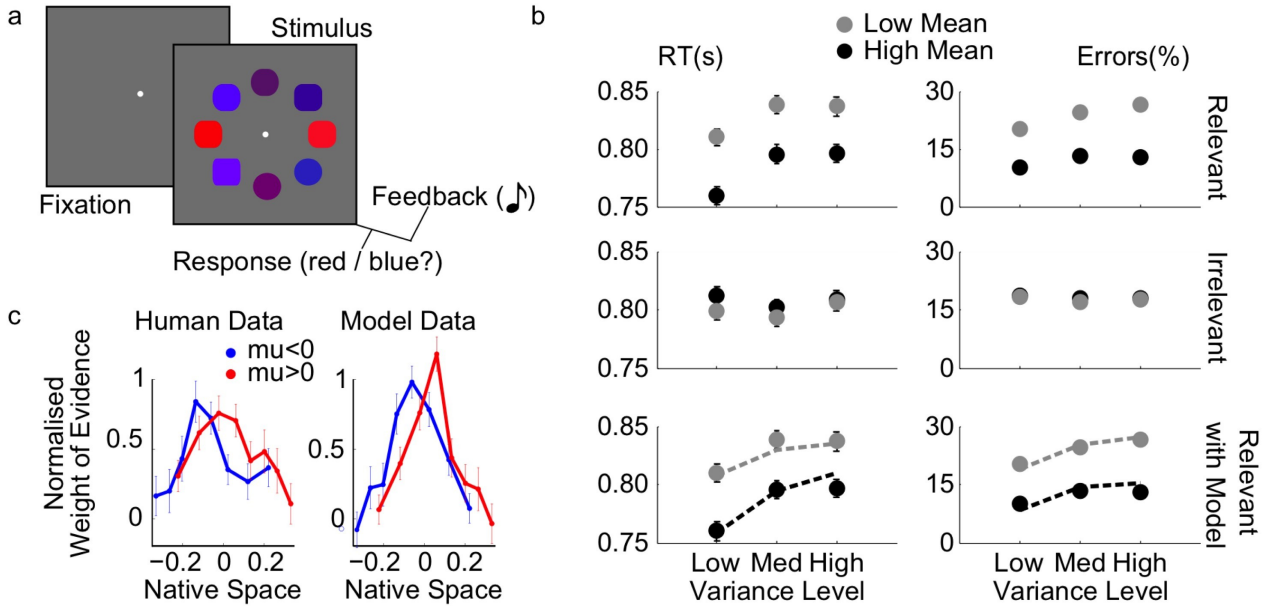

Figure 1. Task, behaviour and modelling. (**a**) Schematic representation of a trial: a white central fixation point, was followed after 500ms by the stimulus array, which participants categorised based on either the average shape (square vs. circle) or average colour (red vs. blue) across all elements, with auditory feedback. (**b**) Response times (RTs, left panels) and error rates (right panels), as a function of the mean and variance manipulations. Low mean and high variance correspond to high uncertainty. Top row: effect of the task-relevant manipulations. Middle row: effect of the irrelevant dimension. Bottom row: best-fitting model data (dashed lines) overlaid on the human data for the task-relevant manipulations. The model was fitted to errors only, and RTs are predictions. (**c**) Coefficients from a logisitic regression in which decision-relevant values, *ranked* in each trial, predicted observers' choices, separately for blue/square (blue) and red/circle (red) stimulus arrays. Higher decision weights appear for inlying vs. outlying elements. The abscissa indicates the average decision value of the elements in each rank, in native space.

FIGURE 2

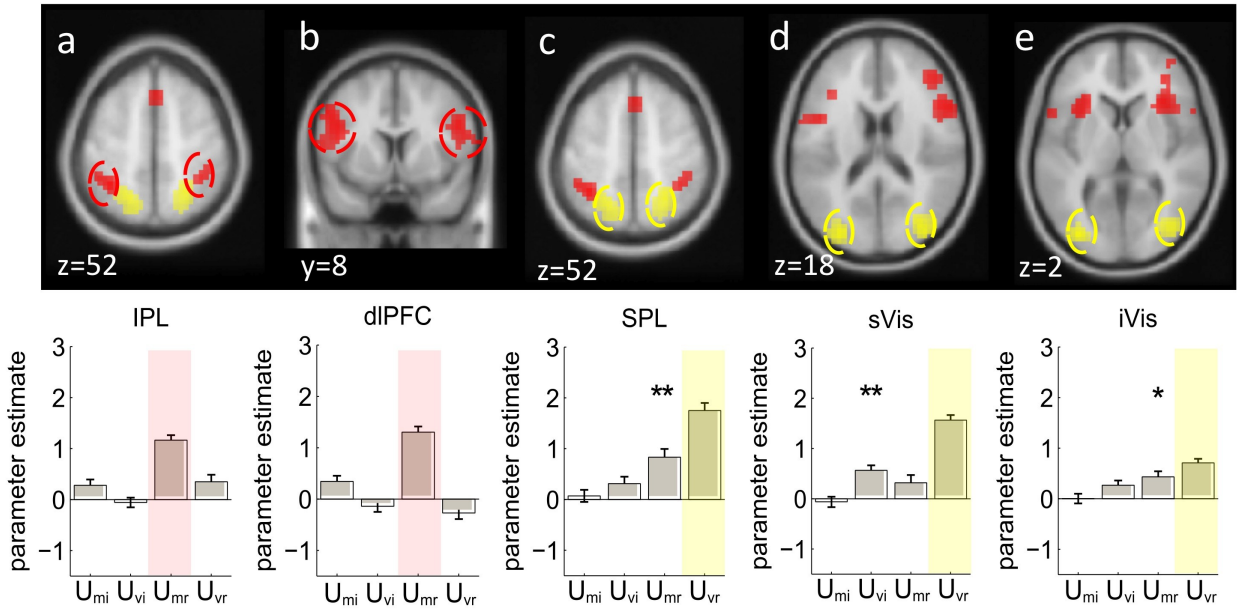

Figure 2. Imaging results from dorsolateral prefrontal, parietal and visual cortices. Top row: voxels where BOLD activity was responding to positively correlated with mean-related uncertainty (U_{Mr}, in red), and positively correlated with variance-related uncertainty (U_{Vr} , in yellow). All activations are rendered on the template brain of the Montreal Neurological Institute with an uncorrected threshold of p<0.001 (see text and tables for full list of activation and peak coordinates). Bottom row: average parameter estimates from a 5 mm sphere centered on the peak activation from the cluster highlighted with a dashed ellipse, for regressors encoding uncertainty due to the mean and variance for relevant and irrelevant dimensions. Stars indicate significance: *p<0.05, **p<0.01, ***p<0.001. Red and yellow shading denotes the condition used to define the ROI. **(a)** Voxels in the inferior parietal cortex (IPL) responding to U_{Mr} (**b**) Voxels in the dorsolateral prefrontal cortex (dlPFC) responding to U_{Mr} (**c**) superior parietal cortex (SPL) showed a positive correlation with U_{vr} (and U_{Mr} – see bar plot). (**d,e**) We subdivided a large region of visual cortex showing activity positively correlated with U_{vr} into superior (sVis) and inferior (iVis) regions.

FIGURE 3

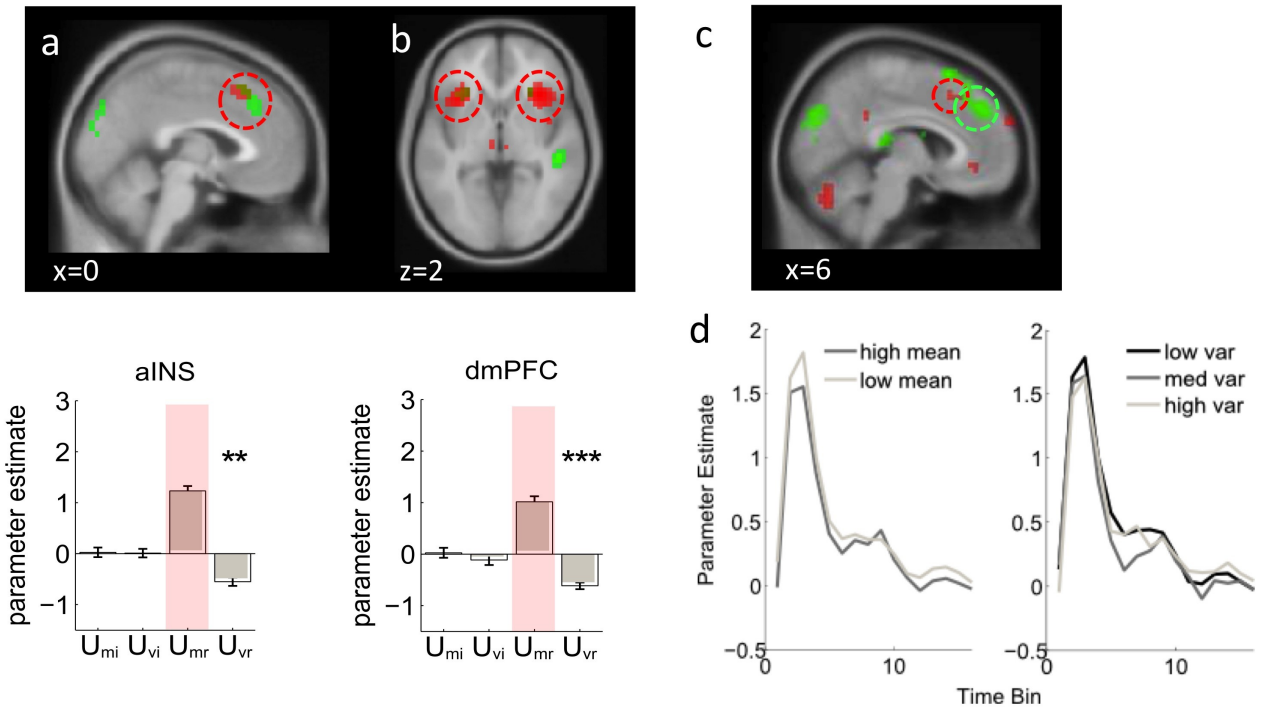

Figure 3. Imaging results from dorsomedial prefrontal cortex and anterior insula. (**a**) Upper panel: voxels responding positively to uncertainty due to the mean (U_{Mr}; red) and negatively to uncertainty due to the variance (U_{Mr}; green) rendered onto a saggital slice of the MNI template brain. The corresponding bar plot shows mean responses extracted a sphere of 5 mm radius around the peak voxel for the highlighted cluster, with stars denoting the statistical significance as in figure 2. **(b)** same results for an axial slice showing the AIC. **(c)** Correlations with native space mean (positive correlation with $|\mu|$, red) and standard deviation (negative correlation with σ, green) in dmPFC, rendered onto a saggital slice at a threshold of $p < 0.005$ uncorrected. The scale indicates the t-value. (**d**) Left panel: haemodynamic response functions (HRFs) generated from a finite impulse response (FIR) model for the dmPFC ROI (5mm sphere extracted from peak of native space activation) for low mean (i.e. $|\mu|$ close to category boundary; light grey) and high mean (dark grey). X-axis shows time in scans (2s). Right panel: HRFs for low (black), medium (dark grey), and high (light grey) variance.

FIGURE 4

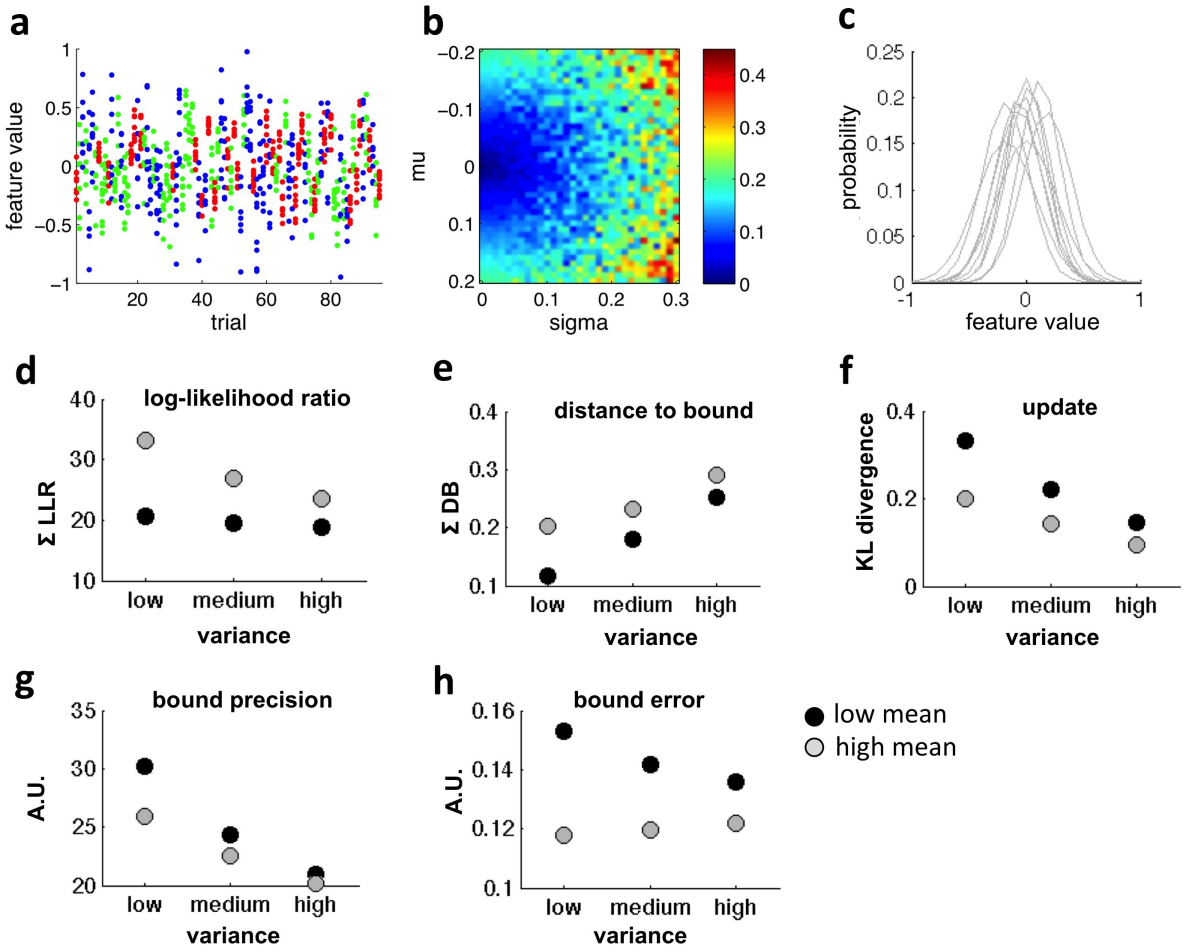

Figure 4. Simulation analysis. (a) Plot of feature values (y-axis) on 100 simulated trials (x-axis). Blue, green and red show high, medium and low variance trials respectively. Distributions were identical to those used in the experiment. **(b)** Total mean distance to category boundary as a function of different values of mu and sigma, for 8 elements. The shortest distance (blue) is observed for trials on which samples were drawn from low variance/low mean distributions. **(c)** Posterior likelihood of the bound position given the stimuli and feedback for 10 trials of the experiment as estimated by the Bayesian learner. **(d)** average sum of log-likelihood ratios across elements for each condition, calculated with respect to the learned bound, for each mean/variance condition **(e)** average sum of distances to bound under the model **(f)** KL divergence beween prior and posterior estimates of the bound following the feedback on each trial type. (g) precision of the bound, i.e. the reciprocal of the square of the standard deviation obtained by fitting a Gaussian to the posterior estimate of the bound. **(g)** the average deviation of the bound from the true objective bound in each condition.

FIGURE 5

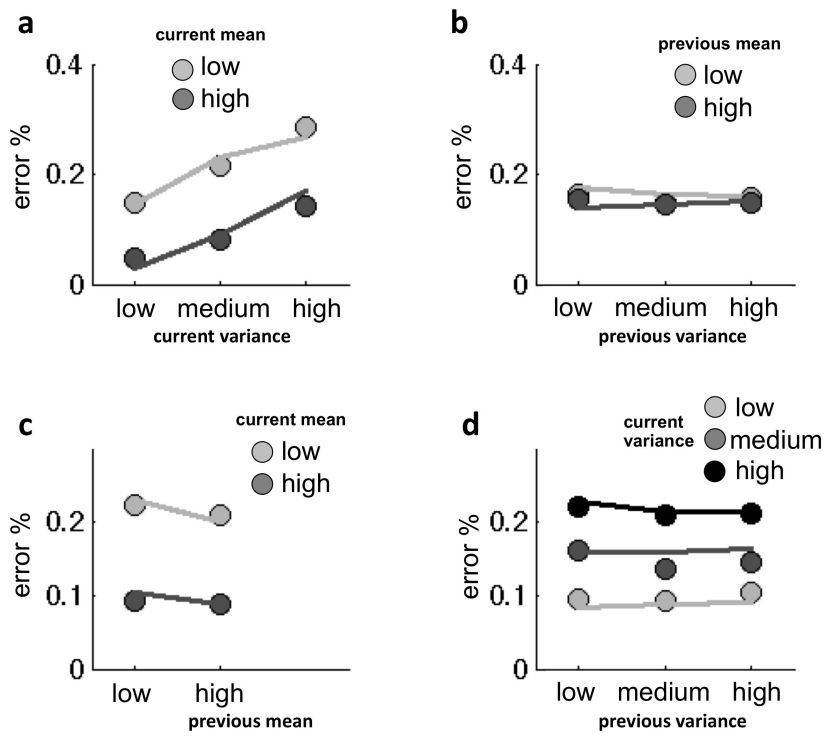

Figure 5. Predicting behavioural adjustments with the bound-learning model. All panels show human (dots) and model predited (lines) error rates as a function of current/previous trial mean and variance. Model performance was estimated with parameters N_{bound} = 0.05 and $N_{decision}$= 0.25 (see online methods). **(a)** Human and model performance as a function of the current trial mean and variance. **(b)** Human and model performance as a function of the previous trial mean and variance. (c) Human and model performance as a function of the current and previous trial mean. **(d)** Human and model performance as a function of the current and previous trial variance.