Encoding of stimulus probability in macaque inferior temporal cortex

Andrew H Bell[1,2,3*], Christopher Summerfield[2*], Elyse L Morin[1], Nicholas J Malecek[1], Leslie G Ungerleider[1]

[1]Laboratory of Brain and Cognition, National Institute of Mental Health, 9000 Rockville Pike, Bethesda, MD 20892, USA
[2]Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford, OX1, 3UD, United Kingdom
[3]MRC Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge, UK CB2 7EF, United Kingdom

* These authors contributed equally to this study.

Corresponding Author: Andrew H Bell, andrew.bell@psy.ox.ac.uk

Number of pages:	26
Number of figures:	6
Number of supplemental figures	6

Summary

Optimal perceptual decisions require sensory signals to be combined with prior information about stimulus probability. Although several theories propose that probabilistic information about stimulus occurrence is encoded in sensory cortex, evidence from neuronal recordings has not yet fully supported this view. We recorded activity from single neurons in inferior temporal cortex (IT) whilst monkeys performed a task that involved discriminating degraded images of faces and fruit. The relative probability of the cue being a face vs. a fruit was manipulated by a latent variable that was not revealed to the monkeys, and which changed unpredictably over the course of each recording session. In addition to responding to stimulus identity (face or fruit), population responses in IT encoded the long-term stimulus probability of whether a face or a fruit stimulus was more likely to occur. Face-responsive neurons showed reduced firing rates to expected faces, an effect consistent with "expectation suppression", but expected stimuli were decoded from multivariate population signals with greater accuracy. These findings support 'predictive coding' theories, whereby neural signals in the mammalian visual system actively encode and update predictions about the local sensory environment.

Keywords: predictive coding; decision-making; neuron; monkey; expectation suppression

Introduction

A long tradition in psychology and neuroscience has cast perception as an inference problem, by which the causes of stimulation are constructed from ambiguous sensory inputs [1-3]. Bayesian models argue that the local and long-term stimulation histories provide prior information about the most likely interpretation of a visual image, which is combined with sensory inputs to guide optimal perceptual decisions [4-6]. According to one influential framework, sensory neurons signal both prior information (i.e., predictions about forthcoming input) and the update signals that allow sensory expectations to be revised in the face of new information (i.e., prediction errors about observed input) [7-10]. This framework has been investigated by human functional neuroimaging studies, which have confirmed that stimuli that are conditionally likely result in globally dampened macroscopic brain responses (i.e., reduced prediction errors), even when low-level sensory adaptation is controlled for [11-20]. For example, BOLD signals in both striate and extrastriate cortex are suppressed when a visual stimulus is predicted by a valid probabilistic cue [11, 14, 17, 21], and BOLD adaptation is attenuated when stimulus repetitions are more rare across a block of trials, even when repetition probabilities are not explicitly signalled [16, 20, 22-27]. These phenomena, whereby expected stimuli elicit reduced brain activity, have been collectively termed "expectation suppression" [28, 29].

Strikingly, however, single-neuron recordings in monkeys have failed to replicate these findings. Kaliukhovich and Vogels trained monkeys to expect certain pairings of stimuli more than others [30]. Rare vs. frequent repetitions elicited comparable adaptation responses from IT neurons, and surprising, deviant visual stimuli elicit firing rates that were indistinguishable from those evoked by rare but conditionally unsurprising images [31]. Olson and colleagues have demonstrated that after repeatedly pairing two visual images, the trailing stimulus will elicit a reduced neural response in the inferior temporal cortex (IT) when preceded by the leading stimulus (relative to another image)

[32-34]. In these studies, however, the stimuli were viewed passively rather than being decision-relevant, and stimulus probabilities were fixed over the course of the experiment. These factors preclude investigation of how short-and long-term fluctuations in expectation influenced perceptual decisions [32, 34]. The finding that visual neurons are sensitive to visual predictions and their updates has the potential to place a strong constraint on computational models of vision in neuroscience and machine learning [35-37].

Here, we asked whether IT neurons encode long-term probabilistic information about stimulus occurrence, and how these neural signals might influence perceptual decisions. Unlike previous studies [30-32, 34], we used a psychophysical approach. Neuronal activity was recorded from individual IT neurons whilst monkeys performed a perceptual decision-making task that required them to indicate whether a noisy cue stimulus was either face or fruit stimulus. The relative probability of the cue being a face vs. a fruit was manipulated by a latent variable that was not revealed to the monkeys, and which changed unpredictably over the course of each recording session. We show that neural signals in IT carry information about the long-term stimulation history, and this information is correlated with perceptual decisions.

Results

Two monkeys performed a perceptual decision-making task that involved discriminating whether a noisy image (cue) was a face or a fruit (**Fig. 1a**). Following a variable delay period, decisions were signalled with a saccadic eye movement to one of two images that occurred randomly on the left and right of the screen: either (i) the same face or fruit stimulus (correct response) or (ii) a stimulus from the opposing category (incorrect response). Cue stimuli were degraded at two levels by the addition of Gaussian noise to the image (low and high noise trials). Cue probability was controlled by a latent variable, *p(face)*, which took on one of 5 levels (0%, 25%, 50%, 75%, 100%), and

changed unpredictably over the course of the experiment (on average every ~50 trials). Recordings were obtained from 253 neurons from areas TEO and TE (lower bank of the STS, between 5-19 mm anterior to the interaural axis); concentrated near/within regions that had been identified by their preference for face stimuli in previous neuroimaging studies [38, 39].

Behavioural data. In **Fig. 1b** (left panels), we pooled the monkeys' average choices as a function of whether faces or fruit were expected (abscissa), for trials where the cue was a face (red lines) or a fruit (blue lines). Both monkeys showed above-chance sensitivity, as indicated by the fact that face and fruit stimuli predicted different choices even when the two classes of stimuli were equiprobable and prior information was thus unavailable (i.e., restricting the analysis to *p(face)* = 0.50: low noise trials: $t_{(118)} = 17.29$, $p < 1 \times 10^{-34}$; high noise trials: $t_{(118)} = 3.73$, $p < 1 \times 10^{-3}$) (analysis performed on z-transformed choice probabilities; see **Fig. S1** for data from each monkey).

Using the z-transformed hit and false alarm data (**Fig.1b**, right panels) from *p(face)* $\in$ [25%, 50%, 75%], we computed *d'* and *c*, decision theoretic statistics that indicate the respective sensitivity and decision criterion of the monkeys in each probability condition. As expected, *d'* was greater for low noise (1.37±0.51) than high noise (0.36±0.32) trials ($F_{(1,81)} = 247.0$, $p < 1 \times 10^{-16}$) but showed no main effect of probability condition ($p = 0.23$). By contrast, the estimate of criterion *c* varied with *p(face)* ($F_{(2,154)} = 96.75$, $p < 1 \times 10^{-14}$), but was insensitive to the visibility condition ($p = 0.55$). In both noise conditions, *c* was significantly below zero in the 25% condition and above zero in the 75% condition (all t-values > 6.7, all p-values $< 1 \times 10^{-9}$) but did not differ from zero in the 50% condition (both t-values < 1, both p-values > 0.24; note that estimates were unavailable for 0% and 100% conditions as only one stimulus class was presented). These data indicate that the monkeys were highly sensitive to the probability manipulation imposed by the task structure.

Neuronal responses to expected vs. unexpected stimuli. Overall, 174 neurons (69%) responded to visual stimuli (comparing baseline vs. post-stimulus activity in low-noise trials; $P<0.05$; see Supplemental Experimental Procedures for details). Of these, 106/174 neurons (61%) had significantly larger responses to faces over fruit; 2/174 neurons (1%) had larger responses to fruit over faces, and the remaining 38% of visually responsive neurons showed no significant difference in response strength between the two stimulus categories. In **Fig. 2a**, we show examples from each monkey of the responses of individual neurons to high and low noise face and fruit images.

Next, we examined whether the response to faces was modulated by face probability. The monkeys were not given any explicit information about the state of the latent variable *p(face)* so we calculated an "ideal" estimate of face probability using a hierarchical Bayesian approach that takes into account the rate of change (volatility) of the environment [40] (see Supplemental Experimental Procedures). The mean of the resulting posterior distribution (which we call $p(face)_{\text{Bayes}}$) served as a proxy for the monkeys' beliefs about the forthcoming stimulation (see **Fig. 1c** and below). To avoid arbitrary selection criteria and guard against circular inference, we conducted our analyses using the entire population of 253 neurons, irrespective of their visual selectivity.

We classified face and fruit cue trials into three conditions: "expected" [$p(face)_{\text{Bayes}} > 0.66$], "unexpected" [$p(face)_{\text{Bayes}} < 0.33$] or "neutral" [$0.33 < p(face)_{\text{Bayes}} < 0.66$] and plotted average firing rates for the example neurons in **Fig. 2b**, and for the whole population in **Fig. 3a-d**. Under low noise conditions, unexpected faces elicited greater activity than neutral faces and neutral faces elicited greater activity than expected faces (**Fig. 3a**). These effects were both significant from about 150 ms onwards. In addition, unexpected faces elicited greater activity than neutral faces regardless of the nature of the previous trial (**Fig. S2a,b**). For consistency with the past literature, in what follows we refer to these two effects collectively as "expectation suppression", although **Fig.**

3a shows that relative to neutral stimuli, neural responses to surprising stimuli were enhanced and neural responses to expected stimuli were attenuated. No such effects were observed for fruit trials (**Fig. 3b,d**) and firing rates were generally much lower under high noise conditions (**Fig. 3c**).

When faces were expected, the previous stimulus was more likely to be a face, and monkeys were thus more likely to have made a saccade to a fully visible face in the immediately preceding trial. Thus, modulation of neural activity by $p(face)_{\text{Bayes}}$ may have been driven by face repetition, rather than expectation of the face. To address this issue, we built a regression model that predicted neuronal activity as a function of a combination of experimental variables, including the current cue stimulus (face or fruit), the latent estimate of face probability $p(face)_{\text{Bayes}}$, and the interaction between these two variables. This approach allowed us to include nuisance predictors encoding the interaction between the current stimulus and previous choices (prevchoice1-3) in the design matrix. Thus, firing rates y at each timepoint were predicted as follows:

$$y = \beta_0 + \beta_1 stimulus + \beta_2 p(face)_{Bayes} + \beta_3 stimulus \times p(face)_{Bayes} + \beta_4 prevchoice_1 \times stimulus + \beta_5 prevchoice_2 \times stimulus + \beta_6 prevchoice_3 \times stimulus + \beta_7 trial \quad \text{(eq. 1)}$$

where *stimulus* indicates whether the cue was a face (*stimulus*=1) or a fruit (*stimulus*=0), $prevchoice_k$ denotes the choice made by the monkey k trials previously, and *trial* is trial number within the given recording session.

The coefficient β3 encodes how neural responsivity to faces vs. fruit was modulated by face probability, and is plotted over time (for low noise trials) in **Fig. 3e**, with red horizontal bars indicating significant timepoints (cluster-corrected over time, $p < 0.05$). Beginning around 150 ms

following cue stimulus onset, there is a statistically significant negative deflection, indicating that average firing rates decrease as the probability of face stimuli increases. Critically, this effect was robust to the inclusion of the monkeys' choices on up to three previous trials (shown in **Fig. 3e** as three separate but largely overlapping red traces). Moreoever, this effect persisted when previous cue, not choice, was used in the model (**Fig. S2c,d**). These data suggest that the effect observed is one of "expectation suppression" rather than "repetition suppression" [30, 41].

Fig. 3f shows the regression coefficient for stimulus category (faces vs. fruit) as a function of time from cue onset. This result is consistent with the population spike density functions (**Fig. 3a**), and confirms the ability of this regression-based approach to reveal the strong stimulus selectivity within the population.

Estimating *p(face)* using a reinforcement learning model. In the analyses described above, we assume that the monkeys estimated the state of the latent variable *p(face)* in a Bayesian fashion, marginalising over a posterior distribution encoding possible values of *p(face)*. However, another possibility is that the monkeys encoded a scalar estimate of *p(face)*, which they updated in proportion to the surprise it engendered at the time of the cue on each trial:

$$\Delta p(face)_{RL_t} = stimulus_t - p(face)_{RL_t} \quad \text{(eq. 2)}$$

In this class of reinforcement learning (RL) model [42], the rate of update that occurs following feedback on trial t is controlled by a free parameter α, known as the learning rate, as follows:

$$p(face)_{RL_{t+1}} = p(face)_{RL_t} + \alpha \left[\Delta p(_{face})_{RL_t}\right] \quad \text{(eq. 3)}$$

In equations 2 and 3, $p(face)_{RL}$ is the reinforcement learning (RL) model-derived estimate of face probability for each trial. To estimate the learning rate that best described behaviour on this task, we searched for values of α that optimised the fit to the data shown in **Fig. 1b** (see Supplemental Experimental Procedures). A learning rate of 1 would imply the monkeys used only the immediately preceding trial to guide future choices, whereas lower learning rates indicate that choices are changing as a function of previous trial history. Averaging over all sessions, the best fit was provided by $\alpha = 0.05$, indicating that monkeys' choices depended on the history of stimulation going back several trials. Model performance is indicated by the dots in **Fig. 1b**, which closely match the monkeys' behaviour. **Fig. 1c** shows an example of how $p(face)_{RL}$ (green line) varies for a single recording session, alongside estimates of $p(face)_{Bayes}$ (turquoise line) and the true generative probability (black line).

We compared the ability of the Bayesian and the RL models to account for choices using maximum likelihood estimation (see Supplemental Experimental Procedures). The RL model provided a substantially better fit than the Bayesian model. Using Bayesian model selection [43], the posterior probability of the RL model, p(m|data), was ~0.92 (~0.08 for the Bayesian model), providing strong evidence that monkeys learned more slowly than an ideal observer.

Neuronal correlates of predictions and prediction errors. We assessed whether IT neurons encoded predictions ($p(face)_{RL}$) and prediction errors ($\Delta p(face)_{RL}$), over and above the effect of sensory stimulation:

$$y = \beta_0 + \beta_1 stimulus + \beta_2 p(face)_{RL} + \beta_3 \Delta p(face)_{RL} + \beta_4 prevchoice_1 \times stimulus + \beta_5 trial \quad \text{(eq. 4)}$$

The resulting coefficients, plotted in **Fig. 4a**, confirmed that the IT population encoded a signed prediction error signal (**Fig. 4a**, centre left panel; showing significantly increased activity when faces were surprising), over and above the variance captured by the stimulus itself. However, the coefficients for $p(face)_{RL}$ failed to diverge significantly from zero (**Fig. 4a**, leftmost panel), suggesting that the population did not show a global increase or decrease in firing rate when faces were more probable than fruit.

Multivariate analyses. The analyses described above treat the IT neurons as a homogenous population. However, a deeper insight into the selectivity for stimulus, probability, and prediction errors can be gained by looking at the correlation between the patterns of selectivity for predictor variables (e.g., stimulus, prediction, prediction error) across the neuronal population [44]. We randomly allocated the trials recorded for each neuron into two independent groups of equal number, and estimated beta coefficients from equation 4 for both halves of the data. Each coefficient provides an independent estimate of a neuron's selectivity for one of the predictor variables (the extent to which the neuron "encodes" that variable) at each point in time. We then plotted the topography of positive (red) or negative (blue) correlations between regression coefficients β1-4 estimated from one half (x-axis) and the other (y-axis) of the data, in **Fig. 4b.**

Although we observed no average firing rate increase among IT neurons when faces were expected (see above), the pattern of activity encoding $p(face)_{RL}$ was conserved across the two halves of the data (**Fig. 4b, leftmost panel**). In other words, some neurons responded positively to $p(face)_{RL}$, and some responded negatively, but this variability was consistent across the two halves of the data and thus not simply driven by noise. We assessed the statistical significance of these correlations by comparing nonparametric Fisher's Z-scores to a null distribution computed from the shuffled data (see Supplemental Experimental Procedures). Averaging across all timepoints, the mean correlation

deviated positively from zero (Fisher's $Z > 10$, $p < 0.001$) as it did for each of the 4 quadrants (pre/post-stimulus periods for each variable; all Fisher's $Z > 6$, $p < 0.001$.

In **Fig. 4c** and **4d** (left panels), we show scatter plots of the parameter estimates for face predictions calculated from the two independent splits of the data, both in the pre-stimulus period (**Fig. 4c**) and post-stimulus period (**Fig. 4d**). As expected, there was a reliable consistency across the population for the encoding of prediction errors, in the post- but not pre-stimulus periods (**Fig. 4c** and **4d**, right panels).

This finding suggests that after measuring the response to $p(face)_{RL}$ from one portion of the data, it would be possible to "decode" the latent variable determining face probability given the neuronal activity from independent test datasets (see below). Furthermore, the existence of a cross-correlation between timepoints taken from the pre- and post-stimulus periods indicates that the neuronal encoding of probability is sustained over the entire cue epoch (-500 ms to +500 ms), including during the pre-stimulus period. Of note, an identical analysis conducted on the predictor prevchoice failed to reach significance for any of the quadrants (**Fig. 4b,** rightmost panel; all p-values > 0.2).

The analysis described above assesses the consistency between two independent estimates of neural encoding of a single predictor (e.g. $p(face)_{RL}$ with $p(face)_{RL}$). However, the approach can also be deployed to assess how the neuronal encoding of one variable (e.g. $p(face)_{RL}$) relates to an independent estimate of encoding of another (e.g. $\Delta p(face)_{RL}$). In a subsequent set of analyses, we explored the interrelationship between population encoding of faces, face predictions and face surprise (**Fig. 5**; see **Fig. S4** for all possible pairings of β1-4). We show these correlations both timepoint-by-timepoint (**Fig. 5a**) and as scatter plots for coefficients estimated from the two

independent splits of the data in the pre-stimulus period (**Fig. 5b**) and post-stimulus period (**Fig. 5c**).

The patterns of encoding for the stimulus (face > fruit) and for the prediction error ($\Delta p(face)_{RL}$) were highly correlated from about 150 ms onwards in the post-stimulus period (**Fig. 5, middle panels**). This effect was significant when averaging over the post-stimulus period of each half of the data (i.e., the lower right quadrant of this panel; Fisher's $Z > 14$, $p < 0.001$) but not when examining data from the pre-stimulus period (other quadrants: all p-values > 0.9). In other words, those neurons that showed the strongest face responses also showed the greatest difference in response magnitude between expected and unexpected face trials, and vice versa. Indeed, when we divided neurons into two halves according to whether they were more face-selective or more fruit-selective, we observed a face prediction error for the former, and a fruit prediction error for the latter (**Fig. S5**), even though fruit selectivity in the dataset was low when defined by conventional approaches.

Furthermore, we observed that in the post-stimulus period, there was a significant positive correlation between the extent to which IT neurons were sensitive to faces and the extent to which they encoded face predictions (Fisher's $Z = 17$, $p < 0.001$; **Fig. 5**, left panels). That is, although face predictions were not encoded on average in the population, they were encoded in those neurons that responded most vigorously to faces over fruit. However, there was no significant correlation between the extent to which neurons encoded face predictions ($p(face)_{RL}$) and face prediction errors ($\Delta p(face)_{RL}$), either in the post-stimulus period or in any other quadrant (**Fig. 5**, right panels). Moreover, there was a significantly stronger correlation between the encoding of faces and face predictions than there was between the encoding of face prediction errors and face predictions, i.e. between the correlation values shown in the leftmost and rightmost panels of Fig. 5a, and this held

for each quadrant tested (all z-values > 4, all p-values < 0.0001). This finding is consistent with influential theories that have argued for distinct populations of neurons encoding predictions and prediction error signals during perceptual decision-making [7, 10].

Neuronal activity at the time of choice modulated by prediction signals. The monkeys indicated their decision by making an eye movement to one of two fully visible probe stimuli, a face or a fruit. The neuronal activity at the time of response was thus heavily influenced by the choice because this determined where the gaze rested. Neuronal activity locked to the behavioural response (i.e., to the start of the 500 ms fixation hold period; shown in **Fig. S6**) were analysed via a new regression model that included the cue, its interaction with $p(face)_{RL,}$ as well as several nuisance quantities:

$$y = \beta_0 + \beta_1 stimulus + \beta_2 p(face)_{RL} + \beta_3 p(face)_{RL} \times stimulus + \beta_4 prevchoice + \beta_5 trial \quad \text{(eq. 5)}$$

In **Fig. 6a**, the coefficients for β1-β3 are shown over time for correct trials only. Here, we averaged over high and low noise trials, because the observed responses were very similar. As expected, the population strongly encoded the main effect of whether the cue was a face or not (i.e., the choice, on correct trials) after response onset (leftmost panel). An inverted effect was observed (blue lines) when we computed the same regression for incorrect trials, which is as expected given the monkey responded with a saccade to a fruit probe on these trials. Critically, we observed an attenuation in firing rate for expected stimuli at the time of choice, just as we did at the time of the cue, as shown by the interaction between choice and probability (rightmost panel). Multivariate analysis at the time of the saccadic response also showed reliable encoding of the choice, face probability, and the cross-correlation between neurons coding for these two variables (**Fig. S6**).

Decoding cue identity with and without expectations. Above, we described a difference in IT firing rates consistent with both "surprise enhancement" and "expectation suppression". This is consistent with previous neuroimaging studies [28]. However, one fMRI study has suggested that although expected stimuli elicit lower overall signal levels, they can be decoded with greater accuracy from multivariate BOLD signals [17]. We tested this hypothesis at the single-neuron level both at the time of the cue and the time of choice. For each neuron, we pseudorandomly allocated trials into a training set (70%) and a testing set (30%), and estimated coefficients that linked the pattern of neural activity on training trials to the identity of the cue stimulus (face or fruit). We then used these coefficients to make predictions about the cue identity on the test set. We performed this analysis separately for those trials for which only weak expectations about stimulus identity were possible ($0.33 < p(face)_{\text{Bayes}} < 0.66$; "neutral" trials) and those where the expectation and the stimulus were congruent ("expectation" trials) We used an undersampling approach to equate trial numbers in each condition, and excluded neurons for which fewer than 20 trials were available. Statistical significance was calculated using a permutation testing approach (see Supplemental Experimental Procedures).

On expectation trials (**Fig. 6b**, red traces), at the time of the cue, it was possible to decode the forthcoming stimulus during the pre-stimulus period because the neural expectation signals carried information about its likely identity. This was not possible on neutral trials (black traces). Under high noise (**Fig. 6c**), neural signals afforded the monkeys an opportunity to decide about the cue on the basis of expectations alone, whereas under low noise (**Fig. 6b**), the relative decoding advantage for expectation vs. neutral trials was diminished, particularly during the post-stimulus period. Similarly at the time of the choice, there was a substantial decoding advantage when expectations were present under both high- and low-noise trials.

Discussion

The best decisions are made by combining sensory signals with prior knowledge of stimulus probabilities. During tasks in which stimuli are perceptually unambiguous but yield a time-varying reward, humans and monkeys dynamically update their behaviour to reflect the statistics of the environment, and regions of the medial prefrontal cortex encode and update probabilistic information about outcome values [40]. Evidence for long-term probabilistic encoding in sensory regions, however, remains weak. Studies employing passive statistical learning designs have demonstrated that BOLD signals in extrastriate cortex are attenuated when a visual stimulus is conditionally probable, based on either an immediately preceding cue or knowledge of the stimulation history [28]. However, these studies have two major limitations. First, passive viewing designs do not provide behavioural indices of statistical learning, making it unclear how changes in brain activity are linked to behaviour. Second, the BOLD signal is an indirect measure of neuronal activity, and it is unknown how the observed effects in sensory cortices relate to underlying firing patterns in sensory neurons.

Our data revealed strong evidence for "expectation suppression" at the single-neuron level: when a face was presented, firing rates depended on whether that face was *a priori* probable or not. We found that this effect was robust to the inclusion of the last three repetitions in the stimulus sequence, excluding the possibility that effects of expectation were a consequence of local adaptation to the information occurring on immediately preceding trials ("repetition suppression"). This finding, coupled with behavioural evidence that the monkey learned slowly about the changing probability of face occurrence, suggests that IT neurons can encode long-term predictions about the likelihood of stimulus occurrence.

Why our data show clear evidence for expectation suppression that has been absent in other single-neuron studies is unclear, but it may be that under passive cueing conditions, the monkeys have only weak incentives to attend to the statistical structure of the task. On the other hand, the attentional requirements of more demanding behavioural tasks such as ours may be important for statistical learning [45] and its neuronal concomitants [20]. Our observed effect began at ~200 ms following cue onset, about 100 ms later than the attenuation that is typically observed when an image is repeated [30, 31] or follows an image with which it has been repeatedly paired [32, 34]. This later onset suggests that our data reflect a distinct expectation-based mechanism, consistent with MEG studies that have reported distinct timecourses for repetition and expectation suppression [46].

The results reported here are consistent with the view that sensory signals are subtractively normalised or "explained away" by probabilistic information stretching back over several trials. Subtractive normalisation allows incoming sensory information to be represented more efficiently, but requires that two quantities – expectation (prediction) and surprise (prediction error) - are encoded at each stage of the sensory processing hierarchy. To test for the encoding of these quantities in IT signals, we fit the behavioural data with a class of learning models often used to understand reward-guided learning, in which expectations are updated as a function of the surprise engendered by a new event. Although this simple model is probably a relatively crude approximation to the underlying dynamics of perceptual inference, scalar prediction error signals captured variance in aggregate neuronal signals, over and above the influence of the stimulus itself. These data provide support for accounts of visual computation collectively known as "predictive coding" [7-9].

388 Although the majority of recorded IT neurons were face-selective, aggregate firing rates did not
389 vary according to subjective estimates of the latent probability of face occurrence. However, a
390 multivariate analysis technique that measures correlations in the neuronal encoding of predictor
391 variables between independent splits of the data revealed that encoding of this quantity was stable
392 across the population and across the trial. In other words, some neuronal responses scaled positively
393 with $p(face)_{RL}$ and others negatively, but this variation was highly consistent over trials (and thus
394 not simply noise). Moreover, although it was possible to decode the to-be-discriminated stimulus
395 from IT neural activity, both where an expectation was present and on neutral trials, this was
396 possible even before stimulus onset on expectation trials – i.e., when *p(face)* positively predicted
397 the identity of the stimulus – because in that case, the population encoded predictive information
398 about the stimulus identity. This was also true on high noise trials, where subsequent decoding of
399 the stimulus identity was weak. The finding that "expected" stimuli are decoded with greater
400 accuracy – which was even more pronounced at the time of choice – is consistent with the
401 observation from neuroimaging studies that expectation suppression is accompanied by increased
402 decoding accuracy for expected stimuli from multivoxel patterns in the visual cortex [17].
403
404 Sensory predictions and prediction errors might be computed by two different classes of neuron;
405 distinguished by their laminar profile and pattern of interconnection within and between cortical
406 stages [10, 47]. One corollary of this view is that those neurons that encode face expectations
407 should not encode face surprise and vice versa. When we examined the relationship between
408 encoding of expectation and surprise across the population, this encoding was uncorrelated. This
409 contrasts with the encoding of the stimulus (face > fruit) and the face prediction, which were
410 significantly correlated across the neuronal population in the post-stimulus period. That is, although
411 face selectivity and face surprise are related, the face prediction variable is only related to that
412 subspace of the selectivity data that is orthogonal to the encoding of face surprise.

In summary, our results demonstrate that IT neurons encode long-term, latent probabilistic information about stimulus occurrence. This information is represented over the course of a perceptual decision including the pre-stimulus period and around the time of the choice. These results support 'predictive coding' theories of perceptual decision-making.

Author Contributions

Conceptualization: AHB and LGU; Investigation: AHB, ELM, and NJM; Formal Analysis: AHB and CS; Writing – Original Draft: AHB and CS; Writing – Review and Editing: AHB, CS, ELM, NJM, and LGU.

Acknowledgements

This work was supported by the National Institute of Mental Health Intramural Research Program (AHB, NJM, ELM, and LGU) and by a European Research Council (ERC) Starter Grant (281628) to CS. The authors would like to thank David Leopold and Mark Stokes for feedback on earlier drafts, and Lucy Guillory and Jennifer Frihauf for their assistance with training the monkeys.

Figure Legends

Figure 1 Behavioural performance in delayed match-to-sample task. (a) Monkeys were trained on a delayed match-to-sample task that required them to identify which of two stimuli (a face or fruit) best matched a previously presented cue. Cue stimuli were pseudorandomly selected on a session-by-session basis from a set of 8 possible images for each category and were degraded at two different levels by the addition of Gaussian noise to the image (low and high noise trials) (N.B. this panel shows an example of a low-noise cue). The trials were arranged into 5 blocks, (presented in random order), each with a different probability of the cue being a "face" vs. a "fruit" (0%, 25%, 50%, 75%, 100%). (b) Left Panels: A robust behavioural effect emerged whereby the subject's choice was strongly biased by prior probability – "face" was chosen more often than "fruit" when trial proportions were biased towards faces and vice-versa. This trend was found at both noise levels. Right Panels: Hit rates and false alarms were more influenced by *p(face)* under high noise conditions as compared to low noise conditions (grey lines). SEM indicated by error bars. See Figure S1 for behavioural group according to monkey. (c) Estimates of *p(face)* over a single

443 recording session. Estimates were derived using a reinforcement learning (delta rule) model (green
444 line) and a Bayesian model (turquoise line), and are compared to the generative probability (black
445 line). Ovals along the bottom of the figure indicate trials where the cue stimulus was a face. See
446 also Figure S1.
447
448
449 **Figure 2** Responses of individual IT neurons to face and fruit stimuli. (a) Four examples of neurons
450 from monkey 1 (left, middle left panel) and monkey 2 (middle right, right panel) to high (dashed
451 lines) and low (solid lines) noise face (red) and fruit (blue) stimuli. In all four examples, low noise
452 face stimuli evoked a strong and persistent response. Fruit stimuli and high noise evoked
453 comparatively weaker responses. (b) Responses of the same four neurons, with trials sorted
454 according to face expectation. Solid lines show responses for trials where faces were expected
455 (p(face)>0.66); dashed lines show responses for trials where face and fruit were equally expected
456 (0.66 > p(face) > 0.33); dotted lines show responses for trials where face were not expected (i.e.,
457 fruit stimuli were expected; p(face)<0.33). All neuronal responses were smoothed with a 10 ms
458 Gaussian kernel (see Supplemental Experimental Procedures).
459
460 **Figure 3** Population **r**esponses in IT to expected, neutral, and unexpected faces (a,c) and fruit
461 stimuli (b,d) for low and high noise trials. A robust and sustained increase in activity in response to
462 face stimuli was observed, beginning around 100 ms following cue onset (a). Black bar indicates
463 cue presentation period. For (a) and (d), red/blue bars indicate timepoints where average firing rates
464 for expected and unexpected faces/fruit were significantly different from one another (light bars,
465 unexpected > neutral; dark bars, neutral > expected). (e) Cue-aligned regression coefficients for the
466 interaction between stimulus category and p(face) (β3), and (f) stimulus category (β3) for the
467 regression analysis from eq. 1. In both cases, three separate lines (largely overlapping) show betas

from separate regressions in which the previous 1 (dark red lines), 2 (medium red) or 3 (light red) trials were included as nuisance covariates. Neuronal responses were smoothed with a 10 ms Gaussian kernel (see Supplemental Experimental Procedures). The red bars show the corresponding significant timepoints for each analysis. To correct for multiple comparisons, a cluster-correction was applied across timepoints ($p<0.05$*). See also Figure S2.

Figure 4 Prediction signals in IT neurons. (a) Cue-aligned regression coefficients for prediction (β_2), prediction error (β_3), stimulus category (β_1), and the interaction between stimulus category and previous choice (β_4) all from equation 4. Black bars indicate stimulus presentation period. Red bars show cluster-corrected significant timepoints. (b) Split-half cross-validation of cue-aligned regression coefficients within each predictor variable. Each plot shows the Pearson's correlation between regression coefficients for two independent splits of the data (x axis, y axis). Each panel shows correlations within a predictor variable (e.g., β_2 and β_2 in left panel) and for a given timepoint with every other timepoint. Points along the diagonal of each panel show correlations in selectivity among equivalent timepoints (e.g., 200 ms post-stimulus in both halves of the data), whereas off-diagonal points indicate correlations between differing timepoints (e.g., 100 ms post-stimulus in one half of the data, with 300 ms in the other). Black contour lines indicate timepoints where a significant correlation was observed ($p < 0.05$), adjusted for multiple comparisons using a cluster correction method. Dashed lines indicate the quadrants used for significance testing in the main text. (c,d) Scatter plots between independently-obtained parameter estimates from $p\mathit{face}_{RL}$ and $\Delta p\mathit{face}_{RL}$ for each neurons in the (c) pre-stimulus and (d) post-stimulus periods. Inset text shows the p-value for the corrected non-parametric test (see Supplemental Experimental Procedures). Red line shows the best-fitting linear trend for significant correlations. See also Figure S3.

Figure 5 Correlations between encoding of faces, face predictions and face prediction errors in IT neurons. (a) cross-validation of cue-aligned regression coefficients. Each panel shows the Pearson's correlation between two different predictor variables for two independent splits of the data (e.g., β_1 and β_2, left panel). Every timepoint is correlated with every other timepoint. Black contour lines indicate timepoints where a significant correlation was observed ($p < 0.01$), adjusted for multiple comparisons using a cluster correction method. Dashed lines indicate the quadrants used for significance testing in the main text. Correlation between parameter estimates for each neuron in the (c) pre-stimulus and (d) post-stimulus periods. Inset text shows the p-value for the corrected non-parametric test (see Supplemental Experimental Procedures). See also Figures S4 and S5.

Figure 6 Encoding of prediction signals at the time of choice. (a) Regression coefficients (from eq. 5) for choice (β_1), prediction (β_2), and their interaction (β_3) at the time the monkey made its choice (aligned on eye-on-target), for correct trials only (red lines) and incorrect trials only (blue lines). Average responses of IT neurons at time of choice shown in Figure S3. Cross-validation of choice-aligned regression coefficients show in Figure S4. (b,c) Decoding accuracy for the cue identity from cue (left panels) and response-aligned activity (right panels) for low noise (b) and high noise (c) trials. Decoding of cue identity was significantly above chance for the majority of the trial duration, in cases where a given cue was expected (i.e., *p(face)*>0.66 and a face stimulus occurred, or where *p(face)*<0.33 and a fruit cue occurred). This was not the case for neutral trials (i.e., *p(face)* > 0.33 & <0.66), where cue identity could not be decoded until after cue onset. Red bars and black bars show timepoints where decoding was significantly above chance for the expected and neutral trials respectively, computed using a permutation testing method. Trial counts were equated for all analysis conditions. See also Figure S6.

References

1. Helmholtz, H.v. (1896). Handbuch der Physiologischen Optik, Dritter Abschnitt., Volume Voss, (Hamburg).

2. Hawkins, J. (2004). On Intelligence, (Times Books).

3. Gregory, R.L. (1980). Perceptions as hypotheses. Philos Trans R Soc Lond B Biol Sci *290*, 181-197.

4. Kersten, D., Mamassian, P., and Yuille, A. (2004). Object perception as Bayesian inference. Annu Rev Psychol *55*, 271-304.

5. Geisler, W.S. (2008). Visual perception and the statistical properties of natural scenes. Annu Rev Psychol *59*, 167-192.

6. Simoncelli, E.P., and Olshausen, B.A. (2001). Natural image statistics and neural representation. Annu Rev Neurosci *24*, 1193-1216.

7. Friston, K. (2005). A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci *360*, 815-836.

8. Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. Biological cybernetics *66*, 241-251.

9. Rao, R.P., and Ballard, D.H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci *2*, 79-87.

10. Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P., and Friston, K.J. (2012). Canonical microcircuits for predictive coding. Neuron *76*, 695-711.

11. den Ouden, H.E., Friston, K.J., Daw, N.D., McIntosh, A.R., and Stephan, K.E. (2009). A dual role for prediction error in associative learning. Cereb Cortex *19*, 1175-1185.

12. Garrido, M.I., Friston, K.J., Kiebel, S.J., Stephan, K.E., Baldeweg, T., and Kilner, J.M. (2008). The functional anatomy of the MMN: a DCM study of the roving paradigm. Neuroimage *42*, 936-944.

13. de Gardelle, V., Waszczuk, M., Egner, T., and Summerfield, C. (2012). Concurrent Repetition Enhancement and Suppression Responses in Extrastriate Visual Cortex. Cereb Cortex.

14. Egner, T., Monti, J.M., and Summerfield, C. (2010). Expectation and surprise determine neural population responses in the ventral visual stream. J Neurosci *30*, 16601-16608.

15. Summerfield, C., Egner, T., Greene, M., Koechlin, E., Mangels, J., and Hirsch, J. (2006). Predictive codes for forthcoming perception in the frontal cortex. Science *314*, 1311-1314.

16. Summerfield, C., Trittschuh, E.H., Monti, J.M., Mesulam, M.M., and Egner, T. (2008). Neural repetition suppression reflects fulfilled perceptual expectations. Nat Neurosci *11*, 1004-1006.

17. Kok, P., Jehee, J.F., and de Lange, F.P. (2012). Less is more: expectation sharpens representations in the primary visual cortex. Neuron *75*, 265-270.

18. Kok, P., Rahnev, D., Jehee, J.F., Lau, H.C., and de Lange, F.P. (2012). Attention reverses the effect of prediction in silencing sensory signals. Cereb Cortex *22*, 2197-2206.

19. Todorovic, A., van Ede, F., Maris, E., and de Lange, F.P. (2011). Prior expectation mediates neural adaptation to repeated sounds in the auditory cortex: an MEG study. J Neurosci *31*, 9118-9123.

20. Larsson, J., and Smith, A.T. (2012). FMRI repetition suppression: neuronal adaptation or stimulus expectation? Cereb Cortex *22*, 567-576.

21. den Ouden, H.E., Daunizeau, J., Roiser, J., Friston, K.J., and Stephan, K.E. (2010). Striatal prediction error modulates cortical coupling. J Neurosci *30*, 3210-3219.

22. Grotheer, M., and Kovacs, G. (2014). Repetition probability effects depend on prior experiences. J Neurosci *34*, 6640-6646.

23. Kovacs, G., and Vogels, R. (2014). When does repetition suppression depend on repetition probability? Front Hum Neurosci *8*, 685.

24. Mayrhauser, L., Bergmann, J., Crone, J., and Kronbichler, M. (2014). Neural repetition suppression: evidence for perceptual expectation in object-selective regions. Front Hum Neurosci *8*, 225.

25. Kovacs, G., Iffland, L., Vidnyanszky, Z., and Greenlee, M.W. (2012). Stimulus repetition probability effects on repetition suppression are position invariant for faces. Neuroimage *60*, 2128-2135.

26. Wacongne, C., Labyt, E., van Wassenhove, V., Bekinschtein, T., Naccache, L., and Dehaene, S. (2011). Evidence for a hierarchy of predictions and prediction errors in human cortex. Proc Natl Acad Sci U S A *108*, 20754-20759.

27. Hansen, K.A., Hillenbrand, S.F., and Ungerleider, L.G. (2012). Effects of Prior Knowledge on Decisions Made Under Perceptual vs. Categorical Uncertainty. Front Neurosci *6*, 163.

28. Summerfield, C., and de Lange, F.P. (2014). Expectation in perceptual decision making: neural and computational mechanisms. Nat Rev Neurosci *15*, 745-756.

29. Den Ouden, H.M., Kok, P., and de Lange, F.P. (2012). How prediction errors shape perception, attention and motivation. Frontiers in Psychology *3*, 548.

30. Kaliukhovich, D.A., and Vogels, R. (2011). Stimulus repetition probability does not affect repetition suppression in macaque inferior temporal cortex. Cereb Cortex *21*, 1547-1558.

31. Kaliukhovich, D.A., and Vogels, R. (2014). Neurons in macaque inferior temporal cortex show no surprise response to deviants in visual oddball sequences. J Neurosci *34*, 12801-12815.

32. Meyer, T., and Olson, C.R. (2011). Statistical learning of visual transitions in monkey inferotemporal cortex. Proc Natl Acad Sci U S A *108*, 19401-19406.

581 33. Ramachandran, S., Meyer, T., and Olson, C.R. (2016). Prediction suppression in monkey inferotemporal
582 cortex depends on the conditional probability between images. J Neurophysiol *115*, 355-362.

583 34. Meyer, T., Ramachandran, S., and Olson, C.R. (2014). Statistical learning of serial visual transitions by
584 neurons in monkey inferotemporal cortex. J Neurosci *34*, 9332-9337.

585 35. Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. Nat
586 Neurosci *2*, 1019-1025.

587 36. Dayan, P., Hinton, G.E., Neal, R.M., and Zemel, R.S. (1995). The Helmholtz machine. Neural Comput
588 *7*, 889-904.

589 37. Gilbert, C.D., and Li, W. (2013). Top-down influences on visual processing. Nat Rev Neurosci *14*, 350-
590 363.

591 38. Bell, A.H., Hadj-Bouziane, F., Frihauf, J.B., Tootell, R.B., and Ungerleider, L.G. (2009). Object
592 representations in the temporal cortex of monkeys and humans as revealed by functional magnetic
593 resonance imaging. J Neurophysiol *101*, 688-700.

594 39. Bell, A.H., Malecek, N.J., Morin, E.L., Hadj-Bouziane, F., Tootell, R.B., and Ungerleider, L.G. (2011).
595 Relationship between functional magnetic resonance imaging-identified regions and neuronal category
596 selectivity. J Neurosci *31*, 12229-12240.

597 40. Behrens, T.E., Woolrich, M.W., Walton, M.E., and Rushworth, M.F. (2007). Learning the value of
598 information in an uncertain world. Nat Neurosci *10*, 1214-1221.

599 41. Grill-Spector, K., Henson, R., and Martin, A. (2006). Repetition and the brain: neural models of
600 stimulus-specific effects. Trends Cogn Sci *10*, 14-23.

601 42. Sutton, R., and Barto, A. (1998). Reinforcement Learning, (MIT press).

602 43. Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., and Friston, K.J. (2009). Bayesian model
603 selection for group studies. Neuroimage *46*, 1004-1017.

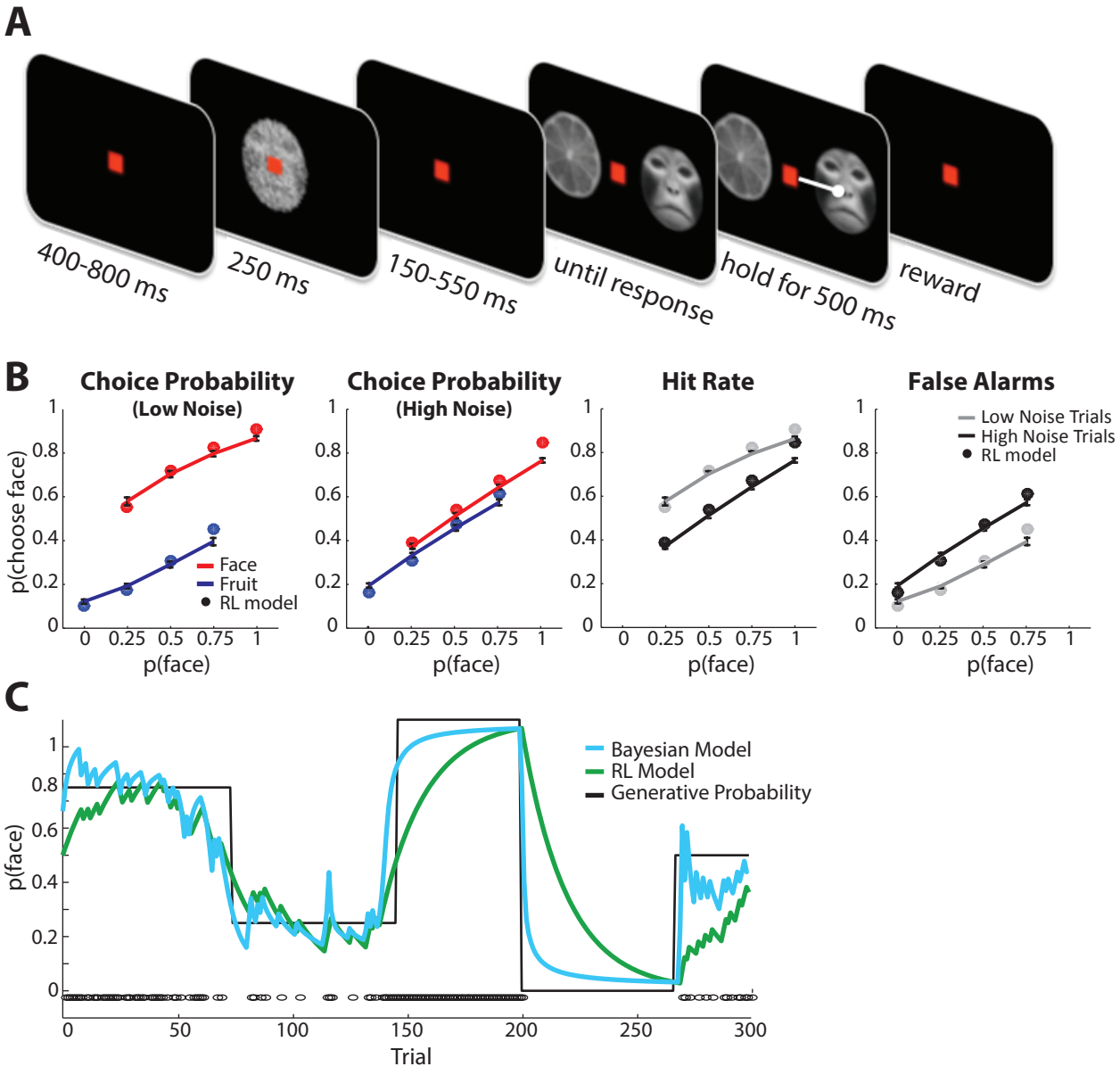
604 44. Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting
605 the branches of systems neuroscience. Front Syst Neurosci *2*, 4.

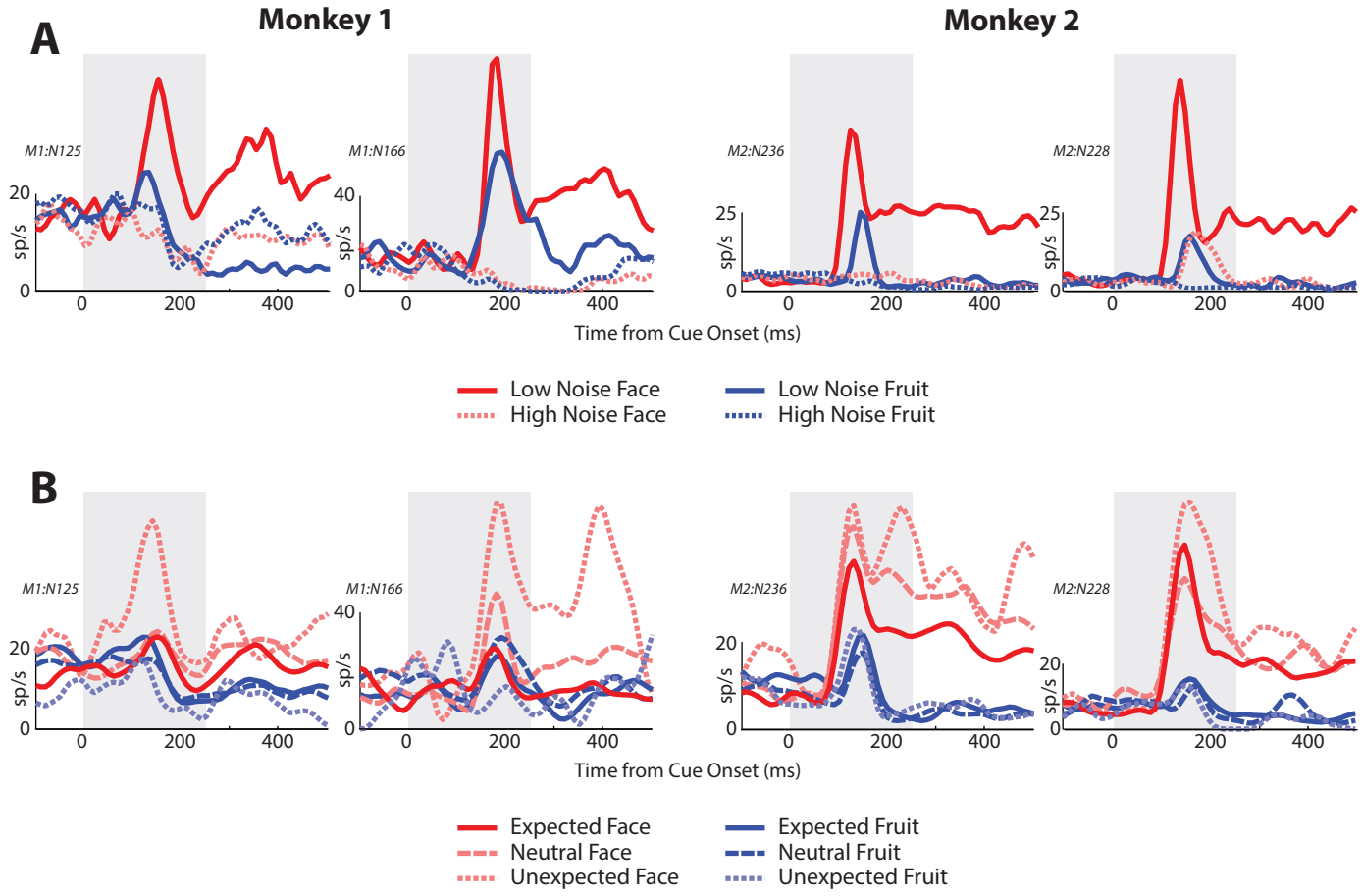
606 45. Turk-Browne, N.B., Junge, J., and Scholl, B.J. (2005). The automaticity of visual statistical learning. J
607 Exp Psychol Gen *134*, 552-564.

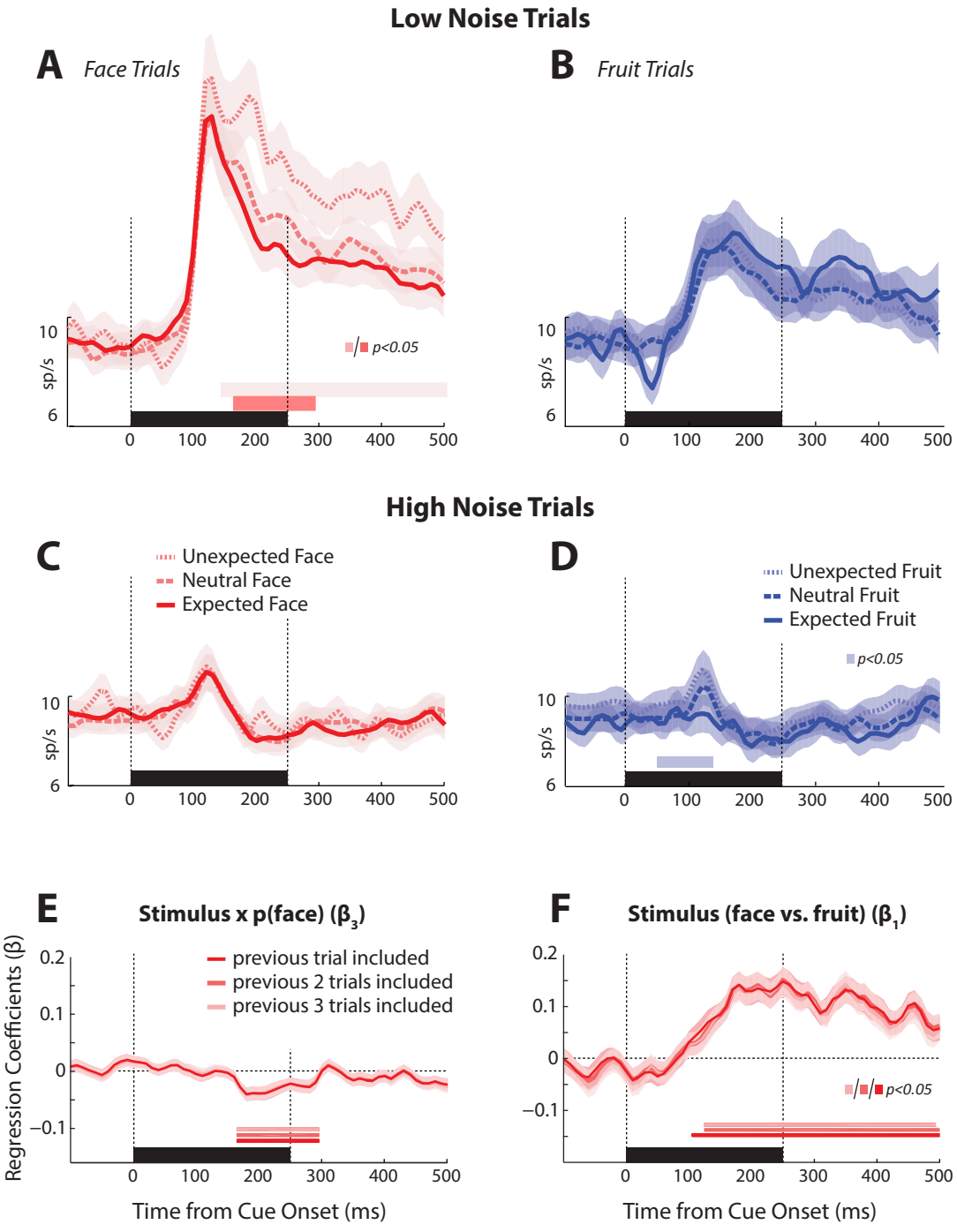
608 46. Todorovic, A., and de Lange, F.P. (2012). Repetition suppression and expectation suppression are
609 dissociable in time in early auditory evoked fields. J Neurosci *32*, 13389-13395.

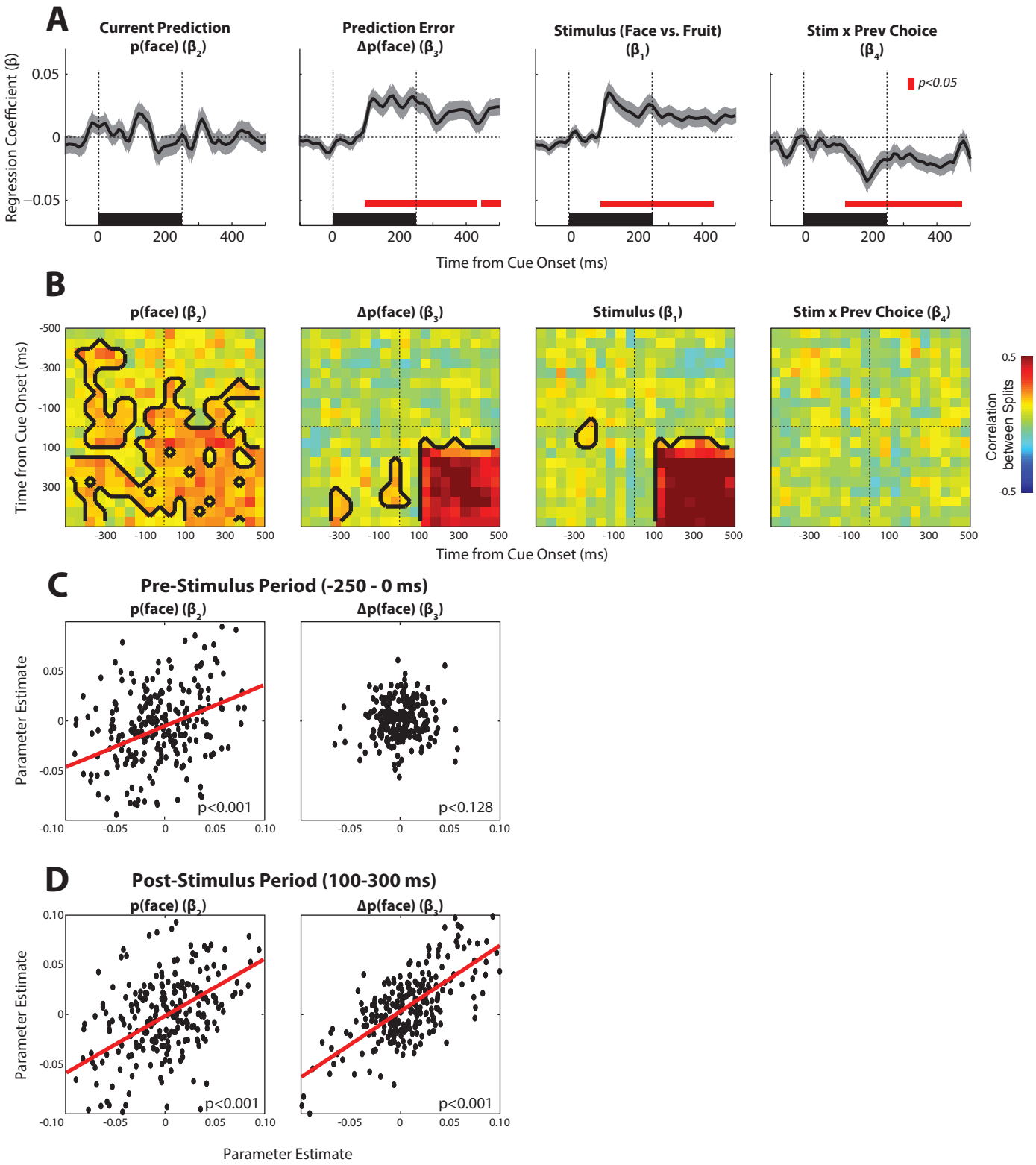
610 47. Kok, P., Bains, L.J., van Mourik, T., Norris, D.G., and de Lange, F.P. (2016). Selective Activation of the
611 Deep Layers of the Human Primary Visual Cortex by Top-Down Feedback. Curr Biol *26*, 371-376.

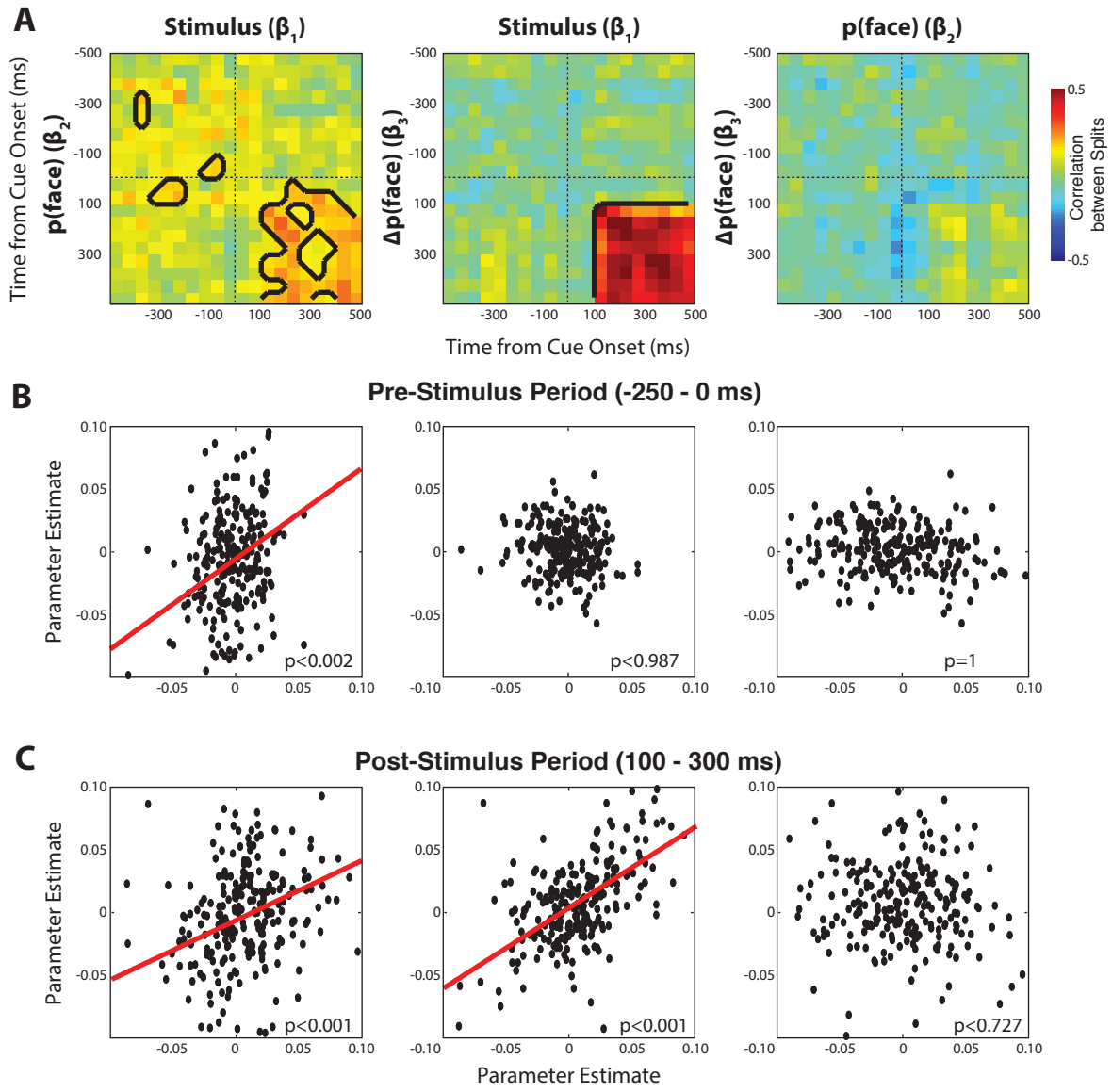
612

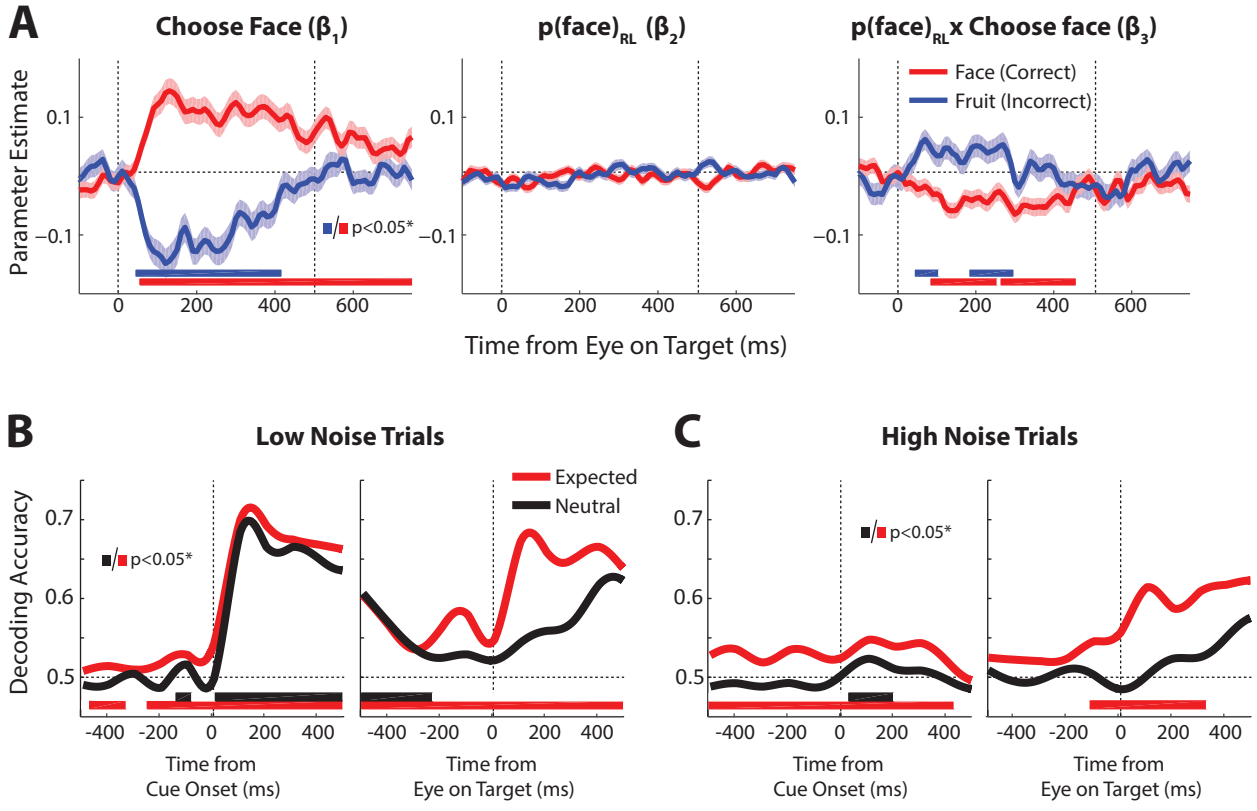

A
400-800 ms
250 ms
150-550 ms
until response
hold for 500 ms
reward
B
Choice Probability
(Low Noise)
Choice Probability
(High Noise)
Hit Rate
False Alarms
p(choose face)
p(face)
Face
Fruit
RL model
Low Noise Trials
High Noise Trials
C
Bayesian Model
RL Model
Generative Probability
p(face)
Trial

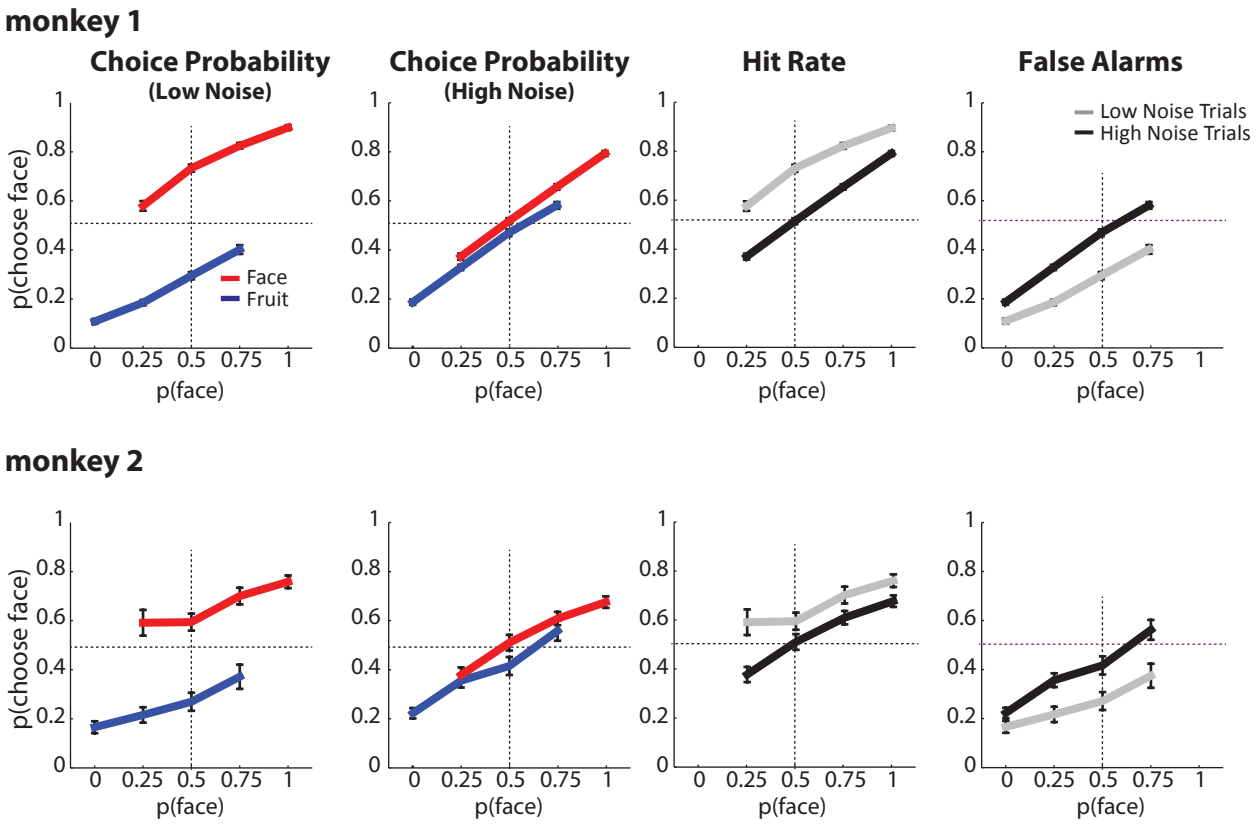

A
Monkey 1
Monkey 2
M1:N125
M1:N166
M2:N236
M2:N228
sp/s
Time from Cue Onset (ms)
Low Noise Face
High Noise Face
Low Noise Fruit
High Noise Fruit
B
Expected Face
Neutral Face
Unexpected Face
Expected Fruit
Neutral Fruit
Unexpected Fruit


Low Noise Trials
A
Face Trials
B
Fruit Trials
10
6
sp/s
0
100
200
300
400
500
p<0.05
High Noise Trials
C
Unexpected Face
Neutral Face
Expected Face
D
Unexpected Fruit
Neutral Fruit
Expected Fruit
p<0.05
E
Stimulus x p(face) (β3)
previous trial included
previous 2 trials included
previous 3 trials included
Regression Coefficients (β)
0.2
0.1
0
−0.1
Time from Cue Onset (ms)
F
Stimulus (face vs. fruit) (β1)
p<0.05
Time from Cue Onset (ms)

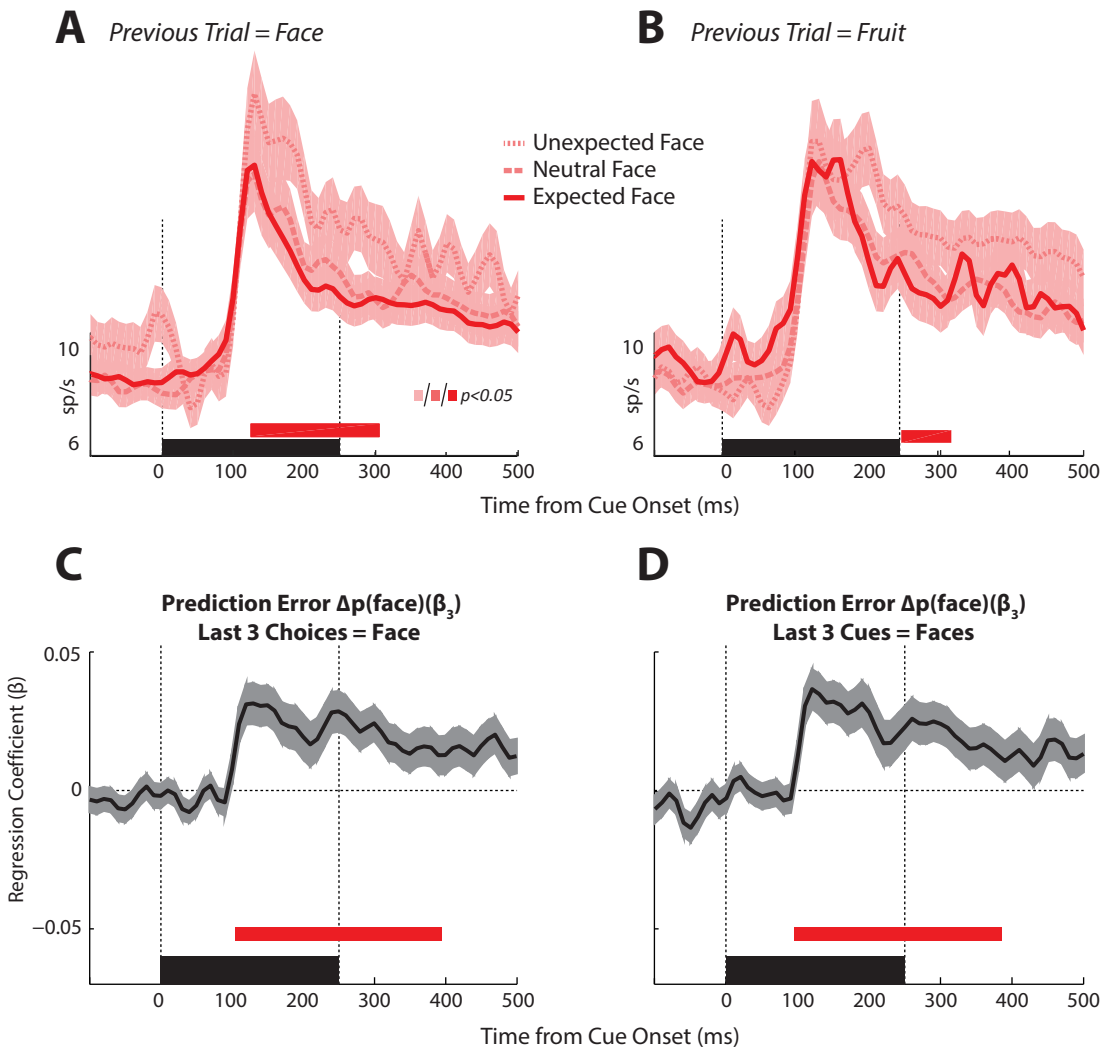

A
Current Prediction
p(face) (β2)
Prediction Error
Δp(face) (β3)
Stimulus (Face vs. Fruit)
(β1)
Stim x Prev Choice
(β4)
p<0.05
Regression Coefficient (β)
0.05
0
−0.05
0
200
400
Time from Cue Onset (ms)
B
p(face) (β2)
Δp(face) (β3)
Stimulus (β1)
Stim x Prev Choice (β4)
Time from Cue Onset (ms)
-500
-300
-100
100
300
500
Time from Cue Onset (ms)
Correlation between Splits
0.5
-0.5
C
Pre-Stimulus Period (-250 - 0 ms)
p(face) (β2)
Δp(face) (β3)
Parameter Estimate
p<0.001
p<0.128
D
Post-Stimulus Period (100-300 ms)
p(face) (β2)
Δp(face) (β3)
Parameter Estimate
p<0.001
p<0.001
Parameter Estimate


A
Stimulus (β1)
Stimulus (β1)
p(face) (β2)
Time from Cue Onset (ms)
p(face) (β2)
Δp(face) (β3)
Δp(face) (β3)
-500
-300
-100
100
300
500
Time from Cue Onset (ms)
0.5
Correlation between Splits
-0.5
B
Pre-Stimulus Period (-250 - 0 ms)
Parameter Estimate
0.10
0.05
0
-0.05
-0.10
p<0.002
p<0.987
p=1
C
Post-Stimulus Period (100 - 300 ms)
Parameter Estimate
p<0.001
p<0.001
p<0.727
Parameter Estimate

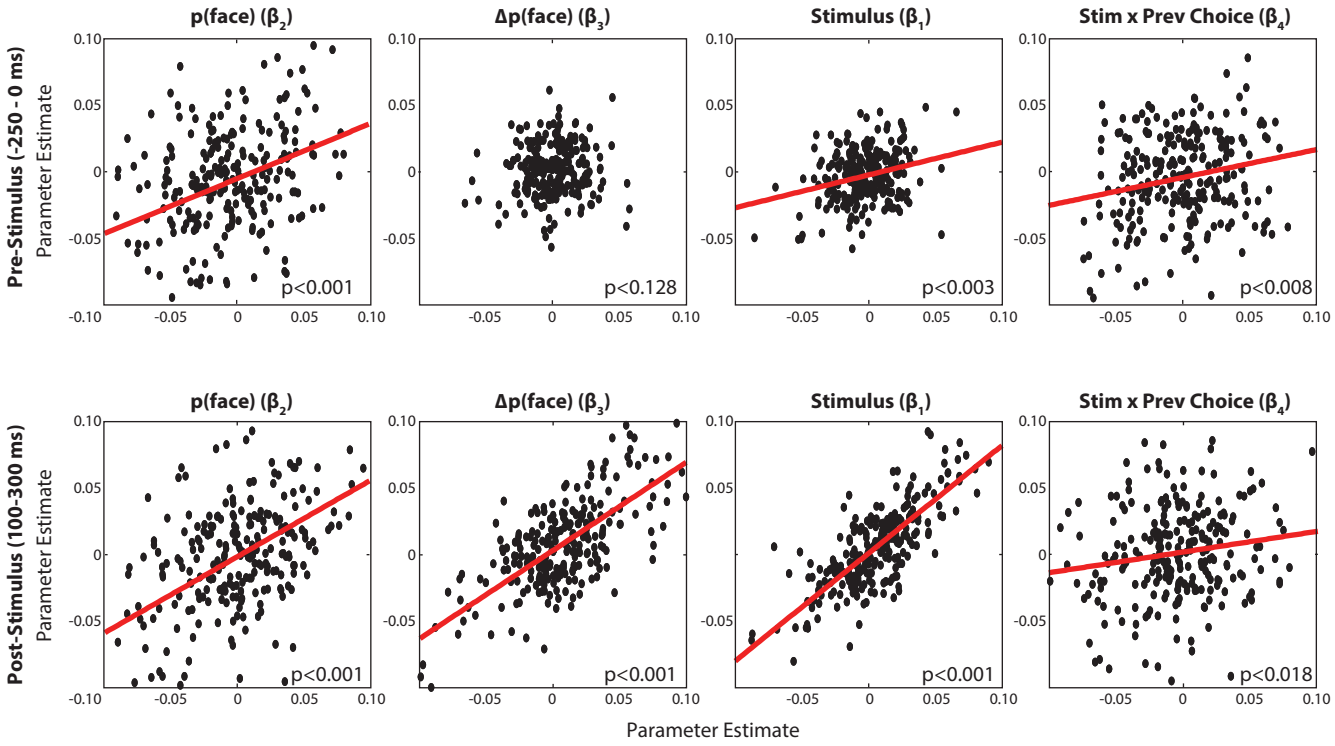

A
Choose Face (β1)
p(face)RL (β2)
p(face)RL x Choose face (β3)
Face (Correct)
Fruit (Incorrect)
Parameter Estimate
0.1
−0.1
p<0.05*
0
200
400
600
Time from Eye on Target (ms)
B
Low Noise Trials
C
High Noise Trials
Expected
Neutral
Decoding Accuracy
0.7
0.6
0.5
-400
-200
Time from
Cue Onset (ms)
Time from
Eye on Target (ms)



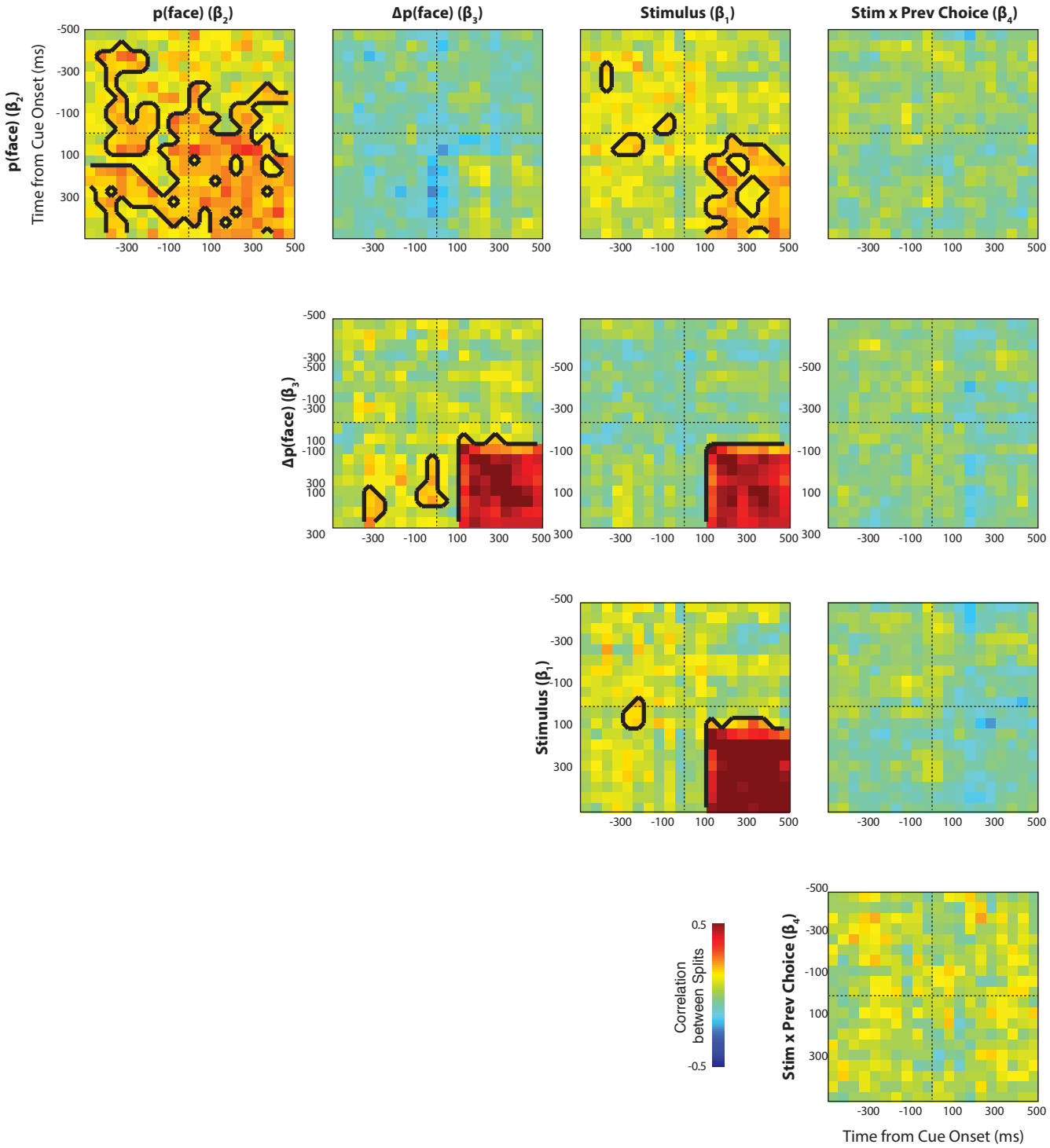
Supplemental Figure 1



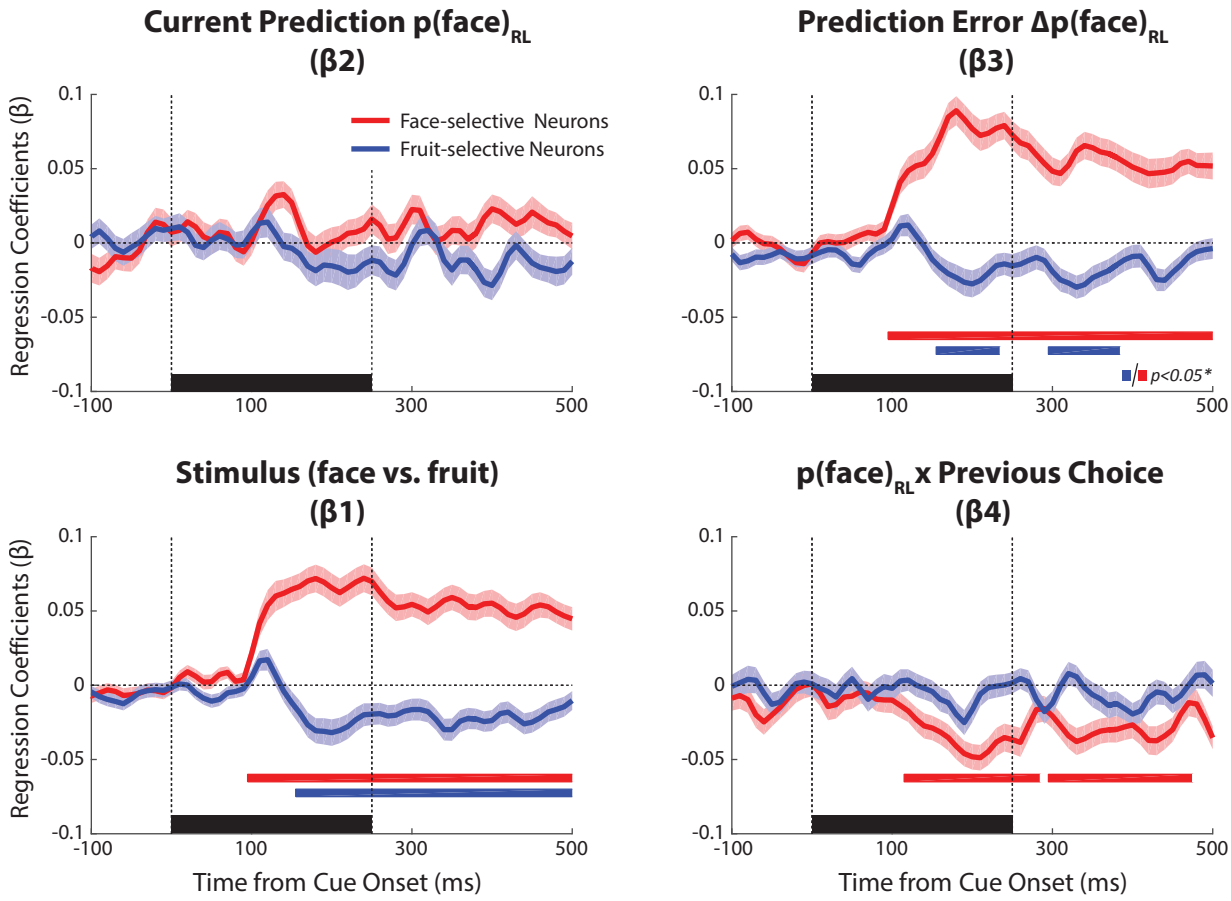
Supplemental Figure 2



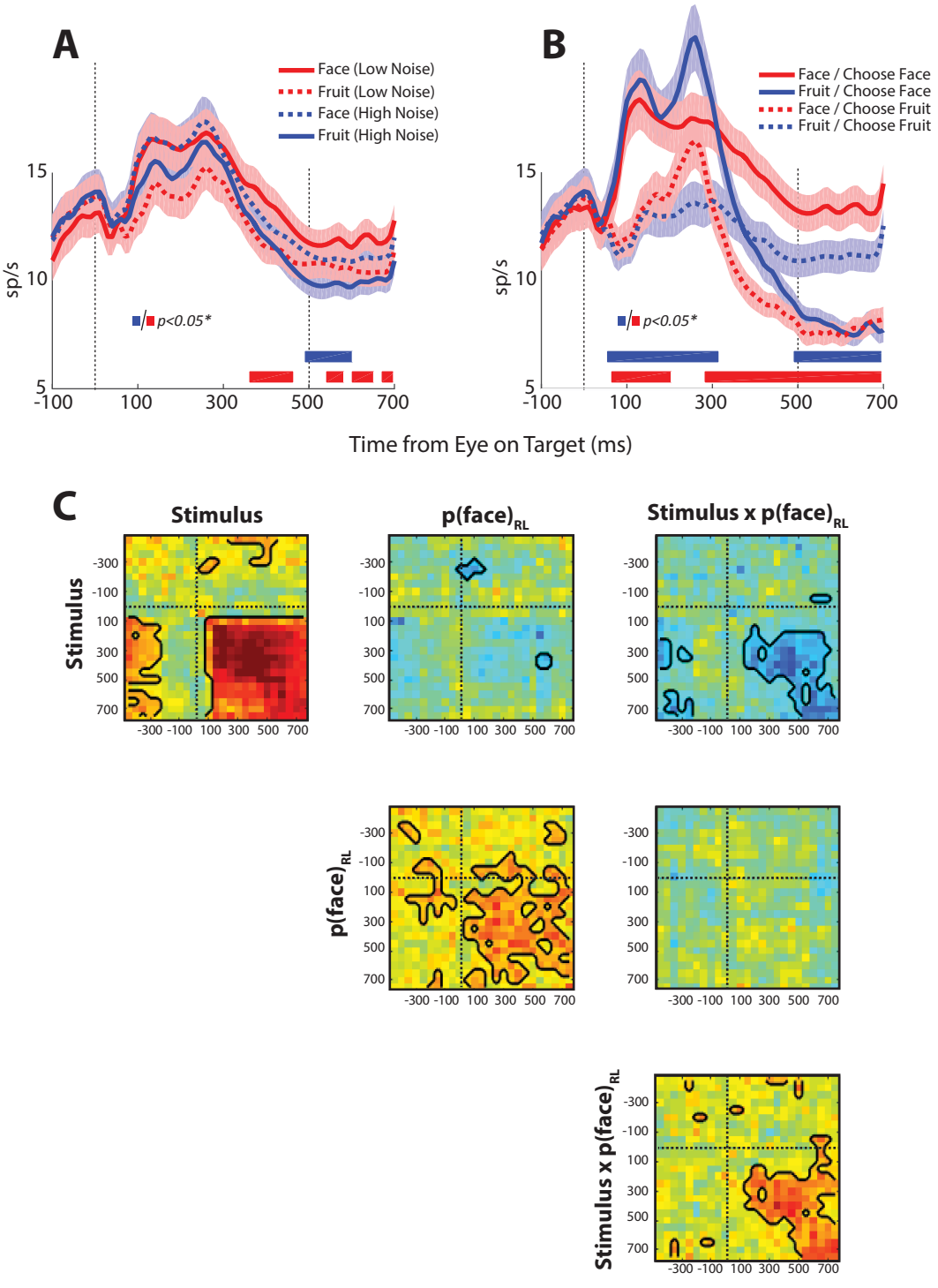
Supplemental Figure 3



Supplemental Figure 4



Supplemental Figure 5



Supplemental Figure 6

Supplemental Figures Legends

Figure S1, related to Figure 1 Behavioural performance in delayed match-to-sample task for each monkey. Left Panels: choice probabilities (for choosing face) in response to low and high noise cues. Right Panels: Hit rates and false alarms were more influenced by *p(face)* under high noise conditions as compared to low noise conditions (grey lines). SEM indicated by error bars.

Figure S2, related to Figure 3 Population **r**esponses in IT to expected, neutral, and unexpected faces, grouped according to the identity of the cue stimulus from the previous trial. Regardless of the nature of the previous trial, a robust and sustained increase in activity in response to face stimuli was observed, beginning around 100 ms following cue onset. Black bar indicates cue presentation period. Red bars indicate timepoints where average firing rates for expected and unexpected faces were significantly different from one another. (c,d) Cue-aligned regression coefficients for the interaction between stimulus category and p(face) ("prediction error", β3) for Eq.1(which includes a regressor based on previous choice) and for the following equation, which includes a regressor for the three previous cue stimuli (prevstim):

$$y = \beta_0 + \beta_1 stimulus + \beta_2 p(face)_{Bayes} + \beta_3 stimulus \times p(face)_{Bayes} + \beta_4 prevstim_1 \times stimulus + \beta_5 prevstim_2 \times stimulus + \beta_6 prevstim_3 \times stimulus + \beta_7 trial$$

The red bars show the corresponding significant timepoints for each analysis. To correct for multiple comparisons, a cluster-correction was applied across timepoints ($p<0.05$*).

Figure S3, related to Figure 4 Scatter plots of the parameter estimates for all coefficients in eq. 4, calculated from the two independent splits of the data, both in the pre-stimulus period (top row) and post-stimulus period (bottom row). Inset text shows the p-value for the corrected non-parametric test (see Experimental Procedures).

Figure S4, related to Figure 5 Cross-validation of all choice-aligned regression coefficients. Each plot shows the Pearson's correlation between regression coefficients for two independent splits of the data. Panels on the diagonal axis show correlations for the same predictor variable (e.g., β_2 and β_2; top row, leftmost panel) whereas off-diagonal panels show correlations for between predictor variables (e.g., β_2 and β_4, top row, rightmost panel). Each panel shows correlations for a given timepoint with every other timepoint. Black contour lines indicate timepoints where a significant correlation was observed ($p < 0.01$), adjusted for multiple comparisons using a cluster correction method.

Figure S5, related to Figure 5 Cue-aligned regression coefficients from eq. 4 for face and fruit. Black bar indicates cue presentation period. Red and blue traces correspond to separate regressions for different subpopulations of neurons, sorted by their preference for faces (red) and fruit (blue) based on a median split on the regression coefficients for a simple regression of stimulus category on neural activity. The red and blue bars show the corresponding significant timepoints for each analysis. To correct for multiple comparisons, a cluster-correction was applied across timepoints ($p<0.05$*).

Figure S6, related to Figure 6 Average responses and regression coefficients for IT neurons at the time of choice. Responses are aligned on the time when the monkey's eye reaches the chosen target stimulus (either an intact face or fruit). The monkey is then required to hold fixation on the target stimulus for 500 ms before a liquid reward is administered. (a) Average response, grouped according to *cue identity*. Red and blue bars indicate timepoints where average responses on high vs. low noise trials were significantly different from one another. (b) Average responses, grouped according to correct vs. incorrect trials. Solid traces indicate trials where the monkey chose "face", dashed traces are those trials where the monkey chose "fruit". Red traces indicate trials where the cue was a face, blue traces indicate trials where the cue was a fruit. Red and blue bars indicate timepoints where average responses for correct vs. incorrect trials were significantly different from one another. To correct for multiple comparisons, a cluster-correction was applied across timepoints ($p<0.05$*). (c) Cross-validation of choice-aligned regression coefficients. Each plot shows the Pearson's correlation between regression coefficients for two independent splits of the data. Panels on the diagonal axis show correlations for the same predictor variable (e.g., β_1 and β_1; top row, leftmost panel) whereas off-diagonal panels show correlations for between predictor variables (e.g., β_1 and β_2, top row, middle panel). Each panel shows correlations for a given timepoint with every other timepoint. Black contour lines indicate timepoints where a significant correlation was observed ($p < 0.01$), adjusted for multiple comparisons using a cluster correction method.

Supplemental Experimental Procedures

All procedures were approved by the National Institute of Mental Health (NIMH) Animal Care and Use Committee and conformed to all NIH guidelines. Two adult, male rhesus macaques (10-12 kg) were prepared for chronic recordings. A post for head restraint and recording cylinders were implanted under aseptic conditions. Circular recording chambers (Crist Instruments, Hagerstown MD) were centred about 12 mm anterior to the inter-aural axis over the right hemisphere in both animals.

Monkeys were trained on a delayed match-to-sample task that required them to identify which of two canonical stimuli (a face or fruit) best matched a previously presented cue (**Fig. 1a**). Each trial began with an initial fixation period of 400-800 ms, after which a cue stimulus was presented for 250 ms. The cue stimulus was either a degraded face or a fruit (face and fruit exemplars were randomly selected on a daily basis from a set of 8 from each category). Stimulus degradation was accomplished through the addition of Gaussian noise (Adobe Photoshop, Adobe Systems, San Jose, CA) at two different levels (20%, "low noise" or 80%, "high noise"). The cue was removed and after an additional delay period (randomised between 150-550 ms), intact face and fruit stimuli (where one matched the identity of the degraded cue stimulus) appeared on the right and left side of the fixation point. The monkey was required to generate a saccade to the matching stimulus and hold fixation for 500 ms before receiving a fluid reward. The next trial began after a 500-1000 ms (randomised) inter-trial interval during which time no stimulus was present and the monkeys were free to move their eyes. The cue and choice stimuli were converted to grayscale and vignetted with an oval window to remove low-level cues such as overall shape. They were approximately 5 degrees in size, and the choice stimuli were positioned about 10 degrees on either side of the fixation point. The respective side on which the canonical face/fruit stimuli appeared was pseudorandomised.

Trials were arranged into 5 blocks, (presented in random order), each with a different probability of the cue being a face vs. a fruit (0%, 25%, 50%, 75%, 100%). The monkeys were given no external cue as to which block they were currently in, nor when blocks transitioned from one probability to the next. The monkeys had to perform at least 3-5 correct trials per condition (approximately 50 trials per block).

Neuronal data were collected and processed using methods described in detail in a previous paper [S1]. Data were collected over 119 recording sessions (monkey 1: 97 sessions, 217 neurons; monkey 2: 22 sessions, 36 neurons). For welfare reasons, we were required to end data collection prematurely in monkey 2, hence the reduced number of recording sessions and neurons. However, no significant differences were observed between the data obtained from monkey 2 as compared to monkey S and so we grouped data across the two animals.

During recording sessions, between 1 and 4 electrodes were lowered into the inferior bank and lip of the superior temporal sulcus (between 5 and 19 mm anterior to the interaural axis), guided by rigid guidetubes that terminated about 10 mm above the targeted area. Waveform data was sampled at 40KHz and later sorted into individual neurons using Offline-Sorter (Plexon Systems). Spiking data was convolved with a Gaussian kernel (σ = 10 ms) to generate spike density functions for each trial.

Behaviour and modelling. Data from both monkeys were pooled as if they were a single observer. Behavioural data were analysed using analysis of variance (ANOVAs). For initial behavioural analyses, we used the objective task structure, measuring the probability of responding "face" as a function of the stimulus (face vs. fruit), the noise level (high vs. low), and the objective probability of a face, *p(face)*. Subsequently, we fit a Bayesian Learner [S2] to the data. Briefly, the Bayesian learner calculates a posterior likelihood distribution over *p(face)* and V, the rate of change of *p(face)*, following each new stimulus, marginalising over V to obtain $p(face)_{Bayes}$ (code adapted from: http://hannekedenouden.ruhosting.nl/RLtutorial/Instructions.html). The delta-rule RL model was implemented as described in the main text. The predictions of delta-rule and Bayesian Learner models were compared using maximum likelihood estimation [S3], calculating single-trial likelihoods p(data|model) by combining psychophysical data (probability that the monkey made a face response in that condition) with model-derived estimates of *p(face)*. Model probabilities and exceedance probabilities were calculated via Bayesian model selection [S4].

Neuronal data. Firing rates on each trial were averaged over bins of 10 ms. Average data were plotted for each trial type with standard errors computed over all 253 neurons. To assess neuronal selectivity, we compared baseline (-250 to 0 ms) to post-stimulus (100 to 300 ms) firing rates for low-noise stimuli only using Wilcoxon signed rank tests for each of the 253 neurons. We used this conventional approach for comparability with previous studies, but very similar results were obtained when using regression-based methods for defining selectivity (e.g., **Fig. 3f**).

Multiple regression was performed for each neuron separately to estimate beta coefficients associated with predictor variables, as described in the main text. Significance was calculated by performing t-tests on the resulting coefficients at each timepoint, using an alpha of $p < 0.05$. To correct for multiple comparisons, we used a cluster permutation approach [S5]. The trial structure within a neuron was shuffled 100 times, and the analysis was repeated (preserving all temporal aspects of the data). For each shuffle, the maximum number of adjacent significant timepoints (cluster size) was logged. Only clusters falling within the 95th percentile of the resulting distribution are reported.

Multivariate analyses were conducted by splitting the data from each neuron into two random sets of trials of equal number, and estimating the coefficients from the regression model separately for each half of the data. The two sets of coefficients were then correlated between independent splits for each timepoint. Significance was calculated by comparing nonparametric (i.e. rank-based) Fisher's Z-scores obtained from this correlation to a null distribution obtained by shuffling the data 1000 times and repeating the analysis. Timepoints where the correlation exceeded $p < 0.05$ (corrected for multiple comparisons) are marked with contours. The same cluster permutation method based on shuffled data was again used to exclude smaller clusters, except that clusters were defined by a surface connectivity criterion [S6].

Multivariate decoding. Neuronal data were first sorted so that trials from each condition (e.g., neutral face stimulus, neutral fruit stimulus) fell in an equal number of adjacent columns, with excess trials discarded. Critically, this ensured equal numbers of trials per condition, ensuring that our analyses were not biased by the larger number of trials in the "expected" condition. Neurons with less than 20 trials in a condition were excluded (~20%). Data were then randomly allocated to training (70%) and test (30%) datasets. Using probit regression, coefficients predicting face stimulus vs. fruit stimulus were estimated from the training set for each timepoint, and used to predict stimulus classes for the corresponding timepoint on the test set, according to whether the resulting predicted value exceeded 0.5. This was carried out separately for neutral trials (where $0.33 < p(face)\text{Bayes} < 0.66$) and expectation trials (where the expectation and stimulus were congruent: i.e., pooling over cases were $p(face)_{\text{Bayes}} > 0.66$ and the stimulus was face, or $p(face)_{\text{Bayes}} < 0.33$ and the stimulus was fruit. This entire process was performed separately for high noise and low noise trials. The resulting decoding accuracy was plotted over time in each condition, in both the cue and response periods.

References

[S1] Bell, A.H., Malecek, N.J., Morin, E.L., Hadj-Bouziane, F., Tootell, R.B., and Ungerleider, L.G. (2011). Relationship between functional magnetic resonance imaging-identified regions and neuronal category selectivity. J Neurosci *31*, 12229-12240.

[S2] Behrens, T.E., Woolrich, M.W., Walton, M.E., and Rushworth, M.F. (2007). Learning the value of information in an uncertain world. Nat Neurosci *10*, 1214-1221.

[S3] Daw, N. (2011). Trial-by-trial data analysis using computational models. In Decision Making, Affect, and Learning: Attention and Performance XXIII, M. Delgado, E.A. Phelps and T.W. Robbins, eds. (Oxford, UK).

[S4] Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., and Friston, K.J. (2009). Bayesian model selection for group studies. Neuroimage *46*, 1004-1017.

[S5] Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. Journal of neuroscience methods *164*, 177-190.

[S6] Kersten, D., Mamassian, P., and Yuille, A. (2004). Object perception as Bayesian inference. Annu Rev Psychol *55*, 271-304.