Behavioral/Systems/Cognitive

Expectation and Surprise Determine Neural Population Responses in the Ventral Visual Stream

Tobias Egner,[1,2] **Jim M. Monti,**[3] **and Christopher Summerfield**[4]

[1]Department of Psychology and Neuroscience, and [2]Center for Cognitive Neuroscience, Duke University, Durham, North Carolina 27708, [3]Department of Psychology, University of Illinois, Beckman Institute, Urbana-Champaign, Illinois 61801, and [4]Department of Experimental Psychology, University of Oxford, Oxford OX1 3UD, United Kingdom

Visual cortex is traditionally viewed as a hierarchy of neural feature detectors, with neural population responses being driven by bottom-up stimulus features. Conversely, "predictive coding" models propose that each stage of the visual hierarchy harbors two computationally distinct classes of processing unit: representational units that encode the conditional probability of a stimulus and provide predictions to the next lower level; and error units that encode the mismatch between predictions and bottom-up evidence, and forward prediction error to the next higher level. Predictive coding therefore suggests that neural population responses in category-selective visual regions, like the fusiform face area (FFA), reflect a summation of activity related to prediction ("face expectation") and prediction error ("face surprise"), rather than a homogenous feature detection response. We tested the rival hypotheses of the feature detection and predictive coding models by collecting functional magnetic resonance imaging data from the FFA while independently varying both stimulus features (faces vs houses) and subjects' perceptual expectations regarding those features (low vs medium vs high face expectation). The effects of stimulus and expectation factors interacted, whereby FFA activity elicited by face and house stimuli was indistinguishable under high face expectation and maximally differentiated under low face expectation. Using computational modeling, we show that these data can be explained by predictive coding but not by feature detection models, even when the latter are augmented with attentional mechanisms. Thus, population responses in the ventral visual stream appear to be determined by feature expectation and surprise rather than by stimulus features per se.

Introduction

"Predictive coding" models of visual cognition propose that perceptual inference proceeds as an iterative matching process of top-down predictions against bottom-up evidence along the visual cortical hierarchy (Mumford, 1992; Rao and Ballard, 1999; Lee and Mumford, 2003; Friston, 2005; Spratling, 2008). Specifically, each stage of the visual cortical hierarchy is thought to harbor two computationally distinct classes of processing unit: representational units that encode the conditional probability of a stimulus ("expectation") and provide predictions regarding expected inputs to the next lower level; and error units that encode the mismatch between predictions and bottom-up evidence ("surprise"), and forward this prediction error to the next higher level, where representations are adjusted to eliminate prediction error (Friston, 2005). These assumptions contrast sharply with more traditional views that cast visual neurons primarily as feature detectors (Hubel and Wiesel, 1965; Riesenhuber and Poggio, 2000), but explicit empirical tests adjudicating between these rival conceptions are lacking.

Here, we exploited the fact that the two models make divergent predictions regarding determinants of neural population responses in category-selective visual regions, like the fusiform face area (FFA) (Kanwisher et al., 1997). Predictive coding suggests that FFA population responses should reflect a summation of activity related to representational units ("face expectation") and error units ("face surprise"), whereas feature detection models suppose the population response to be driven by physical stimulus characteristics ("face features") alone. We adjudicated between these hypotheses by acquiring functional magnetic resonance imaging (fMRI) data from the FFA while independently varying both stimulus features (faces vs houses) and subjects' perceptual expectations regarding those features (low vs medium vs high face expectation) (Fig. 1*A*,*C*). Of note, both the feature detection and predictive coding views also allow for visual neural responses to be scaled by attention. Therefore, the above manipulations were orthogonal to the task demands (the detection of occasional inverted "target" stimuli) (Fig. 1*B*) to control for potential differences in attention across the conditions of interest.

According to predictive coding, FFA activity in this experiment should vary as an additive function of face expectation (high > low) (Fig. 2*A*, left) and face surprise (unexpected > expected faces) (Fig. 2*A*, middle). This would result in an interaction between stimulus and expectation factors (Fig. 2*A* right panel), whereby FFA responses to face and house stimuli should be similar under high face expectation, because both of these conditions would be associated with activity related to face ex-

Received June 1, 2010; revised Sept. 21, 2010; accepted Sept. 28, 2010.

This work was supported by funds granted by the Cognitive Neurology and Alzheimer's Disease Center (Northwestern University) to T.E. We thank Vincent De Gardelle for helpful comments on an earlier version of this manuscript.

Correspondence should be addressed to Tobias Egner, Center for Cognitive Neuroscience, Duke University, LSRC Box 90999, Durham, NC 27708. E-mail: tobias.egner@duke.edu.

DOI:10.1523/JNEUROSCI.2770-10.2010

Copyright © 2010 the authors 0270-6474/10/3016601-08$15.00/0

pectation but no activity related to face surprise. FFA responses to faces and houses should be maximally differentiated under low face expectation, because faces would here be associated with activity related to face surprise while houses would not. (Note, however, that the precise expression of this interaction depends on the relative contribution of face expectation and face surprise units to the population response.) By contrast, the feature detection model would predict a main effect of stimulus type (faces > houses), regardless of expectation conditions (Fig. 2*B*).

To preview the results, the empirical data conformed to the pattern hypothesized by the predictive coding account. We subsequently employed computational modeling to formally show that the observed data can be explained by predictive coding but not by feature detection models, even when the latter are augmented by attentional mechanisms.

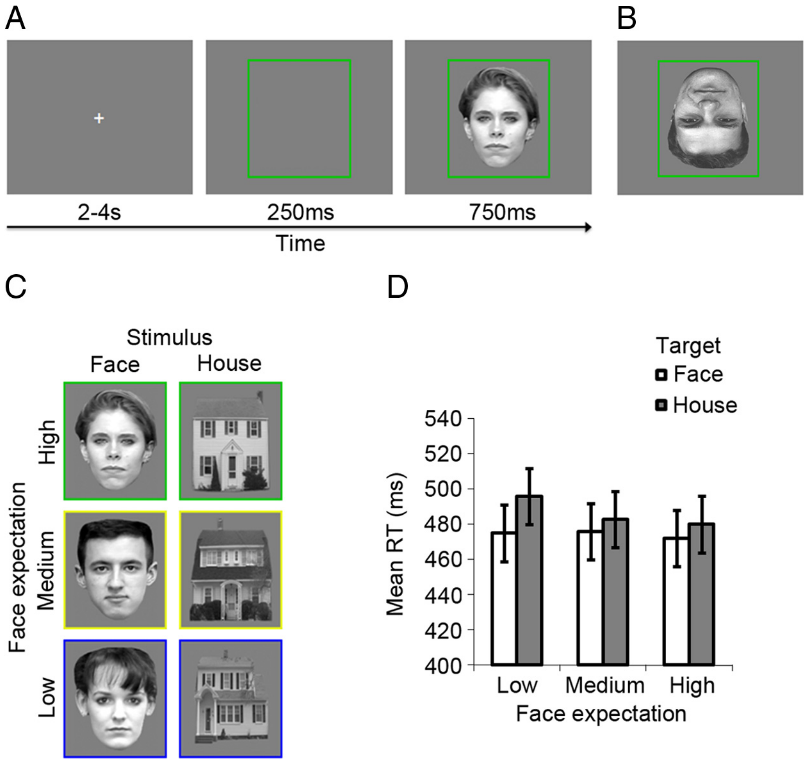

Figure 1. Experimental protocol and behavior. ***A***, Each trial commenced with an intertrial interval during which a fixation cross was presented, varying in duration from 2 to 4 s, drawn from a uniform distribution of 1 s steps (i.e., 2, 3, 4 s). Then, a colored frame (green, yellow, or blue) was presented that briefly preceded (by 250 ms) the addition of either a face or house stimulus inside that frame (for 750 ms). ***B***, It was the subjects' task to detect occasional inverted (upside-down) target stimuli, an example of which is shown here, by performing a speeded right index finger button press. Targets occurred on 10% of all trials, could be either faces or houses, were equally likely to occur in association with each frame color, and the probability of a target being an inverted face or an inverted house stimulus was equal (50%) across the different color frame conditions. ***C***, Orthogonal to task demands, the experimental manipulations of interest concerned nontarget trials, independently varying stimulus features (faces vs houses) and expectation for stimulus features, by probabilistically pairing frame color with stimulus type, with levels of 0.25, 0.50, and 0.75 (low, medium, high) probability of encountering a face stimulus (represented by blue, yellow, and green frames, respectively, in the example depicted). ***D***, Mean RTs (± SEM) for target detection, shown as a function of target type (inverted face vs inverted house) and face expectation condition.

Materials and Methods

Subjects. Sixteen healthy, right-handed volunteers (10 females, 6 males; mean age, 25.3 years; age range, 21–37 years) gave written informed consent to participate in this study, in accordance with institutional guidelines at Northwestern University. All participants had normal or corrected-to-normal vision and were screened by self-report to exclude any subjects reporting previous or current neurological or psychiatric conditions, and current psychotropic medication use. Subjects were paid $30 for participating in a 1 h MRI session.

Stimuli. We employed 300 unique black and white face images (150 females, 150 males) of neutral facial expression, and 300 unique front-view black and white images of houses (Fig. 1*A*, *C*). Each image was presented only once in the course of the experiment. Face images were drawn from the following databases: The Productive Aging Laboratory Face Database (Minear and Park, 2004); the Cohn–Kanade Facial Expression Data Base (Kanade et al., 2000); the Georgia Tech Face Database (http://www.anefian.com/research/face_reco.htm); and a collection by Endl et al. (1998). Using Photoshop (Adobe Systems), face images were cropped, resized, and displayed centrally on a uniform gray background [red-green-blue (RGB) = 128, 128, 128]. House images were culled from the internet and, like the face images, were cropped, resized, and displayed centrally on a uniform gray background (RGB = 128, 128, 128). Mean luminance values were equated across all face and house images. Each face and house image was furthermore paired with a narrow colored frame (Fig. 1*C*) that was either green (RGB = 0, 128, 0), yellow (RGB = 255, 255, 0), or blue (RGB = 0, 0, 255). When presented in the scanner (projected onto a back-projection screen at the head of the bore), face and house stimuli subtended ~10° (height) × 8° (width) of visual angle, and the colored frame outlines subtended ~12° (height) × 10° (width) of visual angle. Of the 600 stimuli, 540 were "regular" stimuli (presented in an upright position) and 60 were target stimuli (Fig. 1*B*), which were presented in an inverted position (upside down).

Experimental protocol. The goal of the experimental manipulations was to evoke perceptual expectations (and violations thereof) regarding the presentation of face and house image stimuli (see Stimuli section). This was achieved by pairing black and white face and house stimuli with colored frames (green, yellow, blue) whose colors were probabilistically predictive of the type of accompanying stimulus. On each trial, a colored frame was first shown for 250 ms by itself, and then a face or house image was added inside the frame for 750 ms, after which both stimulus components were removed from the screen and replaced by a white central fixation cross (Fig. 1*A*). The fixation cross was shown throughout the intertrial interval, whose duration was drawn randomly from a uniform distribution of 2, 3, and 4 s intervals. The experiment consisted of 600 trials, broken down into five runs of 120 trials each. The stimulus set consisted of 300 unique face images and 300 unique house images, none of which were repeated across the experiment.

It was the subjects' task to monitor the sequence of stimuli to perform a speeded button press with their right index finger whenever they spotted an occasional target stimulus. Targets comprised 10% of all stimuli (60 total, 12 per run) and consisted of inverted face and house images (Fig. 1*B*), with 50% of targets being inverted faces, and 50% being inverted houses. Importantly, this task was orthogonal to the manipulation of perceptual expectations. Specifically, the colored frames were not predictive of the type of target stimulus the subject might encounter (i.e., there were equal numbers of inverted face and house targets associated with each frame color). However, frame color was predictive of the stimulus type for the other 90% (540) of regular nontarget (upright) stimulus trials. Specifically, one frame color (e.g., green) was accompanied by face stimuli 75% of the time and by house stimuli 25% of the time (high face expectation), another frame color (e.g., yellow) was accompanied by face stimuli 50% of the time and by house stimuli 50% of the time (medium

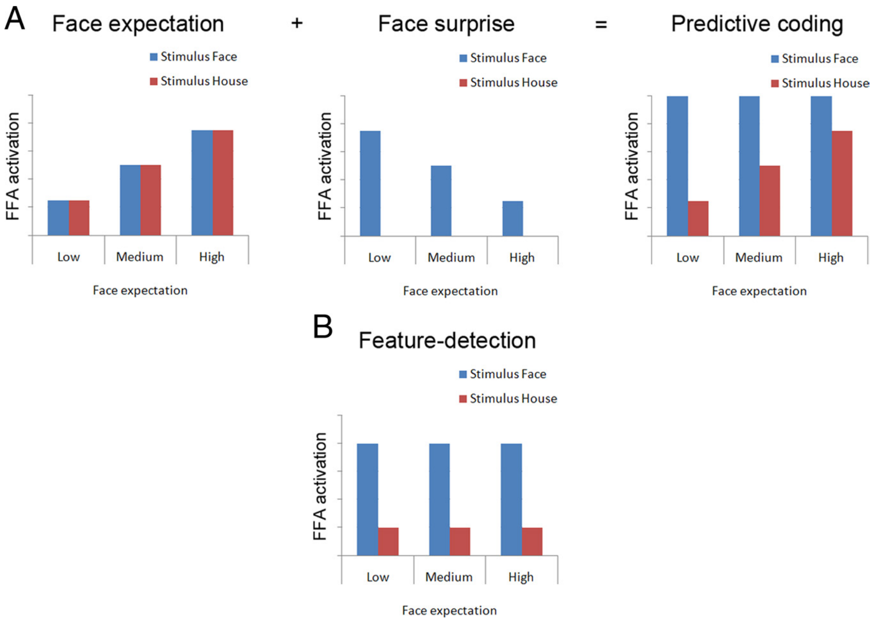

Figure 2. Predicted FFA population response patterns based on predictive coding and feature detection models. ***A***, Predictive coding argues that FFA population responses reflect the sum (right) of activity generated by representation units (face expectation, left) and error units (face surprise, middle). Note that the predicted pattern in the right-hand panel is based on the (hitherto untested) assumption that expectation and surprise contribute equally (50:50) to the FFA population response. Uneven ratios would result either in enhancing (if face surprise contributed more strongly) or attenuating (if face expectation contributed more strongly) this interaction pattern. ***B***, Feature detection views suppose that the FFA population response is driven by stimulus features, with face stimuli eliciting stronger responses than house stimuli.

face expectation), and the remaining frame color (e.g., blue) was accompanied by face stimuli 25% of the time and by house stimuli 75% of the time (low face expectation).

This arrangement resulted in a 2 (stimulus: face vs house) × 3 (face expectation: low vs medium vs high) factorial design for regular trials (Fig. 1*C*). Trial counts in the six cells of this design came to 45 (faces in the low face expectation condition, and houses in the high face expectation condition), 90 (faces and houses in the medium face expectation condition), and 135 (faces in the high face expectation condition, and houses in the low face expectation condition). Target trials were coded as events of no interest in the fMRI analysis (see Image analysis), but behavioral data from these trials were analyzed to confirm that response times (RTs) to targets were not affected by the orthogonal manipulation of perceptual expectations. Reported degrees of freedom and p values are subject to Greenhouse–Geisser correction for sphericity violations, where appropriate. Finally, before the experiment, subjects were made aware of all probabilistic contingencies in the protocol (by verbal instruction). The reasons for this were twofold. First, in this study we were not interested in distinguishing between implicit and explicit formation of perceptual expectations, and to control for individual differences in subjects noticing the contingencies we made them explicit to all subjects. Second, by explaining and emphasizing the irrelevance of the probabilistic association between frame color and stimulus category to the subjects' task, we sought to discourage the subjects upfront from forming misguided differential attentional strategies in the different expectation conditions.

Subsequent to the main task, subjects also performed a standard localizer task to define the FFA (Kanwisher et al., 1997). The localizer consisted of a 1-back task during block-wise presentation of face and house stimuli on a black background and required subjects to push a button whenever two identical stimuli were presented in a row. Face and house stimuli subtended ~10° × 8° of visual angle. Each block consisted of 15 stimuli (including 1–2 repetitions), with each stimulus presented for 750 ms followed by 250 ms fixation, and 10 s fixation periods between blocks. The task consisted of 12 blocks shown in ABAB order.

Image acquisition. Images were recorded with a Siemens Trio 3-Tesla scanner, using a 12 channel birdcage headcoil. Functional blood oxygenation level-dependent (BOLD) images were acquired parallel to the anterior commissure–posterior commissure line with a T2*-weighted echo planar imaging sequence of 38 contiguous axial slices [repetition time (TR) = 2000 ms; echo time (TE) = 20 ms; flip angle = 80°; field of view (FOV) = 220*220 mm; array size 64*64] of 3.0 mm thickness and 3.4 × 3.4 mm in-plane resolution. Structural images were acquired with a T1-weighted spoiled gradient-recalled acquisition in a steady-state sequence (TR = 19 ms; TE = 5 ms; flip angle = 20°; FOV = 220*220 mm), recording 124 slices at a slice thickness of 1.5 mm and in-plane resolution of 0.86 × 0.86 mm.

Image analysis. All preprocessing and statistical analyses were carried out using SPM5 (http://www.fil.ion.ucl.ac.uk/spm/software/spm5/). For each subject, functional data were slice time corrected and spatially aligned to the first volume of the first run. Each subject's structural scan was coregistered to a mean image of their realigned functional scans and then used to calculate transformation parameters for normalizing the functional images to the Montreal Neurological Institute template brain. The normalized functional images (resampled at 3 mm^3) were spatially smoothed with a Gaussian kernel of full width at half-maximum of 9 mm^3. The first five volumes of each run were discarded before building and estimating the statistical models of the task. A 128 s temporal high-pass filter was applied to remove low-frequency artifacts. Temporal autocorrelation in the time series data was estimated using restricted maximum likelihood estimates of variance components with a first-order autoregressive model, and the resulting nonsphericity was used to form maximum likelihood estimates of the activations.

The statistical models for the main task consisted of six regressors (vectors of stick functions) coding for the onset time and duration (1 s) of each trial in each of the experimental conditions (faces stimulus/low face expectation, face stimulus/medium face expectation, face stimulus/high face expectation, house stimulus/low face expectation, house stimulus medium face expectation, house stimulus/high face expectation), as well as a nuisance regressor coding for target trials. These models were convolved with the canonical hemodynamic response function (HRF) of SPM5 and then regressed against the observed fMRI data. For the localizer task, block regressors coding for onsets and durations of face and house blocks were convolved with the canonical HRF and regressed against the observed fMRI data. Subsequently, in each subject, we contrasted activity associated with face blocks with that associated with house blocks, and then employed the resulting contrast images in a second-level random-effects group analysis, to determine a group FFA region of interest (ROI) (Fig. 3*A*). We defined the FFA group ROI as a 6 mm diameter sphere centered on the group peak activation in the fusiform gyrus for the faces > houses localizer contrast. For a control analysis, we also defined a "parahippocampal place area" (PPA) (Epstein and Kanwisher, 1998) group ROI as a 6 mm diameter sphere centered on the group peak activation in the parahippocampal gyrus for the reverse, houses > faces localizer contrast. We then used these ROIs to extract (using Marsbar, http://marsbar.sourceforge.net/) from each subject's data estimates of the activity (mean β parameters) associated with each trial type during the main task, and entered these estimates into a 2 (stimulus: face vs house) × 3 (face expectation: low vs medium vs high) repeated-measures ANOVA. Reported degrees of freedom and p values are subject to Greenhouse–Geisser correction for sphericity violations, where appropriate.

Modeling. To formally quantify how well the predictive coding and feature detection models could account for the observed FFA fMRI data

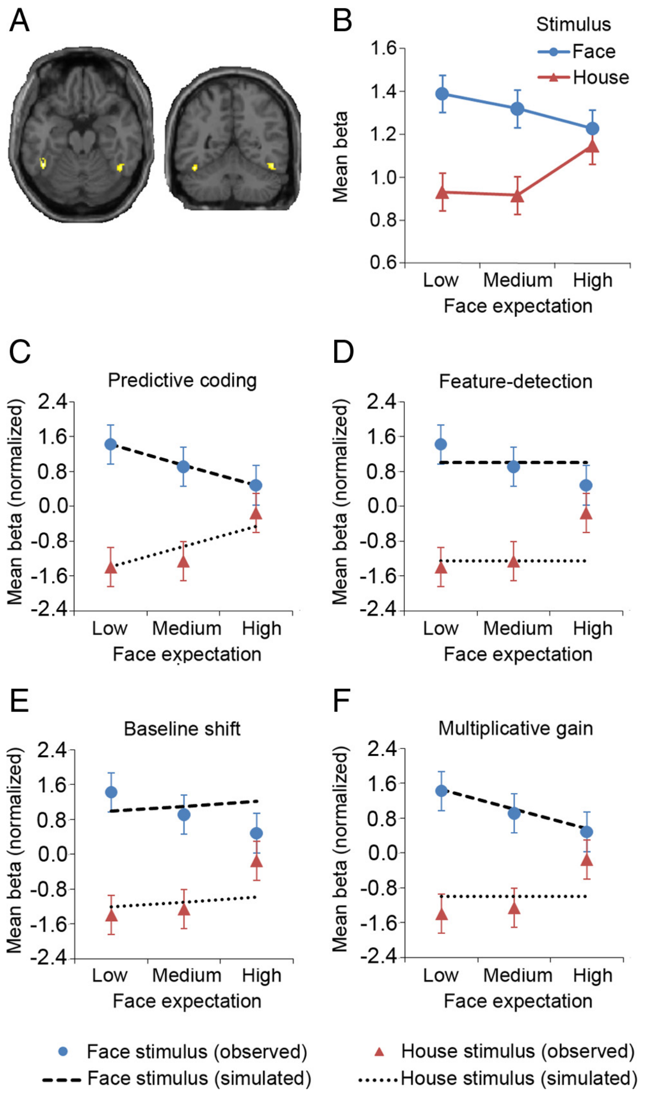

Figure 3. Functional MRI and computational modeling data. ***A***, FFA localizer group results in the fusiform gyrus are displayed on axial and coronal sections of a single-subject normalized brain. Data are shown at a false discovery rate-corrected threshold of $p < 0.05$ ($t = 3.53$) (left FFA peak: $x = -44$, $y = -50$, $z = -22$, cluster = 54 voxels; right FFA peak: $x = 48$, $y = -52$, $z = -18$, cluster = 52 voxels). ***B***, Mean group activation estimates (β parameters ± SEM) for each condition of the main experimental protocol are shown for the group FFA peak ($x = -44$, $y = -50$, $z = -22$) defined in the localizer task. ***C–F***, Observed (colored markers) and best-fit simulated (black lines) FFA BOLD responses to face and house stimuli based on a predictive coding model (***C***), a feature detection model (***D***), and feature + attention models (***E***, ***F***), where face expectation could impose either an additive baseline shift with varying levels of face expectation (***E***) or a multiplicative gain to faces on FFA responses (***F***) (see Materials and Methods).

(Fig. 3*B*), we conducted a set of computational simulations. Initially, we compared the performance of the predictive coding model, in which FFA BOLD responses are the product of a weighted sum of expectation and surprise responses, to a model in which FFA BOLD responses were driven by the face or house stimulus alone (feature detection model). Subsequently, we introduce two further models that simulate the influence of attention on FFA BOLD signals (feature + attention models, baseline shift and contrast gain variants). To eliminate heterogeneity in the global FFA response between subjects and to facilitate model fitting and comparison, we normalized each individual subject's BOLD responses to have a mean of zero.

Predictive coding model versus feature detection model. In the predictive coding model, the BOLD response is simulated as the weighted sum of expectation (R) and error (E) responses for each stimulus s (face or house) and conditional probability c of face occurrence (low, medium, and high) as follows:

$$Y = w_1 \times E_{s,c} + w_2 \times R_{s,c} \qquad -5 < w_1, w_2 < 5. \qquad (1)$$

Note that, based on the literature on prediction error neurons in other domains (Schultz et al., 1997; Schultz and Dickinson, 2000), we assume FFA prediction error units to be activated by positive prediction error (the occurrence of an unexpected face), but not by negative prediction error (the unexpected omission of a face, corresponding to an unexpected house stimulus). For both stimulus conditions, conditional probability values R were those employed in the actual experiment (i.e., 0.25, 0.5, or 0.75). For faces, surprise values E were simply $1 - R$; for houses, they were all zero, reflecting the fact that we would not expect the FFA to respond to the surprising appearance of a house. We used the true conditional probability values to simplify the model, and because subjects had been instructed about the meaning of the colored frames. However, fitting a delta rule implementation of prediction error minimization, such as a reinforcement learning model (Sutton and Barto, 1998; den Ouden et al., 2009) to the actual trial sequence employed (with learning rates of about ~0.1), and averaging the face expectation and prediction error values across all trials yielded indistinguishable values for E and R in each condition (see supplemental Fig. 1, available at www.jneurosci.org as supplemental material).

In the feature detection model, FFA activity (regardless of expectation condition c) was simulated with parameters b_1 and b_2, reflecting the FFA response to faces and to houses:

$$Y = f \times b_1 + h \times b_2 \qquad -5 < b_1, b_2 < 5, \qquad (2)$$

where f is the presence ($f = 1$) or absence ($f = 0$) of a face stimulus and h is the presence ($h = 1$) or absence ($h = 0$) of a house stimulus.

We used maximum likelihood estimation to provide a statistical comparison between the fits of these models to mean parameter estimates in the FFA. We report log-likelihoods associated with each model, and χ^2 statistics and p values associated with comparisons between models (see Results). Because we fit the models to the group means, the analyses reported here constitute fixed-effects analyses, and as such are vulnerable to the influence of outliers. To guard against this possibility, we conducted additional control analyses documenting that results were not driven by any one individual subject. Specifically, an alternative approach would be to fit the model to each individual subject and compare the model evidences at the group level. This latter approach constitutes a full random-effects analysis that is more immune to the effects of outlying subjects. However, using this approach, we were only able to demonstrate a marginal advantage for the predictive coding model over the feature detection model ($t_{(15)} = 1.31$, $p < 0.1$). Exploration of the individual model fits suggested that this was because of interindividual differences in factors that did not arbitrate between the models (such as the magnitude of the FFA response on neutral trials, which was not explicitly modeled by any of our competing models) introduced error into the fit and weakened statistical comparison. It is also possible that our single-subject FFA data were simply too noisy to support model fitting at the individual level.

However, to guard against the possibility that our data were driven by outliers, we conducted a further analysis aimed at verifying that each of our subjects contributed to the effects described. To this end, we refit the model to the mean calculated from exhaustive combinations of 15 of 16 subjects. This allowed us to determine whether the presence or absence of any one single subject was decisive for the success of the predictive coding model. In fact, each and every one of these analyses yielded a statistical advantage for the predictive coding model over the other models (all p values < 0.008). While this analysis is not entirely conclusive, it provides support for the view that our results were not driven by outlying subjects.

Predictive coding model versus feature + attention models. While it appears unlikely that attentional biasing mechanisms would mediate the observed FFA response pattern in the current experiment (see Results), for completeness we fit two further feature detection models to the observed data that included an additional attentional parameter. One model treated attentional biasing as exerting an additive baseline shift

(Luck et al., 1997; Boynton, 2009), and the other as exerting a multiplicative contrast gain (Reynolds and Chelazzi, 2004; Williford and Maunsell, 2006) on FFA responses. In the baseline shift model, FFA responses were modeled as

$$Y = f \times b_1 + (1 - f) \times b_1 + A \times p(\text{face}), \tag{3}$$

where the attentional effect (the final term) is linearly related to the probability that a face will occur, and the parameter A is fitted to the data (b and f are as above). This model allows for responses to be additionally enhanced ($A > 0$) or suppressed ($A < 0$) when faces are anticipated (Fig. 3*E*). In the contrast gain variant of this model, a similar equation is used, as follows:

$$Y = f \times b_1 + (1 - f) \times b_1 + f \times A \times p(\text{face}). \tag{4}$$

This model differs only in that attentional modulation is limited to face-present trials, as if face anticipation will increase sensitivity to faces in a multiplicative (rather than an additive) fashion (Fig. 3*F*). To limit all models to two free parameters and thereby simplify model comparison, both feature + attention models fit the differential FFA response to faces versus houses as a single parameter. Goodness-of-fit statistics were calculated for these models in the same way as for the predictive coding and feature detection models described above.

Results

Behavioral data

The goal of the main task was to independently manipulate stimulus features (faces vs houses) and subjects' perceptual expectations regarding those features, while keeping attentional demands constant across conditions. On each trial, subjects viewed a briefly presented face or house stimulus (750 ms) that was preceded (by 250 ms) and accompanied by a colored frame (Fig. 1*A*). The frame color (green, yellow, or blue) was indicative of the likelihood that the stimulus presented would be a face (with face stimulus probability at 0.25, 0.5, and 0.75), resulting in a 2 (stimulus: face vs house) × 3 (face expectation: low vs medium vs high) factorial design (Fig. 1*C*). Importantly, the task requirements were orthogonal to these manipulations, as subjects were asked to monitor the stimulus stream for occasional target trials (Fig. 1*B*), consisting of inverted face or house stimuli (10% of all trials, with equal numbers of face and house targets), to which a speeded right index finger response was required. Inverted face and house targets were equally likely to occur in any of the six experimental conditions. Thus, the stimulus frame color was predictive neither of the likelihood of a target occurring nor of whether the target would be a face or house stimulus, such that subjects could not gain any performance benefit from using the frames to guide attentional processes. To further discourage the deployment of such strategies, we also explicitly explained the experimental contingencies to the subjects before scanning (see Materials and Methods).

To ascertain that deployment of attention was not affected by the perceptual expectations induced by the colored frames, we analyzed target trial RT data as a function of stimulus type and expectation (Fig. 1*D*), in a 2 (target stimulus: inverted face vs inverted house) × 3 (face expectation: low vs medium vs high) repeated-measures ANOVA. A main effect of stimulus type ($F_{(1,15)} = 9.3$, $p < 0.01$) revealed that subjects were faster at identifying inverted faces than inverted houses as target stimuli. Crucially, however, this effect did not interact with expectation condition ($F_{(2,30)} = 2.3, p > 0.1$), and we observed no main effect of expectation either ($F_{(2,30)} = 2.4, p > 0.1$). Thus, in line with the fact that the manipulation of perceptual expectation was irrelevant to the subjects' task, their attention toward the stimuli, and toward particular stimulus (face or house) features, did not appear to vary across the different expectation conditions (Fig. 1*D*). Response accuracy in this task was at ceiling (<1% missed targets), and error rates were therefore not subjected to inferential statistics.

fMRI data

Based on the independent localizer task (see Materials and Methods), we defined a group FFA ROI, given by the peak activation in the fusiform gyrus for the face > house blocks contrast (Fig. 3*A*). Subsequently, we extracted activation estimates (mean β parameters) from this ROI for each experimental condition in the main task from each subject, and entered these estimates into a 2 (stimulus: face versus house) × 3 (face expectation: low vs medium vs high) repeated-measures ANOVA. According to predictive coding, FFA activity should display an interaction between stimulus and expectation factors (Fig. 2*A*), with FFA responses to faces and houses being most similar under high face expectation, since both of these conditions are associated with face expectation and neither is associated with face surprise, and they are maximally differentiated under low face expectation, since here faces are associated with face surprise while houses are not. (Note that the a priori predictions displayed in Fig. 2*A* assume an equal contribution of expectation and surprise units to the FFA population response.) By contrast, the feature detection model predicts a main effect of stimulus (faces > houses), regardless of expectation condition (Fig. 2*B*).

The empirical data are shown in Figure 3*B*. FFA BOLD responses displayed a main effect of stimulus features, as face stimuli, on average, elicited higher activation than house stimuli ($F_{(1,15)} = 9.4, p < 0.01$), but there was no main effect of expectation ($F_{(1,15)} = 0.5, p > 0.1$). Crucially, the stimulus and expectation factors interacted ($F_{(2,30)} = 3.6, p < 0.05$), as the strength of the stimulus feature effect varied with expectation conditions (Fig. 3*B*). Specifically, faces elicited higher FFA responses than houses in the low face expectation condition ($t_{(15)} = 4.1$, $p < 0.001$) and the medium face expectation condition ($t_{(15)} = 4.2$, $p < 0.001$), but not in the high face expectation condition where FFA responses to face and house stimuli were statistically indistinguishable ($t_{(15)} = 0.4$, $p > 0.1$) (Bonferroni-corrected $\alpha = 0.016$). Polynomial contrast analyses showed that the stimulus × expectation interaction displayed a significant linear effect ($F_{(1,15)} = 5.0$, $p < 0.05$), confirming the impression that the differential FFA activation to face versus house stimuli increased linearly across the three levels of the expectation factor, from high to medium to low face expectation (Fig. 3*B*). These results are clearly incommensurate with a feature detection account of the FFA population response (compare Fig. 2*B*), but they are compatible with a predictive coding account that assumes the population response to reflect a summation of face expectation and face surprise responses (compare Fig. 2*A*, right). An equivalent pattern of results found in the analysis based on the group peak FFA ROI above was also observed when employing mean activation values across the entire group FFA ROI, as well as when defining FFA ROIs individually for each subject (data not shown).

As a test of the generality of these results, we exploited the symmetry of our experimental design, which allowed us to conduct the equivalent analyses on data extracted from the PPA, a visual region thought to be selective for scene and place stimuli (Epstein and Kanwisher, 1998). If predictive coding constituted the general coding strategy in visual cortex, one would expect the PPA to display the inverse pattern of the responses found in the FFA. As can be seen in supplemental Figure 2 (available at www.jneurosci.org as supplemental material), this was indeed the case.

Specifically, PPA responses displayed a main effect of stimulus (houses > faces, $F_{(1,15)} = 81.0$, $p < 0.001$), no main effect of expectation ($F_{(2,30)} = 0.1$, $p > 0.1$), and a marginal stimulus × expectation interaction ($F_{(2,30)} = 3.1$, $p = 0.058$), due to larger differences in the response to house versus face stimuli under low expectation for houses (high face expectation) than under high expectation for houses (low face expectation). These results again conform more closely to the predictions of the predictive coding model than to those of the feature detection model.

Modeling data

To formally quantify and compare the ability of predictive coding and feature detection models to account for the observed fMRI data, we conducted a set of computational simulations on the FFA data (see Materials and Methods). Specifically, we first contrasted model performance between a predictive coding model where free parameters weighted hypothetical FFA expectation and surprise unit responses, and a feature detection model where free parameters reflected the FFA responses to face and house stimuli. Simulated values from the best-fitting variants of these models are shown in Figure 3, *C* and *D*, respectively, in relation to observed (mean corrected) BOLD data from the FFA. The comparison between log-likelihoods for the best-fitting variants of the predictive coding model (−3.37) and the feature detection model (−8.62) revealed that the former provided a reliably better fit to the FFA data ($\chi^2 = 10.51$, $p < 0.002$).

Prima facie, it appears highly unlikely that the pattern of effects we observed in the FFA is due to attentional effects. First, the task performed by the subjects was orthogonal to the expectation and stimulus feature manipulations. Second, RTs neither differed between expectation conditions nor varied with the interaction between stimulus and expectation factors (Fig. 1*D*). Finally, neural models of attention posit an enhancement of neural responses to anticipated or relevant stimulus features (Summerfield and Egner, 2009), whereas in the current dataset an increase in the probability of face occurrence was actually associated with less activity in the FFA (Fig. 3*B*). Nevertheless, we fit two further feature + attention models to the data (see Materials and Methods). One of the models embodied the possibility of attention exerting an additive baseline shift effect (Luck et al., 1997; Kastner et al., 1999; Boynton, 2009) on FFA BOLD responses, as a function of face expectation (Fig. 3*E*). The other model applied the notion of attention modulating stimulus-evoked responses via a multiplicative gain mechanism (Williford and Maunsell, 2006), as espoused by contrast gain models of attention (Reynolds and Chelazzi, 2004) (Fig. 3*F*). In either case, the predictive coding model provided a closer fit to the observed data (baseline shift model: $\chi^2 = 7.0$, $p < 0.009$; contrast gain model: $\chi^2 = 10.6$, $p < 0.002$).

It should also be noted that the best-fit parameter values for the multiplicative gain feature + attention model follows the opposite pattern from that which might be expected from a feature-based attention account (Treue and Martínez Trujillo, 1999), with anticipated faces leading to reduced rather than increased FFA activity. This is because we observed face surprise, not face expectation, to contribute most robustly to FFA BOLD responses. Specifically, the ratio of best-fitting representation versus error unit weighting parameters for the predictive coding model was 1:2, suggesting that in the current experiment face surprise contributed about twice as strongly to the FFA BOLD response as face expectation.

Finally, at a reviewer's suggestion, we tested a fifth account for our data, which draws upon elements of both the predictive coding and the feature detection models. Under this hybrid account, FFA responses are the weighted sum of (1) a differential response to faces and houses and (2) a differential response to expected and unexpected stimuli (reflecting a generic surprise signal). In other words, the FFA responds preferentially to face features but also displays an additional gain in activation for any surprising event, be it a surprising face or house. This feature + surprise model was implemented as follows:

$$Y = 1[f \times b_1 + (1 - f) - b_1] + [(1 - f) \times b2 \times p(\text{face}) + f \times b_2 \times (1 - p(\text{face})], \quad (5)$$

where parameter b_1 encodes the relative response to faces and houses, and b_2 encodes the extra response associated with a surprising event. This hybrid model can explain the data in a manner that is quantitatively similar to the predictive coding model (log-likelihood, −3.38; difference with predictive coding model, n.s.) (see supplemental Fig. 3*A*, available at www.jneurosci.org as supplemental material). Therefore, we conducted an additional analysis to distinguish which of these models ultimately offers the more likely account of our data. Specifically, the feature + surprise model requires that there is a generic boost to FFA responses whenever a surprising stimulus occurs. It follows that the FFA response to unexpected faces and to unexpected houses would be positively correlated across the subject cohort. The predictive coding model, however, makes a different prediction. It assumes that the response to unexpected faces is driven by face surprise (error units) and the response to unexpected houses is driven by face expectation (representational units). Given that we have mean-corrected FFA responses, one would expect these two factors to have a reciprocal influence on FFA responses, that is, for the response to unexpected faces and unexpected houses to be negatively correlated. Empirically, the latter prediction was confirmed, as FFA responses to surprising faces and houses were strongly negatively correlated ($r = -0.72$, $p < 0.002$) (supplemental Fig. 3*B*, available at www.jneurosci.org as supplemental material), thus arguing in favor of the predictive coding account.

Discussion

We adjudicated between two views of the nature of visual cognition in the posterior brain. On the one hand, a traditional feature detection view envisions visual neurons as specialized bottom-up feature detectors whose response varies as a function of how well present stimulus features match the detector's preferred feature (Hubel and Wiesel, 1965; Riesenhuber and Poggio, 2000). On the other hand, a heavily top-down predictive coding view casts visual cognition as a predictive matching process that entails two distinct neurocomputational component processes, the encoding (and top-down propagation) of conditional probabilities of stimulus features (predictions) and the encoding (and bottom-up propagation) of mismatches between predictions and bottom-up input (prediction error) (Mumford, 1992; Rao and Ballard, 1999; Lee and Mumford, 2003; Friston, 2005; Spratling, 2008). While the former model asserts that neural population responses in a category-selective visual area, like the FFA, should be driven by stimulus features per se (Fig. 2*B*), the latter model suggests the population response to be an additive function of feature expectation and expectation violation (surprise) (Fig. 2*A*). We tested how well these rival hypotheses could account for FFA neural population responses, as measured by BOLD fMRI, in relation to independently varied stimulus features and expectations. FFA activity displayed an interaction of stimulus feature and expectation factors, where the differentiation between FFA responses to face and house stimuli decreased linearly with increasing levels of face expectation, with face- and house-evoked signals being indistinguishable under high face expectation (Fig. 3*B*).

These results are clear-cut, in that the observed interaction effect of stimulus feature and expectation factors on FFA responses was hypothesized on the basis of the predictive coding model but is incompatible with predictions of the feature detection model. In addition to showing that the FFA response pattern qualitatively conforms to the predictive coding account but not the feature detection account, we performed a set of computational model simulations to provide a formal quantitative comparison of how well the two models could account for the observed BOLD data. The predictive coding model clearly outperformed the feature detection model (Fig. 3*C*,*D*). This is because the predictive coding model assumes the population response to reflect a weighted sum of face expectation and face surprise responses, a conception that can naturally accommodate the observed interaction between stimulus features and feature expectation, including the highly counterintuitive finding that face and house stimuli elicited identical FFA population responses under conditions of high face expectation. By contrast, the feature detection model assumes FFA population activity to reflect a differential response to face features over nonface features, and it therefore cannot explain an interaction of stimulus features with expectation.

However, it could be argued that the above characterization of the feature detection view of visual neurons is impoverished, as this view traditionally also allows for bottom-up, feature-driven responses to be modulated by attention. A natural question to ask of the current results, therefore, is whether they may reflect attention-modulated feature detection responses rather than a sum of feature expectation and surprise responses, in particular because expectations are typically assumed to direct attention (Posner et al., 1980). We are confident that this is not the case, for four reasons. First, the manipulation of perceptual expectations was, by design (and known to the subjects), irrelevant to the task, such that no benefits could be derived from attending to expected (or unexpected) stimulus features. Second, in line with the experimental manipulations, the behavioral data showed that attention toward house- or face- related stimulus features neither varied across expectation conditions nor exhibited a stimulus by expectation interaction. Third, we nevertheless entertained the possibility of attention-modulated feature detection and formally tested two feature + attention models, one incorporating an additive baseline-shift parameter (Luck et al., 1997; Kastner et al., 1999; Boynton, 2009), and the other incorporating a multiplicative gain parameter (Reynolds and Chelazzi, 2004; Williford and Maunsell, 2006) into the feature detection model. The predictive coding model provided a better fit of the observed data than either of these augmented models. Fourth, the best-fit version of the feature + attention models actually required attention to modulate FFA responses in the opposite direction from what one would customarily expect (Fig. 3*F*), in that the modulation led to smaller FFA responses under high face expectation than under low face expectation. While the current results are thus unlikely to be confounded by attentional influences, the general question of how predictive coding schemes may relate to attention is the subject of ongoing debate (Rao and Ballard, 2005; Spratling, 2008; Friston, 2009) and, in our view, represents a highly interesting question for future research (Summerfield and Egner, 2009).

Another possible interpretation of the FFA data pattern we observed might be that these responses combine feature detection (faces > houses) with a surprise signal that elevates responses both to unexpected face and unexpected house stimuli. We found that such a feature + surprise model can, in principle, account for the data to a similar degree as the predictive coding model. However, additional analyses suggested this account to be a less likely explanation of the empirical data than predictive coding. Specifically, FFA responses to surprising face and house stimuli were found to be negatively correlated across individuals, which argues strongly against a generic surprise signal but is commensurate with the predictive coding account, where FFA responses to surprising faces and houses are mediated by distinct processing units. It could also be argued that the feature + surprise model is a conceptually less parsimonious account, since it would require the FFA to harbor feature detectors that are feature selective, but prediction error units that produce a surprise response that is not (or is much less) feature selective, to produce the equally large surprise responses to face and house stimuli that characterize our data (compare Fig. 3*B*).

The current results represent a key advance in the investigation of predictive coding processes in visual cognition. A number of recent studies have employed a variety of elegant manipulations to demonstrate that some portion of population activity in visual cortices can be attributed to prediction error signals, in general support of predictive coding models (Murray et al., 2002, 2003; Summerfield et al., 2006, 2008; Summerfield and Koechlin, 2008; den Ouden et al., 2009, 2010; Alink et al., 2010). However, the current study is to our knowledge the first investigation to formally and explicitly demonstrate that population responses in visual cortex are in fact better characterized as a sum of feature expectation and surprise responses than by bottom-up feature detection (with or without attention), thus providing solid support for a crucial neurocomputational tenet of predictive coding (Mumford, 1992; Rao and Ballard, 1999; Lee and Mumford, 2003; Friston, 2005; Spratling, 2008). An interesting additional finding from our computational simulations was the nature of the best-fit expectation-to-surprise ratio (i.e., the best weighting parameter) for the predictive coding model, which indicated that FFA population activity in the current experiment was best accounted for by the assumption that surprise (prediction error) processing units contributed approximately twice as strongly to the BOLD response as expectation (representational or prediction) processing units. Theoretical accounts of predictive coding have been largely agnostic about the relative prevalence and/or metabolic demands of prediction versus error processing units, so the present data represent an intriguing explanandum for these models and future studies. One complication with interpreting these data though is that it is not currently known whether the BOLD signal itself may index top-down versus bottom-up inputs to a brain region to different degrees. Furthermore, if the coding of prediction and/or prediction error were to interact with attention, then the metabolic prediction/error demand ratio would likely vary with task demands, in that it would differ depending on whether (un)expected stimulus features were relevant (attended) or irrelevant (unattended) to the task.

In sum, the current data strongly support predictive coding models of visual cognition and add further credence to the emerging notion that the encoding of predictions (based on internal forward models) and prediction errors may be a ubiquitous feature of cognition in the brain (Schultz and Dickinson, 2000; Bubic et al., 2010; Friston, 2010) rather than a curiosity of reward learning (Schultz et al., 1997) or motor planning (Wolpert and Kawato, 1998). A crucial question for future research in this area is how perceptual prediction and error processing are embodied at the cellular level; for instance, whether representational and error units map onto functionally distinct groups of neurons (Summerfield and Egner, 2009).

References

Alink A, Schwiedrzik CM, Kohler A, Singer W, Muckli L (2010) Stimulus predictability reduces responses in primary visual cortex. J Neurosci 30:2960–2966.

Boynton GM (2009) A framework for describing the effects of attention on visual responses. Vision Res 49:1129–1143.

Bubic A, von Cramon DY, Schubotz RI (2010) Prediction, cognition and the brain. Front Hum Neurosci 4:25.

den Ouden HE, Friston KJ, Daw ND, McIntosh AR, Stephan KE (2009) A dual role for prediction error in associative learning. Cereb Cortex 19: 1175–1185.

den Ouden HE, Daunizeau J, Roiser J, Friston KJ, Stephan KE (2010) Striatal prediction error modulates cortical coupling. J Neurosci 30:3210–3219.

Endl W, Walla P, Lindinger G, Lalouschek W, Barth FG, Deecke L, Lang W (1998) Early cortical activation indicates preparation for retrieval of memory for faces: an event-related potential study. Neurosci Lett 240:58–60.

Epstein R, Kanwisher N (1998) A cortical representation of the local visual environment. Nature 392:598–601.

Friston K (2005) A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci 360:815–836.

Friston K (2009) The free-energy principle: a rough guide to the brain? Trends Cogn Sci 13:293–301.

Friston K (2010) The free-energy principle: a unified brain theory? Nat Rev Neurosci 11:127–138.

Hubel DH, Wiesel TN (1965) Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. J Neurophysiol 28: 229–289.

Kanade T, Cohn JF, Tian Y (2000) Comprehensive database for facial expression analysis. Proceedings of the Fourth IEEE International Conference of Automatic Face and Gesture Recognition, pp 46–53. IEEE: Washington, DC.

Kanwisher N, McDermott J, Chun MM (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception. J Neurosci 17:4302–4311.

Kastner S, Pinsk MA, De Weerd P, Desimone R, Ungerleider LG (1999) Increased activity in human visual cortex during directed attention in the absence of visual stimulation. Neuron 22:751–761.

Lee TS, Mumford D (2003) Hierarchical Bayesian inference in the visual cortex. J Opt Soc Am A Opt Image Sci Vis 20:1434–1448.

Luck SJ, Chelazzi L, Hillyard SA, Desimone R (1997) Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. J Neurophysiol 77:24–42.

Minear M, Park DC (2004) A lifespan database of adult facial stimuli. Behav Res Methods Instrum Comput 36:630–633.

Mumford D (1992) On the computational architecture of the neocortex. II. The role of cortico-cortical loops. Biol Cybern 66:241–251.

Murray SO, Kersten D, Olshausen BA, Schrater P, Woods DL (2002) Shape perception reduces activity in human primary visual cortex. Proc Natl Acad Sci U S A 99:15164–15169.

Murray SO, Olshausen BA, Woods DL (2003) Processing shape, motion and htree-dimensional shape-from-motion in the human cortex. Cereb Cortex 13:508–516.

Posner MI, Snyder CR, Davidson BJ (1980) Attention and the detection of signals. J Exp Psychol 109:160–174.

Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2:79–87.

Rao RP, Ballard DH (2005) Probabilistic models of attention based on iconic representations and predictive coding. In: Neurobiology of attention (Itti L, Rees G, Tsotsos JK, eds), pp 553–561. New York: Elsevier Academic.

Reynolds JH, Chelazzi L (2004) Attentional modulation of visual processing. Annu Rev Neurosci 27:611–647.

Riesenhuber M, Poggio T (2000) Models of object recognition. Nat Neurosci [3 Suppl]:1199–1204.

Schultz W, Dickinson A (2000) Neuronal coding of prediction errors. Annu Rev Neurosci 23:473–500.

Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275:1593–1599.

Spratling MW (2008) Predictive coding as a model of biased competition in visual attention. Vision Res 48:1391–1408.

Summerfield C, Egner T (2009) Expectation (and attention) in visual cognition. Trends Cogn Sci 13:403–409.

Summerfield C, Koechlin E (2008) A neural representation of prior information during perceptual inference. Neuron 59:336–347.

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

Summerfield C, Trittschuh EH, Monti JM, Mesulam MM, Egner T (2008) Neural repetition suppression reflects fulfilled perceptual expectations. Nat Neurosci 11:1004–1006.

Sutton R, Barto A (1998) Reinforcement learning. Cambridge, MA: MIT.

Treue S, Martínez Trujillo JC (1999) Feature-based attention influences motion processing gain in macaque visual cortex. Nature 399:575–579.

Williford T, Maunsell JH (2006) Effects of spatial attention on contrast response functions in macaque area V4. J Neurophysiol 96:40–54.

Wolpert DM, Kawato M (1998) Multiple paired forward and inverse models for motor control. Neural Netw 11:1317–1329.