Behavioral Neuroscience
2012, Vol. 126, No. 4, 575–581

0735-7044/12/$12.00 DOI: 10.1037/a0029029

BRIEF COMMUNICATION

Explaining Neural Signals in Human Visual Cortex With an Associative Learning Model

Jiefeng Jiang, Nestor Schmajuk, and Tobias Egner
Duke University

"Predictive coding" models posit a key role for associative learning in visual cognition, viewing perceptual inference as a process of matching (learned) top-down predictions (or expectations) against bottom-up sensory evidence. At the neural level, these models propose that each region along the visual processing hierarchy entails one set of processing units encoding predictions of bottom-up input, and another set computing mismatches (prediction error or surprise) between predictions and evidence. This contrasts with traditional views of visual neurons operating purely as bottom-up feature detectors. In support of the predictive coding hypothesis, a recent human neuroimaging study (Egner, Monti, & Summerfield, 2010) showed that neural population responses to expected and unexpected face and house stimuli in the "fusiform face area" (FFA) could be well-described as a summation of hypothetical face-expectation and -surprise signals, but not by feature detector responses. Here, we used computer simulations to test whether these imaging data could be formally explained within the broader framework of a mathematical neural network model of associative learning (Schmajuk, Gray, & Lam, 1996). Results show that FFA responses could be fit very closely by model variables coding for conditional predictions (and their violations) of stimuli that unconditionally activate the FFA. These data document that neural population signals in the ventral visual stream that deviate from classic feature detection responses can formally be explained by associative prediction and surprise signals.

Keywords: associative learning, predictive coding, visual cortex, fMRI, attention

Building on Helmholtz' pivotal insight that visual cognition necessitates a contextually informed "unconscious inference" regarding the most likely explanation for a given percept (Helmholtz, 1876), a number of modern-day models have proposed a central role for learned associations in actively informing the interpretation of visual signals (Grossberg, 1980). Specifically, "predictive coding" models assert that perceptual inference proceeds as an iterative matching process of top-down predictions against bottom-up evidence along the visual cortical hierarchy (Friston, 2005; Lee & Mumford, 2003; Mumford, 1992; Rao & Ballard, 1999; Spratling, 2008). To implement this matching process, each visual processing stage is thought to harbor two computationally distinct classes of neural processors. First, representational units encode the conditional probability of a stimulus ("expectation") and provide predictions regarding expected inputs to the next lower level. Second, prediction error units encode the mismatch between predictions and bottom-up evidence ("surprise") and forward this error to the next higher level, where representations are adjusted accordingly (Friston, 2005, 2010).

Thus, predictive coding models suggest that neural processing in visual cortex is driven largely by top-down information derived from associative learning, a position that contrasts with traditional views of visual neurons acting solely as bottom-up feature-detectors (e.g., Hubel & Wiesel, 1965; Riesenhuber & Poggio, 2000). In support of the central tenet of the predictive coding hypothesis, a number of recent functional MRI (fMRI) studies in human participants have shown that neural population responses in visual cortex are indeed susceptible to manipulations of expectation and surprise (Alink, Schwiedrzik, Kohler, Singer, & Muckli, 2010; den Ouden, Daunizeau, Roiser, Friston, & Stephan, 2010; den Ouden, Friston, Daw, McIntosh, & Stephan, 2009; Egner et al., 2010; Summerfield et al., 2006; Summerfield & Koechlin, 2008; Summerfield, Trittschuh, Monti, Mesulam, & Egner, 2008).

In one study that directly pitted the traditional feature-detector view against the predictive coding perspective, Egner and colleagues (Egner et al., 2010) acquired fMRI data from the "fusiform face area" (FFA), a region of the ventral visual stream that specializes in face processing (Kanwisher, McDermott, & Chun, 1997), while independently varying physical stimulus features (faces vs. houses) and participants' perceptual expectations regarding those features (low vs. medium vs. high face expectation) by means of probabilistic cues. At the same time, the study attempted to control for differential allocation of attention across these conditions by occupying participants with an incidental task. According to predictive coding, FFA activity should reflect the summation of face expectation (high > low) and face surprise

Jiefeng Jiang and Tobias Egner, Department of Psychology & Neuroscience and Center for Cognitive Neuroscience, Duke University; Nestor Schmajuk, Department of Psychology & Neuroscience, Duke University.

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

(unexpected > expected faces). This would result in an interaction between stimulus and expectation factors 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 expectation 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. By contrast, the feature-detection model would predict only a main effect of stimulus type (faces > houses).

The neural data displayed a stimulus by expectation interaction effect (Figure 2a), a pattern of results that, qualitatively, matches the hypotheses of the predictive coding account. However, even though this FFA data pattern provides a descriptive match to the type of result anticipated on the basis of the predictive coding hypothesis, it is not certain whether these data could be explained quantitatively by the assumed underlying mechanisms of learned cue-stimulus associations (and violations thereof). Here, we applied such a stringent, quantitative test of the predictive coding account: If the predictive coding view were accurate, it should be possible to explain the neural FFA responses via formal associative learning variables derived from trial-by-trial estimation of cue-stimulus probability distributions in individual participants, analogous to the way that neural responses in the striatum can be modeled by reinforcement learning variables encoding reward prediction and reward prediction error signals (O'Doherty et al., 2004). Thus, we here tested whether the FFA data could be fit by variables of a formal, mathematical neural network model of associative learning, as represented by the Schmajuk, Lam and Gray (SLG) model (Schmajuk, 2010; Schmajuk et al., 1996), which has previously proved effective in explaining fMRI data patterns related to fear learning (Dunsmoor & Schmajuk, 2009). In the context of the SLG model, we assumed that the FFA signal reflects two variables: (1) the conditioned response (CR), which is proportional to the prediction of a face based on the association between Frame Color (considered a conditioned stimulus, CS) and Face (considered the unconditioned stimulus, US), and (2) the error of that prediction, given by the occurrence of a Face minus the prediction of a face by the Frame Color (Figure 1). In the model, the CR on each trial depends on the attention paid to the CS and its prediction of the US. We thus used the CR as an estimate of the neural expectation signal of the predictive coding hypothesis, and putative surprise signals were approximated by the error between this prediction and the presence or absence of the US (face).

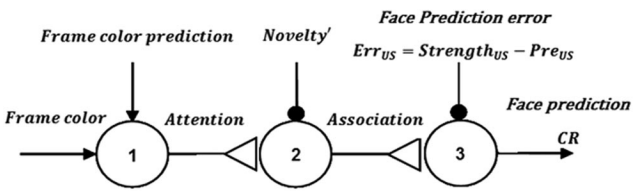

Figure 1. A simplified diagram of the SLG model applied to describing neural activity in the fusiform face area (FFA) in relation to the study of Egner et al. (2010). We assume that activity in the FFA reflects two variables in the model: (1) the CR, which is proportional to the prediction of a face based on the association between Frame Color (the CS) and Face (the US), and corresponds to the output of node 3, which is a regularized version of face prediction (Pre_{US}); (2) the error of that prediction, given by the occurrence of a face minus the prediction of a face, $Err_{US} = Strength_{US} - Pre_{US}$, that is, the Face prediction error modulating node 3 activity.

Our goals were twofold: first, we aimed to test whether the SLG model could produce a close fit of the empirical data, which would provide formal support for the idea that visual neural population activity can be viewed as reflecting an additive mixture of prediction and surprise signals. Second, the fact that the SLG model entails a node that explicitly models the level of attention allocated to incoming stimuli allowed us to test whether differential allocation of attention would be required for the SLG model to fit the fMRI data. Specifically, while Egner et al. (2010) had rendered cue-stimulus associations irrelevant to the participants' task (and thus assumed to have held attention constant across conditions), it is nevertheless possible that attention played a role in mediating effects of predictions or prediction errors, and some authors have in fact proposed that attention is an integral component of predictive coding (Feldman & Friston, 2010; Friston, 2010; Spratling, 2008). We were thus also interested in determining whether and in what way attention in the SLG model would vary with the associative learning variables in fitting the FFA fMRI data.

Method

Experimental Protocol

The fMRI acquisition and analysis parameters and the results are described in detail in Egner et al. (2010). Briefly, 16 healthy volunteers (mean age = 25.3ys) underwent conventional fMRI scanning while viewing black and white images of faces and houses, displayed centrally on a gray background. The goal of the experiment was to induce perceptual expectations (and violations thereof) regarding the presentation of face and house image stimuli. This was done by pairing 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 (CS) was first shown for 250 ms by itself, and then a face or house image (US) 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 for a jittered intertrial interval of 2–4 s.

It was the participants' 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 and consisted of inverted (upside-down) face and house images. To control for attention effects, this task was orthogonal to the manipulation of perceptual expectations, as the colored frames were not predictive of occurrence or type of target stimuli. However, frame color was predictive of the stimulus type for the other 90% 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 face expectation), and the remainder 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). Subsequent to this task, participants were scanned on a standard localizer task to define the "fusiform face area" (FFA)

(Kanwisher et al., 1997). The main data analyses then concerned the fMRI signal recorded from the FFA pertaining to the main task manipulation for regular (nontarget) trials, that is, the FFA data were analyzed according to a 2 (stimulus: face vs. house) × 3 (face expectation: low vs. medium vs. high) ANOVA. In line with the predictive coding model (see Introduction), FFA responses displayed a main effect of stimulus features, as face stimuli elicited higher mean activation than house stimuli, and an interaction between stimulus and expectation factors, as the strength of the stimulus feature effect varied with expectation conditions (Figure 2a). Here, we tested whether this pattern of responses can be formally accounted for within the framework of the SLG model of associative learning.

The SLG Model

The Schmajuk, Lam, & Gray model (SLG, 1996) is an attentional-associative model of classical conditioning (for a more detailed description, please refer to the Appendix). The network incorporates (a) a mechanism that modulates the efficacy of the processing (via attention) of the conditioning stimuli (CSs) in proportion to the total novelty detected in the environment, and (b) a network that forms CS–CS and CS–US excitatory and inhibitory associations, according to a real-time competitive rule. The model assumes that total novelty increases when (a) a predicted CS or predicted US is absent, or (b) an unpredicted CS or unpredicted US is present. Figure 1 shows a simplified diagram of the model as applied to the present data set, illustrating the different mechanisms involved in the generation of a conditioned response (CR, proportional to Face Prediction) when a given CS (Frame Color) is presented. Node 1 receives input from a short-term memory (STM) trace of the Frame Color and the prediction of that Frame Color by the context (CX). To modulate Attention to Frame Color processing in proportion to the novelty detected in the environment, the output of Node 1, (Frame Color + Frame Color Prediction), becomes associated (with the association represented by the first triangle) with the normalized value of the total novelty detected in the environment, Novelty′. Node 3 receives input from Node 2, (Frame Color + Frame Color Prediction) * Attention to Frame Color, as well as from the error term (Face prediction error). The synaptic weight connecting Node 2 to Node 3 reflects the (excitatory or inhibitory) association between the Frame Color with a Face. Changes in Frame Color–Face associations are proportional to an error term (Face prediction error), which reflects the difference between the predicted (Pre_{US}) and the real value of the Face ($Strength_{US}$). Presentation of a Frame Color activates Node 1, which activates Node 2 through the Attention connection, and the output of Node 2 activates Node 3 through the Frame Color-Face association.

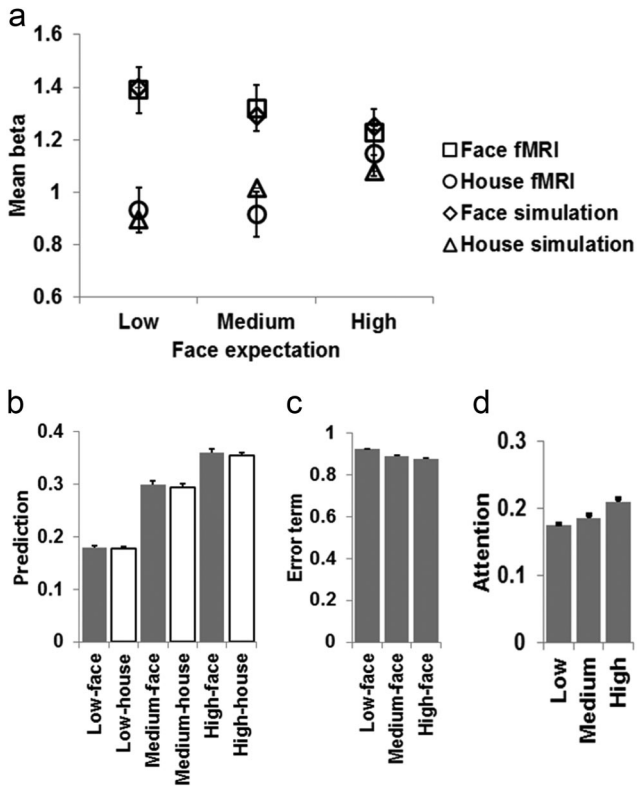

Figure 2. Empirical data and simulation results. a, Neural responses (mean beta weights ± *SEM*) in the FFA are plotted as a function of stimulus (faces vs. house) and expectation for face stimuli, and best-fit simulated values derived from the SLG model are overlaid on the empirical data. b, Mean (± *SEM*) simulated activation values related to face expectation are plotted as a function of experimental conditions (Low-face = low face expectation, face stimulus; Low-house = low face expectation, house stimulus; etc.). c, Mean (± *SEM*) simulated activation values related to surprise (prediction error) in response to face stimuli are plotted as a function of face expectation. d, Mean (± *SEM*) simulated activation values in the SLG model's Attention node are plotted as a function of face expectation.

Simulations

The simulations comprised two steps: (1) simulating face prediction and prediction errors using the SLG model and (2) using them to fit the FFA activation observed in the fMRI data. The simulation parameters used in the first step (the salience of the CX and the CS, and strength of the US) were fixed and based on values employed in recent application of this model to behavioral conditioning experiments (Schmajuk & Kutlu, 2011; Kutlu & Schmajuk, 2012). Specifically, the salience of the CX (Sal_{CX}) was set to 0.1, the salience of the CSs (Sal_{CS}) was set to 1, and the strength of the US ($Strength_{US}$) was set to 1. The SLG model simulated face prediction and prediction errors using the actual trial sequences that participants were exposed to in the empirical study. For the current application, the onset times and condition specifications of each trial of the Egner et al. (2010) experiment were used as input to the SLG model, which estimated predictions and error terms for each trial. Specifically, for each participant, a trial sequence file was generated that reflected the trial sequence experienced by that participant. Each trial had a duration of 30 time units (t.u) and was modeled with three stimuli, namely, a CX,[1] a CS (the colored frames), and a US (face stimuli). The CX, simulating the experiment environment, was presented throughout the whole simulated trial (from 1 t.u to 30 t.u for every trial). Cues were simulated using

[1] The CX can be regarded as an additional CS and sets the context in which the conditioned and unconditioned stimuli are presented. In our set up, this context was constant and CX was the same for all trials.

3 different CSs (one for each frame cue color that indicated the three different levels of face probability) and presented from 5 t.u to 25 t.u, to approximate the timing of cue presentation in the fMRI experiment. In each trial, one CS was presented, based on the trial sequence from the fMRI experiment. All CSs shared the same salience. Face stimuli were simulated using a US, presented from 10 t.u to 25 t.u., again in line with the timing of the fMRI experiment. Because we were interested in simulating the activation pattern of the FFA, which responds specifically to face stimuli, we only presented the US when a face was presented in a given trial, whereas presentation of a house stimulus was represented by the absence of a US. To quantify model based predictions, face expectation was approximated by using the conditioned response (CR) of the SLG model measured at 9 t.u., that is, the level of face expectation just prior to US onset. We used the CR instead of Pre_{US} to simulate face expectation because this study simulates fMRI activations rather than pure prediction values. By contrast, face surprise (prediction error) was approximated by $Strength_{US}$—Pre_{US}, where $Strength_{US}$ reflects the potency of the unconditioned stimulus and Pre_{US} is the prediction value of the US measured in the SLG model (see Appendix). The prediction error term was assessed upon the onset of US presentation at 10 t.u. and did not incorporate the updating of predictions based on the appearance of the US, which would, however, affect the estimates in the subsequent trial.

In the second step, the simulated face expectation and prediction errors were used to fit the FFA activations using the linear model:

$$\beta_i = w_1 CR_i + \delta_i w_2 Err_i + C_1 + \delta_i C_2 \quad (1)$$

Where β_i is the averaged fMRI activation of trial type i (defined by the combination of the stimulus category and the frame) and where CR_i and Err_i are the sample mean of simulated face expectation and prediction error of trial type i, respectively. δ_i is 1 when this trial type contains face stimuli, 0 otherwise. This parameter is used to simulate face specific activation of error terms in the FFA. w_1 and w_2 are the scaling coefficients for face expectation and prediction errors, respectively. C_1 and C_2 represent the constant part of the FFA activity related to face expectation and prediction errors, respectively. Thus, the six observations from the 2×3 design of the fMRI study were modeled using four independent coefficients (w_1, w_2, C_1, C_2).

Results

The SLG model produced a sum of squared error of 0.0174, indicating an excellent fit to the data (see below). Figure 2a displays the empirical data and the fit of the current simulation results based on modeling trial-by-trial associative learning of cue-stimulus events within the SLG model. Figure 2b and 2c displays face expectation (prediction) and surprise (error term) model values for different trial types, averaged across trials and participants. Figure 3 unpacks the averaged prediction values by plotting face expectation terms as a function of time over the course of the experiment (Figure 3a), averaged across all different trial sequences. It can be seen that for all conditions, face prediction starts at zero (no expectation for faces) but in the course of the first 5–10 trials, the values for the three different frame conditions start to diverge (Figure 3b). This documents that the SLG model learned the cue–stimulus probability distributions within a relatively short period of time and sustained them robustly thereafter. Because in our model prediction (CR) was measured before the onset of a US, the category of the presented stimulus (face vs. house) has no effect on face expectation per se during the ongoing trial (but it does of course affect predictions for forthcoming trials). In addition, as expected, face expectations rose with increased cued probability of a forthcoming face stimulus, reflecting the model's associative learning of the CS–US contingencies across the experiment. By contrast, the prediction error term (face surprise) elicited by an unanticipated US decreased with increasing face probability (Figure 2c).

The SLG model captures the data very well, with all of the best-fit values falling within one mean standard error of the empirical data points (Figure 2a), thus providing support for the central tenet of the predictive coding view of visual processing,

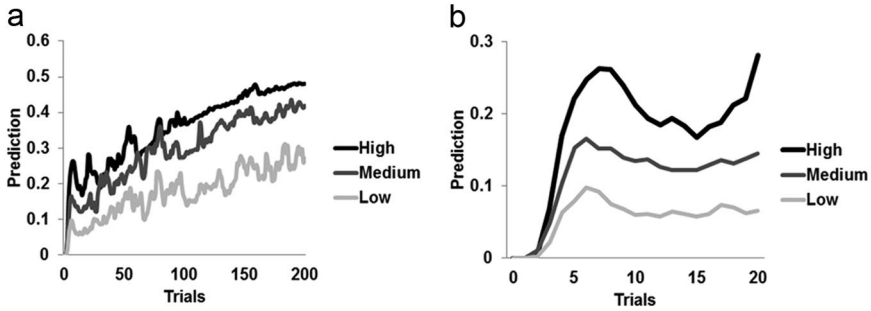

Figure 3. Temporal evolution of simulated face-prediction responses in the FFA across experimental trials. Shown are mean simulated time-courses (averaged across all different trial sequences) of FFA conditioned responses (CRs, Figure 1), that is, predictions of face stimuli, as elicited by each of the three types of conditioning stimuli (CSs; i.e., the different frame colors associated with 25%, 50% and 75% likelihood of the occurrence of a face stimulus). Face prediction values are displayed for the entire experiment time course in A, whereas B zooms in on the first 20 trials, to highlight initial learning by the SLG model. At the start of the experiment, expectations for faces are at zero for all three conditions, but within the first 5–10 trials, the conditioned responses produced by the different CS types start to diverge markedly due to rapid learning of the CS–US associations by the model.

namely, that visual processing is primarily driven by internally generated predictions regarding forthcoming stimulation and their interaction with that stimulus, rather than by bottom-up stimulus features alone. Finally, as the original fMRI study design sought to control for differential recruitment of attention across the experimental conditions (Egner et al., 2010), we also assessed whether the Attention node of the SLG model was differentially activated in the different cue conditions in the current simulations. As shown in Figure 2d, and in apparent contrast to the assumptions of the original paper (Egner et al., 2010), Attention node activation actually displayed a positive association with levels of face expectation (One-way ANOVA, $F(2, 45) = 19.1$, $p < .00001$), with low face prediction associated with less Attention than medium, $t(15) = 2.15$, $p < .05$ or high face prediction, $t(15) = 6.56$, $p < .00001$, and medium face prediction associated with less Attention than high face prediction, $t(15) = 5.44$, $p < .0001$.

Discussion

Egner et al. (2010) reported neural population signals in the human FFA as a function of independently manipulated stimulus features (face vs. house stimuli) and participants' expectations for those features. The data pattern produced by these manipulations qualitatively matched predictions derived from a predictive coding model of visual cognition, where FFA signal would be driven by learned cue-stimulus associations and prediction error. However, that study did not incorporate a formal test of whether an associative learning model acquiring trial-by-trial cue-stimulus associations could in fact provide a quantitative account for these data. Here, we performed such a test and showed that expectation and prediction error signals derived from an associative learning model created to account for behavioral conditioning phenomena can explain the FFA population responses very closely.

Specifically, by applying the attention-associative SLG model (Schmajuk et al., 1996) to simulate the fMRI data, we obtained two main findings. First, we found that SLG model variables expressing learned expectation (CR) of face stimuli, and the violation of these expectations (prediction error), together provided a very close fit to the empirical data, thus showing that neural population signals in the ventral visual stream can be formally accounted for in terms of associative learning variables. Specifically, before stimulus presentation, FFA activity is driven by acquired expectations regarding the likely appearance of a face stimulus. In the case of high face expectation, presentation of a face stimulus elicits little or no prediction error, and thus little or no additional FFA activation occurs. On the other hand, if faces are unexpected, the presentation of a face stimulus elicits a large prediction error response, thus producing additional activity in the FFA. This provides strong formal support for the basic tenets of the predictive coding perspective on visual cognition; namely, that processing at a given stage of the visual hierarchy reflects a summation of expectation and surprise responses associated with a particular visual stimulus or feature, rather than being driven by the mere physical presence of that feature (Friston, 2005; Lee & Mumford, 2003; Mumford, 1992; Rao & Ballard, 1999; Spratling, 2008; Summerfield & Egner, 2009). While our results are in line with these basic hypotheses derived from predictive coding models, we are not in a position to conduct any formal model comparisons, as these models were not designed to simulate the acquisition of cue-stimulus associations of the type we modeled in the current study.

Second, these prediction and surprise signals in the ventral visual stream appear to be partly mediated by attention, as the SLG model's Attention node activation scaled positively with face prediction in producing the fMRI data fit. This finding suggests that, even though Egner and colleagues (2010) sought to equate attention across conditions by rendering the experimental manipulations incidental to the participants' task, predictive processing in the FFA may nevertheless have interacted with attention, whereby cues signaling a higher likelihood of face occurrence resulted in greater attentional activation. Specifically, CSs that signaled a higher probability of face occurrence were associated with greater attention. In the model, this would in turn enhance face prediction activity and, in the case where predictions are violated, elicit a larger prediction error response. The positive association between expectation and attention in the present simulations appears broadly consistent with recent proposals that view attention as an inherent consequence of expectations in the predictive coding framework (Feldman & Friston, 2010; Friston, 2010). These findings also highlight the importance of addressing the precise relationship between expectation and attention explicitly in future empirical investigations, ideally by manipulating variables affecting stimulus probability and stimulus relevance in an orthogonal fashion (Summerfield & Egner, 2009; Wyart, Nobre, & Summerfield, 2012).

Conclusion

We showed that neural population responses in the ventral visual stream can be quantitatively explained by formal associative learning parameters coding for predicted percepts and mismatches between predictions and percepts (prediction error), thus providing strong support for predictive coding models of visual cognition.

References

Alink, A., Schwiedrzik, C. M., Kohler, A., Singer, W., & Muckli, L. (2010). Stimulus predictability reduces responses in primary visual cortex. *Journal of Neuroscience, 30,* 2960–2966.

Baker, A. G. (1974). Conditioned inhibition is not symmetrical opposite of conditioned excitation: Test of Rescorla-Wagner Model. *Learning and Motivation, 5,* 369–379.

den Ouden, H. E., Daunizeau, J., Roiser, J., Friston, K. J., & Stephan, K. E. (2010). Striatal prediction error modulates cortical coupling. *Journal of Neuroscience, 30,* 3210–3219.

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

Dunsmoor, J., & Schmajuk, N. (2009). Interpreting patterns of brain activation in human fear conditioning with an attentional-associative learning model. *Behavioral Neuroscience, 123,* 851–855.

Egner, T., Monti, J. M., & Summerfield, C. (2010). Expectation and surprise determine neural population responses in the ventral visual stream. *Journal of Neuroscience, 30,* 16601–16608.

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free energy. *Frontiers in Human Neuroscience, 4,* 215. doi:10.3389/fnhum.2010.00215

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 360,* 815–836.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Review Neuroscience, 11,* 127–138.

Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review, 87,* 1–51.

Helmholtz, H. (1876). *Handbuch der physiologischen Optik* (Vol. 9). Leipzig, Germany: Leopold Voss.

Hubel, D. H., & Wiesel, T. N. (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the Cat. *Journal of Neurophysiology, 28,* 229–289.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17,* 4302–4311.

Kutlu, M. G., & Schmajuk, N. A. (2012). Deactivation and reactivation of the inhibitory power of a conditioned inhibitor: Testing the predictions of an attentional-associative model. *Learning & Behavior, 40,* 83–97.

Larrauri, J. A., & Schmajuk, N. A. (2008). Attentional, associative, and configural mechanisms in extinction. *Psychological Review, 115,* 640–676.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision, 20,* 1434–1448.

McLaren, I. P. L., & Mackintosh, N. J. (2000). An elemental model of associative learning: I. Latent inhibition and perceptual learning. *Animal Learning & Behavior, 28,* 211–246.

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

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science, 304,* 452–454.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2,* 79–87.

Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience, 3 Suppl,* 1199–1204.

Schmajuk, N. A. (2009). Attentional and error-correcting associative mechanisms in classical conditioning. *Journal of Experimental Psychology-Animal Behavior Processes, 35,* 407–418.

Schmajuk, N. A. (2010). Mechanisms in classical conditioning: A computational approach. New York, NY: Cambridge University Press.

Schmajuk, N. A., Buhusi, C., & Gray, J. A. (1996). An attentional-configural model of classical conditioning. *Journal of Mathematical Psychology, 40,* 358–358.

Schmajuk, N. A., Gray, J. A., & Lam, Y. W. (1996). Latent inhibition: A neural network approach. *Journal of Experimental Psychology: Animal Behavior Processes, 22,* 321–349.

Schmajuk, N. A., & Kutlu, M. G. (2011). Latent inhibition and compound conditioning: A reply to Holmes and Harris (2009). *Journal of Experimental Psychology: Animal Behavior Processes, 37,* 254–260.

Schmajuk, N. A., Lam, Y. W., & Gray, J. A. (1996). Latent inhibition: A neural network approach. *Journal of Experimental Psychology-Animal Behavior Processes, 22,* 321–349.

Schmajuk, N. A., & Larrauri, J. (2008). Associative models can describe both causal learning and conditioning. *Behavioral Processes, 77,* 443–445.

Schmajuk, N. A., & Larrauri, J. A. (2006). Experimental challenges to theories of classical conditioning: Application of an attentional model of storage and retrieval. *Journal of Experimental Psychology-Animal Behavior Processes, 32,* 1–20.

Spratling, M. W. (2008). Predictive coding as a model of biased competition in visual attention. *Vision Research, 48,* 1391–1408.

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., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Science, 13,* 403–409.

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

Summerfield, C., Trittschuh, E. H., Monti, J. M., Mesulam, M. M., & Egner, T. (2008). Neural repetition suppression reflects fulfilled perceptual expectations. *Nature Neuroscience, 11,* 1004–1006.

Wyart, V., Nobre, A. C., & Summerfield, C. (2012). Dissociable prior influences of signal probability and relevance on visual contrast sensitivity. *Proceedings of the National Academy of Sciences of the United States of America, 109,* 3593–3598.

Zimmer-Hart, C. L., & Rescorla, R. A. (1974). Extinction of Pavlovian conditioned inhibition. *Journal of Comparative Physiology. A, Sensory, Neural, and Behavioral Physiology, 86,* 837–845.

Appendix

SLG Model Details

The Schmajuk, Lam, and Gray (1996) Model

Here, we offer a summarized description of the SLG model, a neural-network theory that describes (and has successfully predicted) many features that characterize classical conditioning (Larrauri & Schmajuk, 2008; Schmajuk, 2009; Schmajuk & Kutlu, 2011; Schmajuk & Larrauri, 2006). In the SLG model, the learning process is accounted for jointly by an attentional mechanism and an associative mechanism. Note that because the CX works in the same way as the CS (except that the CX is presented throughout the experiment), the following equations only model the variables related to the CS to reduce repetition. Those equations can be extended to describe the CX related variables by replacing CS related variable with the corresponding CX related variables.

Attentional Mechanism

The model assumes that presentation of a conditioning stimulus (CS) activates a STM trace, τ_{CS}. The dynamics of τ_{CS} is modeled by:

$$\frac{d\tau_{CS}}{dt} = K_1 \, (Sal_{CS} - \tau_{CS}) \tag{A1}$$

This STM trace, τ_{CS}, is added to Pre_{CS}, defined as the prediction of the CS by other CSs, the context (CX), and the CS itself. The sum of τ_{CS} and Pre_{CS}, $(\tau_{CS} + Pre_{CS})$, activates a synaptic weight proportional to the positive value of attention, z_{CS}. The value of z_{CS} is computed as the association between $(\tau_{CS} + Pre_{CS})$ with the value of Novelty′. Changes in z_{CS} during 1 time unit (t.u.) are given by

$$\frac{dz_{CS}}{dt} \sim (\tau_{CS} + Pre_{CS})(Novelty'(1 - z_{CS}) - (1 + z_{CS})) \tag{A2}$$

where Novelty′ is proportional to the sum of the novelties of all stimuli present or predicted at a given time. The novelty of a given CS, CX, or the unconditioned stimulus (US) is computed as the absolute value of the difference between the average observed value of the CS, CX, or the US, and the average of their corresponding predicted value. By Equation [A2], z_{CS} increases to $+1$ when Novelty′ is relatively large, and decreases to -1 otherwise.

The attention-modulated representation of the CS (X_{CS}) is given by

$$X_{CS} \sim (z_{CS} + K_5)(\tau_{CS} + Pre_{CS}) \tag{A3}$$

where K_5 represents a nonmodifiable connection between $(\tau_{CS} + Pre_{CS})$ and X_{CS}, and z_{CS} assumes only positive values.

Associative Mechanism

Changes in the excitatory or inhibitory association (V_{CS-US}), between X_{CS} and the US, are proportional to

$$\frac{dV_{CS-US}}{dt} \sim X_{CS}(Strength_{US} - Pre_{US})(1 - |V_{CS-US}|)) \tag{A4}$$

where $Strength_{US}$ is the strength of the US, Pre_{US} is the aggregate prediction of the US by all X's active at a given time, $Strength_{US} - Pre_{US}$) is the prediction error term, and the individual error term $(1 - |V_{CS-US}|)$ constrains V_{CS-US}, $-1 < V_{CS-US} < +1$. Associations V_{CS-US} increase when the prediction error term is positive and decrease when it is negative.

Because presentation of a conditioned inhibitor does not decrease its inhibitory power (Zimmer-Hart & Rescorla, 1974), the model assumes that when $Pre_{US} < 0$ then $Pre_{US} = 0$, and when $Pre_{CS} < 0$ then $Pre_{CS} = 0$ (see also McLaren & Mackintosh, 2000). Most importantly, this assumption correctly predicts that a neutral stimulus does not become excitatory when presented with an inhibitory stimulus (Baker, 1974).

Performance

The aggregate prediction of the US by all CSs with representations active at a given time, Pre_{US}, is given by

$$Pre_{US} = \sum X_{\mathrm{CS}} V_{CS-US} \tag{A5}$$

The strength of the CR is given by

$$CR = \frac{Pre_{US}^2}{(Pre_{US}^2 + K_{11}^2)} \tag{A6}$$

where K_{11} represents a nonmodifiable parameter that constrains CR within the range of 0 and 1.

Notice that attention z_{CS} and X_{CS} (Equations A2 and A3) control the formation of V_{CS-US} and CS-CS (V_{CS-US}) associations during conditioning (Equation A4), and the activation of V_{CS-US} (Equation A5) and the CR (Equation A6).

Parameter Values

A detailed description of the model's differential equations is offered in Schmajuk, Lam, and Gray (1996). The present simulations use parameter values identical to those used in previous publications (Schmajuk & Larrauri, 2008; Schmajuk & Larrauri, 2006), which have been applied to a vast number of classical conditioning paradigms: $K_1 = .2$, $K_5 = .02$ and $K_{11} = .15$ (values of all parameters can be found in Schmajuk et al.1996). In the present study, face prediction was simulated by CR; prediction error was simulated using $Strength_{US} - Pre_{US}$; and attention was simulated by z_{CS}.

Received January 30, 2012
Revision received May 14, 2012
Accepted May 19, 2012 ■