Journal of Mathematical Psychology 76 (2017) 80–93

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Integrating theoretical models with functional neuroimaging

Michael S. Pratte [a,b,*], Frank Tong [b]

[a] Department of Psychology, Mississippi State University, United States

[b] Department of Psychology and the Vanderbilt Vision Research Center, Vanderbilt University, United States

ARTICLE INFO

Article history:
Available online 25 July 2016

ABSTRACT

The development of mathematical models to characterize perceptual and cognitive processes dates back almost to the inception of the field of psychology. Since the 1990s, human functional neuroimaging has provided for rapid empirical and theoretical advances across a variety of domains in cognitive neuroscience. In more recent work, formal modeling and neuroimaging approaches are being successfully combined, often producing models with a level of specificity and rigor that would not have been possible by studying behavior alone. In this review, we highlight examples of recent studies that utilize this combined approach to provide novel insights into the mechanisms underlying human cognition. The studies described here span domains of perception, attention, memory, categorization, and cognitive control, employing a variety of analytic and model-inspired approaches. Across these diverse studies, a common theme is that individually tailored, creative solutions are often needed to establish compelling links between multi-parameter models and complex sets of neural data. We conclude that future developments in model-based cognitive neuroscience will have great potential to advance our theoretical understanding and ability to model both low-level and high-level cognitive processes.

1. Introduction

The general goal of cognitive psychology has been to understand psychological processes at what Marr (1982) would call the *algorithmic* or *representational level* (see Love, 2015). In order to explore the algorithms and representational structures that might underlie processes such as attention or memory, cognitive psychologists often propose formal theoretical models of these processes, and test them by assessing predicted patterns in behavior. For decades, doing so has provided remarkable insights into how the mind works. During the same time, the field of neuroscience has made strides in understanding what Marr calls the *implementation* level, or, how these processes are implemented in the biological machinery that makes up the brain. More recently, the merger of these fields into a unified *cognitive neuroscience* has resulted in part from the development of new neuroimaging techniques, such as functional magnetic resonance imaging (fMRI), which have made investigating the biological substrates of human cognition possible. The more targeted approach of combining theoretical modeling and neuroscience has been termed *computational neuroscience*, a field that is often credited as originating from Marr's work. In this review, we consider a newly emerging endeavor that has arisen from the merger of cognitive psychology, theoretical modeling and neuroscience: using theoretical models in conjunction with human neuroimaging to study psychological processes.

Advances in mathematical and computational approaches have played a key role in fMRI since its invention. Some of these developments include analytical approaches for extracting relevant information from the BOLD signal across the temporal domain, such as the use of temporal phase-encoded designs (Engel, 2012; Engel et al., 1994; Sereno et al., 1995) and the development of de-convolution approaches for fast-event related designs (Boynton, Engel, Glover, & Heeger, 1996; Buckner et al., 1996; Glover, 1999). Developments in inferential statistical techniques have produced a number of tools that have helped make fMRI mapping studies so successful, especially regarding techniques to account for what is arguably the most serious multiple comparisons problem in psychology (see Nichols & Hayasaka, 2003). Given the multivariate nature of fMRI data, correlation-based approaches (Haxby et al., 2001), machine learning techniques (Kamitani & Tong, 2005; Norman, Polyn, Detre, & Haxby, 2006; Tong & Pratte, 2012) and voxel-based modeling approaches (Brouwer & Heeger, 2009; Kay, Naselaris, Prenger, & Gallant, 2008; Serences & Saproo, 2009) have been used to capture the complexities of these high-dimensional data sets, providing powerful new ways of identifying perceptual

* Correspondence to: Mississippi State University, Department of Psychology, 255 Lee Blvd, Starkville, MS 39762, United States.
E-mail address: prattems@gmail.com (M.S. Pratte).

http://dx.doi.org/10.1016/j.jmp.2016.06.008
0022-2496/

information contained in brain activity patterns (Harrison & Tong, 2009; Haynes & Rees, 2005; Kamitani & Tong, 2005; Serences & Boynton, 2007), as well as evidence of semantic tuning properties (Huth, Nishimoto, Vu, & Gallant, 2012; Mitchell et al., 2008). The correlational structure of activity patterns has also been used to characterize object representations in the ventral temporal lobe (Haxby et al., 2001; Kriegeskorte et al., 2008). Correlation-based approaches have also proven useful for delineating the functional connectivity of the human brain (e.g., Honey et al., 2009), including resting state networks (Fox et al., 2005), and more recent studies of brain connectivity have benefitted from the application of graph theoretical models (Bullmore & Sporns, 2009) and other model-based approaches (Tavor et al., 2016). Such advances in analytic methods continue to expand the ways in which fMRI can be used to study the brain.

More recently, research on new mathematical approaches for fMRI has evolved beyond the goal of simply developing more powerful analytic methods, to that of integrating and testing cognitive models. This model-based approach to cognitive neuroscience represents an exciting development that goes beyond the simpler goals of "brain mapping", identifying correlations between individual differences and brain activity, or information-based approaches to characterize cortical function. Instead, the goal of model-based cognitive neuroscience lies in describing the perceptual or cognitive processes that underlie behavior in a mathematically precise manner, and determining the neural processes that underlie these computations.

Cognitive process models have a long history in the study of human performance. For example, models based on signal detection theory have served as the foundation for studying perception (Green & Swets, 1966), attention (Lu & Dosher, 1998) and memory (Kintsch, 1967). Early cognitive research also demonstrated that stochastic accumulator models can accurately predict patterns of choice reaction times across numerous behavioral paradigms (Ratcliff & Rouder, 1998; Stone, 1960). While some cognitive process models have focused on identifying and quantifying a few key parameters to capture patterns of cognitive performance, other models rely on general learning principles to train complex networks with numerous parameters to perform a cognitive task. For example, neural network models (e.g., McClelland & Rogers, 2003) have been developed to characterize high-level processes including speech perception (McClelland & Elman, 1986), categorization (Ashby & Maddox, 1993; Nosofsky & Palmeri, 1997), cognitive control (Botvinick, Braver, Barch, Carter, & Cohen, 2001), and human memory (Polyn, Norman, & Kahana, 2009).

One might expect that the application of theoretical models to neuroimaging data would be a naturally obvious and fruitful endeavor. However, most cognitive neuroscientists have not rushed to meet this challenge until recently. Why has this been the case? A central challenge lies in establishing strong links between the parameters of a cognitive model and particular brain responses embedded within a cognitive experiment. Cognitive models typically rely on latent constructs of presumed psychological processes that must somehow be translated into a predicted pattern of brain responses. If the model leads to clear predictions regarding how the univariate BOLD responses should change over time, such models may be more straightforward to test using standard fMRI analysis procedures. Earlier applications of model-based fMRI have relied on such approaches to identify the neural correlates of reward prediction error (e.g., O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003; O'Doherty et al., 2004; O'Doherty, Hampton, & Kim, 2007) and response conflict (e.g., Botvinick, Cohen, & Carter, 2004). However, fMRI data is very high-dimensional, such that establishing links between cognitive models and the information contained in multivariate brain activity patterns is considerably more challenging. Even if a model can be positively related to an information-based metric of brain processing, the next step of determining whether a particular model provides a compelling fit of the high-dimensional brain data can be difficult to demonstrate.

In this review, we highlight several recent studies that have successfully combined theoretical models with fMRI data, addressing diverse questions spanning lower-level perceptual processes to higher-level cognitive processes. A central theme across these studies is the goal of identifying compelling relationships between brain, model and behavior (see Fig. 1(A)), often with the cognitive model serving as the intermediary for mapping between brain and behavior. However, as we will see, there are many possible options and approaches for establishing these links, as a model fitted to behavioral data might be used to predict brain responses, or brain data might be incorporated into a model to predict behavior. Moreover, intermediate processing steps may take place before links are established, such as methods to reduce the high dimensionality of brain data to lower-dimensional measures that can be more directly related to model predictions.

We begin this review by discussing an application of the normalization model to the visual perception of orientation (Brouwer & Heeger, 2011). In this work, fMRI data from early cortical visual areas was first transformed into interpretable constructs using a multivariate modeling approach, and the normalization model was then fitted to the resulting measurements. We then describe an application of models of visual attention to fMRI data (Pratte, Ling, Swisher, & Tong, 2013). Here, a multivoxel pattern classification approach was used to transform multivariate fMRI data, obtained from multiple levels of the visual hierarchy, into an interpretable measure of information representation, and the theoretical model was fitted to the result. Whereas in both of these studies, a formal model was fitted to information decoded from the multivariate fMRI signal (Fig. 1(B)), we next consider a study in which the time course of the fMRI signal on individual trials was incorporated within a theoretical model of behavioral memory performance (Fig. 1(C)), to determine whether this neural signal can lead to more accurate predictions of free recall performance (Kragel, Morton, & Polyn, 2015). An fMRI study of categorization highlights yet another approach (Fig. 1(D)), by assessing the degree to which competing models of behavioral categorization performance can account for observed patterns of neural data (Mack, Preston, & Love, 2013). Finally, we review a study of cognitive control that demonstrates how the application of a theoretical model to fMRI data can reveal new insights about neural processing that would have been impossible without the model (Ide, Shenoy, Yu, & Li, 2013). Here, a model of behavioral performance in the stop-signal task was incorporated within the fMRI analysis (Fig. 1(E)), and the results suggest that the function of the anterior cingulate is more specific than has been suggested by previous studies.

This collection of studies demonstrates both the feasibility and potential of model-based cognitive neuroscience. The approaches are remarkably diverse, both in how the data are used to inform the model, and in the technical solutions employed to establish compelling relationships between brain, model and cognitive performance (see Fig. 1). Of particular interest is that fact that none of the reviewed works share exactly the same strategy to link a theoretical model with functional imaging data. Rather, these examples underscore how individual studies have relied on clever innovations that are custom-built for a particular model or experimental paradigm. As such, we neither foresee nor prescribe a one-size-fits-all approach to model-based cognitive neuroscience. Instead, we believe that the diversity in attempts to integrate modeling and functional brain imaging will forward the advance of theoretical models with a momentum that would not happen if these fields remained isolated from one another.

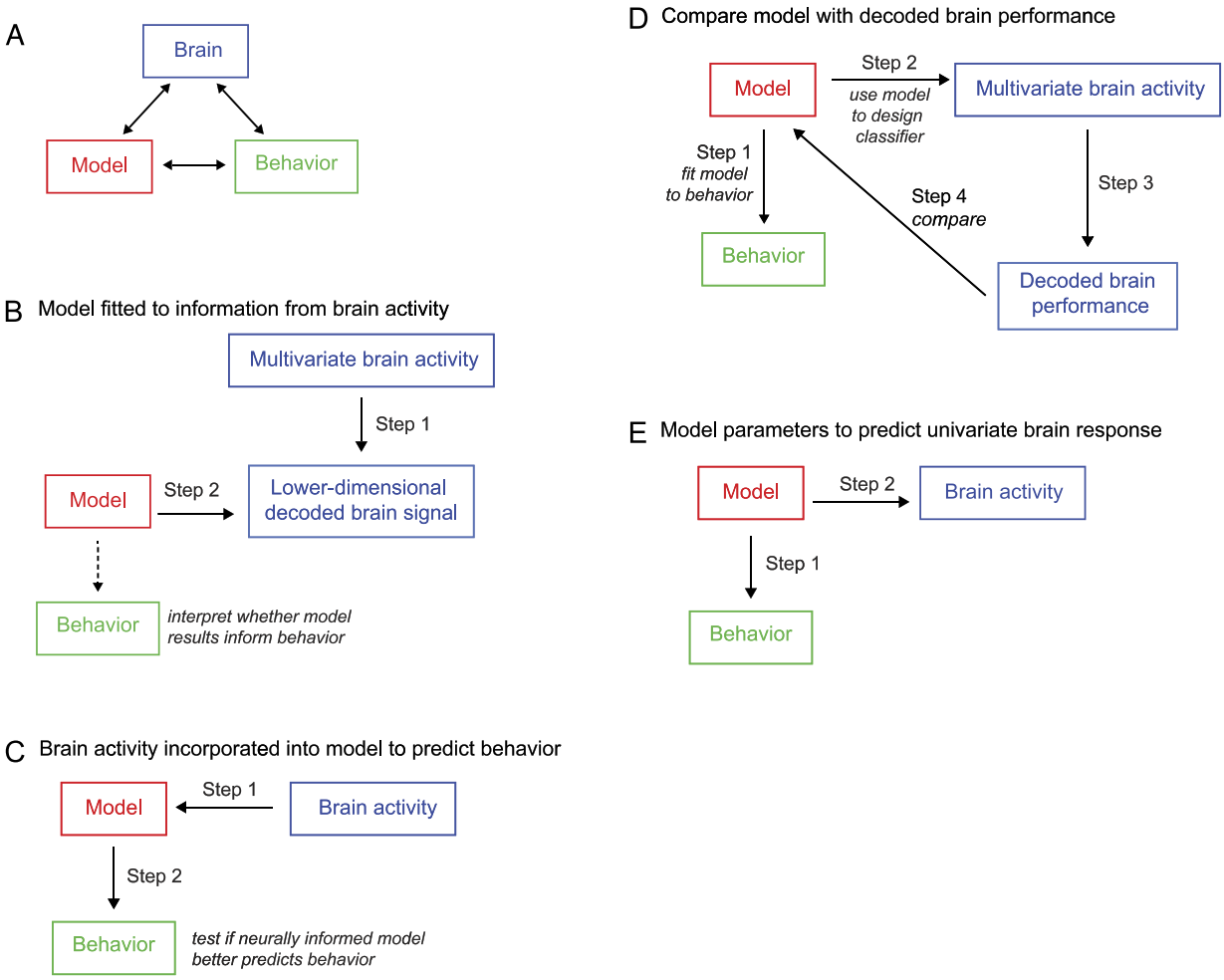

Fig. 1. Approaches To Linking Models With Brain and Behavior. (A) The goal of a model-based cognitive neuroscience is to develop theory by forming links between brain activity, behavior, and a theoretical model that describes these data. (B) In the approach taken by Brouwer and Heeger (2011) and Pratte et al. (2013), a multivariate pattern analysis is first used to transform the brain activity data into a lower-dimensional space, and the model is then applied to these more interpretable data. (C) In the approach taken by Kragel et al. (2015) the brain activity signal is directly incorporated within the model, and statistical tests are used to determine whether this inclusion produces a more accurate account of the behavioral data than a model without the neural data. (D) Mack et al. (2013) first fitted a model to behavioral data, and used the fitted model to decode patterns of brain activity. They then compared this decoding performance across different models to determine which model provided the best correspondence with the neural data. (E) Ide et al. (2013) first fited a model to behavioral data, and then used a regression analysis to identify fMRI voxels that co-varied with various model constructs over the course of the experiment.

2. Perception: the normalization model

A major goal in the study of visual perception is to understand how neural responses to a stimulus are affected by the presence of other stimuli in the visual field. One of the most widely studied models of such stimulus interactions, the *normalization model* (see Carandini & Heeger, 2012 for a review) has been rigorously tested using behavioral, neurophysiological, and more recently, human neuroimaging approaches. The normalization model posits that the response of a neuron to a stimulus will depend on its sensitivity to the stimulus, and further that this response is divisively reduced by the concurrent activity of other neighboring neurons. The activity of these other neurons is driven largely by the presence of other stimuli, such that the model accounts for a myriad of behavioral and neural effects resulting from stimulus interactions. The purported benefits of such a normalization process to the perceptual system include reducing the overall neural and metabolic demands within a local region, de-correlating the responses of neurons with different stimulus preferences, and providing for more efficient neural coding.

Before considering interactions between multiple stimuli, a simpler case of normalization can be understood by considering how the response of a single neuron changes as a stimulus changes size. Neural responses in many brain areas, such as the lateral geniculate nucleus (LGN) and striate cortex (V1), increase monotonically as a function of stimulus contrast. These responses typically follow a sigmoidally shaped response function, with a signature compressive effect at high stimulus contrasts (Fig. 2(A)). For example, consider a neuron in the LGN that responds to stimulation within its receptive field, in a manner that depends on the stimulus contrast (C). According to the normalization model, the activity within a pool of nearby cells that also respond to this stimulus (C_m) will divisively dampen the response of the target neuron $R(C)$,

$$R(C) = \frac{C}{\sigma + \sum C_m}$$

where parameter σ determines the shape of the contrast-response function.[1] The population of neurons in the denominator (C_m) is referred to as the normalization pool, and the summed responses

[1] These models often include parameters such as an intercept, scale and non-linearity to account for the various measurements that are used. These parameters are omitted here for simplicity.

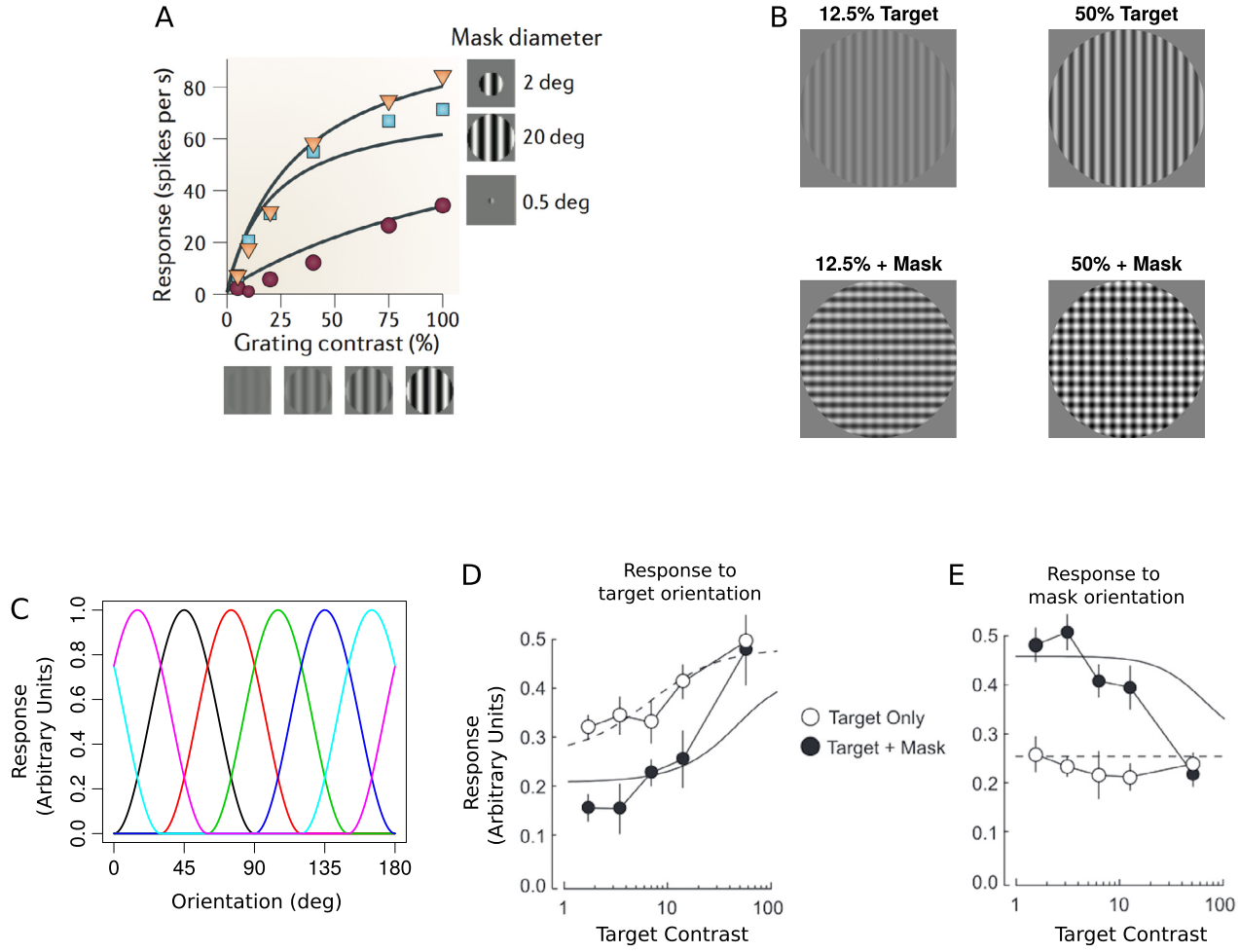

Fig. 2. Normalization Model of Perception. (A) Responses of an LGN neuron to a stimulus within its receptive field, plotted as a function of stimulus contrast (Data from Bonin et al., 2005, figure adapted from Carandini & Heeger, 2012). Colors denote responses for various stimulus sizes; lines show fits of the normalization model. (B) Example of cross-orientation suppression stimuli used by Brouwer and Heeger (2011), where targets of various contrast are presented alone (top) or with a superimposed mask (bottom). (C) Orientation tuning curves that serve as basis functions for the forward encoding model. The orientation-tuned response of each voxel is modeled as a weighted sum of these response curves, with weighs estimated from the data. (D) Target-tuned responses in V1 as a function of target contrast presented alone (open points) or with a superimposed mask (filled points) from Brouwer and Heeger (2011). (E) Mask-tuned responses in V1 as a function of target contrast when the mask was not presented (open points) and when it was superimposed on the target (filled points). Panels D and E adapted from Brouwer and Heeger (2011). (Color figures available in the online version of this article.)

of these neurons act to *normalize* the response of the target neuron. When the response of a neuron is plotted as a function of contrast, increased activity in the neuron's normalization pool will act to suppress responses of the target neuron by an amount that depends on the stimulus contrast. For example, Fig. 2(A) shows that the response of an LGN neuron is suppressed as the stimulus is enlarged from 2° to 20°, growing beyond that neuron's receptive field (Bonin, Mante, & Carandini, 2005). According to the normalization model, a larger stimulus will activate more nearby LGN neurons that make up the target neuron's normalization pool, thereby suppressing the response of the target neuron. When the stimulus is made to be smaller than 2°, the response again decreases as the stimulus no longer occupies the entire receptive field. Overall, the relationship between stimulus size and the response of the neuron is non-monotonic at high stimulus contrasts, and the normalization model provides an accurate account of the shape of the contrast-response function, and the complex ways in which its shape changes across many stimulus manipulations.

The normalization model not only accounts for the neural interactions that may be evoked by a single stimulus, it provides a parsimonious account of the non-linear interactions that can take place when multiple stimuli are presented, such as superimposed gratings of differing orientation (see Fig. 2(B)). Consider a target population of neurons with responses that depend on the contrast (C) and the orientation (θ) of a stimulus within the population's receptive field, where responses to various orientations are described by some function $f(\theta)$. The responses of nearby neurons will also depend on the contrast and orientation of the stimulus according to their own tuning curves $f_m(\theta)$. These tuning curves are incorporated into the normalization model by assuming that the response of a neuron to a particular stimulus is weighted by the neuron's tuning function and divisively normalized by the response of neighboring neurons to that same stimulus,

$$R(C,\theta) = \frac{f(\theta)C}{\sigma + \sum f_m(\theta)C_m}.$$

A major goal of the normalization model is to account for neural responses to simultaneously presented stimuli. A prominent example is *cross-orientation suppression*, in which responses to a preferred target orientation are suppressed by the presence of a masking grating with an orthogonal orientation. For example, consider the case when a vertical target stimulus of various contrasts is presented alone, or superimposed with a horizontal mask of 50% contrast (Fig. 2(B)). The normalization model is easily amended to account for this situation. Let C_1 and C_2 denote the contrast of target and mask, respectively, and θ_1 and θ_2 denote their

orientations. The total response of the target population to both stimuli becomes:

$$R(C_1, C_2, \theta_1, \theta_2) = \frac{f(\theta_1)C_1 + f(\theta_2)C_2}{\sigma + \sum f_m(\theta_1)C_{1,m} + \sum f_m(\theta_2)C_{2,m}}.$$

The model simply states that the total response reflects the summed response to both target and mask, but that this response is dampened by the summed responses of neurons in the normalization pool to both stimuli.

Brouwer and Heeger (2011) tested the predictions of the normalization model for cross-orientation suppression in humans using fMRI, following the steps shown in Fig. 1(B). Each measured fMRI voxel reflects the pooled response from many neurons, and in an early visual area such as the primary visual cortex (V1), most neurons exhibit strong selectivity for local orientation. The orientation-selective response of a voxel that encompasses many such neurons can be described by a combination of the tuning curves of the underlying neurons. Brouwer and Heeger (2011) developed a *forward* modeling approach (Brouwer & Heeger, 2009; Kay et al., 2008) in which the responses of a voxel to different orientations are modeled as a linear combination of orientation-tuned basis functions. For example, Brouwer and Heeger modeled orientation responses using 6 evenly spaced raised cosine profiles to cover the space of all orientations (Fig. 2(C)), and the amplitude of response for each of these 6 tuned populations could then be estimated for each voxel by using linear regression applied to a subset of the fMRI data. The fitted model is then inverted and applied to voxel responses in another data set, providing an overall measure of how strongly particular orientations are activated across the entire population of voxels in V1. This analysis allowed the researchers to separately measure the strength of V1 responses to various orientations. Consequently, in a cross-orientation scenario in which two stimuli are presented simultaneously, the combined response of voxels in V1 could be decomposed to provide estimated responses to each stimulus separately.

In their experiment, Brouwer and Heeger (2011) showed vertically oriented target stimuli of various contrasts either alone, or superimposed with a horizontally oriented mask stimulus of 50% contrast (Fig. 2(B)). Orientation preferences were first determined for voxels in V1, and then this estimated forward model was used to provide separate measures of V1 responses to target and mask stimuli. Fig. 2(D) shows the resulting contrast response functions for V1 responses to the target orientation, and the fitted normalization model. In the absence of a mask, the target-tuned response rises with target contrast in a way that is accounted for well by the model. The addition of the superimposed horizontal mask reduces target-tuned responses, and this gross effect is the signature of cross-orientation suppression. Specifically, the effect of the mask is to reduce target-tuned responses for low-contrast targets, but it has a weaker suppressive influence when the target grating's contrast exceeds that of the masking grating, as predicted by the normalization model.

Fig. 2(E) shows V1 responses to the mask-tuned orientation, as a function of target contrast. In the absence of a mask (Target Only condition), these responses are invariant to target contrast as expected. However, the presence of the horizontal mask leads to large horizontally tuned responses for low target contrasts, and these responses systematically decrease as the target contrast is increased. This negative-going curve is also predicted by the normalization model: as the target-tuned cells become more strongly activated by increasing target contrast, these target-tuned responses will act to normalize responses to the mask, leading to greater inhibition of the mask-tuned cells at high target contrasts.

The normalization model provides a simple and accurate account of how fMRI responses behave in this rather complex stimulus paradigm. This study relied on a sophisticated approach to decompose multivariate V1 activity into target-selective and mask-selective responses. This approach therefore provides a starting point for further investigations of the normalization model, such as its application to other feature domains (e.g., Moradi & Heeger, 2009) and to higher-level processes such as attention (e.g., Herrmann, Heeger, & Carrasco, 2012).

3. Attention: the perceptual template model

The effect of attending to a particular stimulus is often characterized as enhancing the perceived *quality* of the attended stimulus. A prominent example of this effect is when the to-be-attended stimulus is embedded in external noise, such as trying to focus on a person's voice at a crowded party or trying to see through a rain-spattered windshield. In an effort to describe the mechanisms that underlie our ability to enhance the representation of an attended stimulus in the presence of perceptual noise, Lu and Dosher (1998) constructed a model based on core concepts of signal detection theory and signal processing, which they termed the *Perceptual Template Model* (PTM). The PTM model assumes that attending to a noisy stimulus has the effect of increasing the signal-to-noise ratio of the representation of that stimulus. As such, attention can do one of two things: increase the strength of the incoming input (*amplification*), or decrease the irrelevant noise (*noise reduction*). Whereas the amplification mechanism simply increases responses to all of the incoming information (i.e., both signal and noise), the noise-reduction mechanism selectively decreases responses to the external noise. To do so, the model assumes that there exists a *perceptual template* of the relevant signal, which allows the brain to discriminate between signal and noise. Since its inception, this model has been applied and tested behaviorally in many contexts (see Carrasco, 2011 for a review). More recently, the PTM model has been used with fMRI to determine the neural correlates of these purported attentional processes.

In the vast majority of fMRI studies, the effect of attention is to increase the mean BOLD signal, for example, within retinotopic visual areas corresponding to the attended part of space (e.g., Buracas & Boynton, 2007; Kastner, Pinsk, De Weerd, Desimone, & Ungerleider, 1999). The PTM model, however, makes the opposite prediction for a specific region of the stimulus space: If attention acts as a noise-reduction mechanism in a brain region, then attending to a low-contrast stimulus embedded in high levels of external noise should reduce the total response to signal plus noise, as responses to the noise are reduced but responses to the signal are not affected. If the mean BOLD signal in a region reflects the total activity of cells responding to both signal and noise, then attention should reduce mean BOLD responses for low-contrast signals presented in noise. Lu, Li, Tjan, Dosher, and Chu (2011) found support for this prediction, showing that attending to low-contrast gratings in high levels of external noise led to reduced mean BOLD responses in V1, whereas attention increased responses to the low-contrast grating when it was presented alone.

These results illustrate that it is possible to test theoretical models with fMRI by making predictions about mean BOLD, especially by exploring predicted interactions with parametric manipulations of the stimulus. However, the formal PTM model specifies how amplification and noise reduction affect the signal-to-noise ratio of stimulus information, rather than the total neural response as is measured with mean BOLD. In particular, consider a stimulus with contrast C that is embedded in visual noise (termed *external noise*) with contrast N_E. The goal is to predict some measure of the sensitivity to a signal relative to the noise (d') for a neural population tuned to some property of the stimulus (e.g. orientation). Sensitivity varies as a function of the contrast of the stimulus and the contrast of the noise. In the absence of

attention, sensitivity is proportional to the stimulus strength, and lowered by external noise:

$$d'(C, N_E) = \frac{C}{\sqrt{N_E^2 + N_I^2}}$$

where N_I is the level of *internal noise* inherent in the perceptual system, and determines the shape of the response which varies as a function of stimulus contrast (C) or external noise contrast (N_E). Within the denominator, external and internal noise sources are combined, and the combined effects act to degrade the representation of the stimulus.

According to the PTM model, there are two ways that attention can act on the incoming signal and noise. An amplification mechanism (a) will multiplicatively enhance the entire input signal, thus increasing responses to both the signal and the external noise. As a consequence, amplification can improve the representation of the stimulus if external noise ranges from negligible levels to levels not much greater than that of internal noise. In contrast, a selective noise-reduction mechanism (r) will act solely to decrease the level of external noise without affecting the stimulus:

$$d'(C, N_E) = \frac{aC}{\sqrt{\left(\frac{aN_E}{r}\right)^2 + N_I^2}}.$$

This model makes strong predictions about how attention should affect sensitivity to a signal in a population of cells. A noise-reduction mechanism would significantly improve the quality under conditions of high external noise by reducing responses to the noise, but in the absence of noise, a noise-reduction mechanism would have nothing to do and would lead to no net benefit. This pattern may be contrasted with the effects of a non-selective amplification mechanism, which will improve visual processing in low levels of external noise and comparatively high levels of internal noise: This is achieved by boosting the representation of both signal and external noise, such that internal noise has less of an impact. However, amplification will be ineffective when external noise levels are higher than internal noise, as both the signal and noise will be increased, leading to no net benefit.

In order to assess the correspondence between this model's predictions and fMRI responses, Pratte et al. (2013) adopted a multivariate pattern classification approach to measure the amount of stimulus information contained at multiple levels of the cortical visual hierarchy about the viewed stimulus category. The accuracy of fMRI classification served as a measure of discrimination sensitivity to link to the predictions of the model (see Fig. 1(B)).

The stimuli were comprised of line drawings of faces, houses, shoes and chairs embedded in varying levels of external noise (Fig. 3(A)). Participants either attended to the stimuli by performing a one-back matching task on the images, or they performed a demanding task on letters presented at fixation, effectively withdrawing attention away from the object images. To measure neural sensitivity to the stimuli in a group of voxels, a multivariate pattern classifier was trained to identify which stimulus type (face, house, shoe or chair) was being viewed in a given block, based on the measured fMRI activity patterns in each cortical visual area. The ability to predict the viewed stimulus category based on activity patterns served as a measure of stimulus-specific sensitivity. This measure of classification accuracy was assessed separately for each external noise level and each attention condition.

Fig. 3(B) shows classification performance for V1 as a function of external noise level, separately for when the stimulus was attended and when attention was withdrawn from the object images. As expected, adding external noise to the stimulus systematically lowered classification performance, regardless of whether the stimulus was attended to or not. The influence of attention, however, depended on the external noise level: Attention enhanced classification performance when stimuli were embedded in high levels of noise, but had no effect for stimuli presented in low or moderate noise levels. This pattern of results implies that a pure noise-reduction mechanism operates in V1: attention is only effective in the presence of high levels of external noise. This can be contrasted with the results found in the fusiform face area (FFA, Fig. 3(C)), an area in the ventral temporal cortex that shows strong selectivity for faces (Kanwisher, McDermott, & Chun, 1997). Unlike V1, attention increased the sensitivity of the FFA to stimuli presented in both low- and high-noise regimes. The attentional enhancement observed across all noise levels suggests that both noise-reduction and amplification mechanisms impact the contents of visual representations in these high-level brain regions.

The lines in Fig. 3(B) and (C) show predictions of the fitted PTM model, which captures the effects of both external noise and attention on fMRI classification performance quite well. The model was fitted to data from several brain regions in the visual cortex, providing estimates of noise reduction and amplification in each. Fig. 3(D) and (E) shows the resulting parameter estimates of noise reduction and amplification, respectively. Whereas noise reduction appears to occur throughout the early visual cortex, amplification is absent in early areas V1 and V2, but becomes more prominent in higher-level areas. These results have important implications for how attention may operate to improve visual processing, first by relying on attentional templates tuned to the target to filter out external noise, and subsequently by amplifying these filtered responses to enhance overall responses, effectively overcoming the effects of internal noise.

Despite tackling a very different problem, the methodological approach of Pratte et al. shares similarities with the previous study by Brouwer and Heeger (2011). Both relied on an analytic approach to isolate stimulus-specific components of the fMRI responses, and then compared these observed responses with those predicted by a fitted cognitive model. In Pratte et al.'s study, a multivariate classification approach was used to estimate the amount of stimulus information in a visual area, which could then be mapped onto the predictions and specific parameters of the PTM model. We believe that using multivariate pattern analyses as a transformational step between the dense spatiotemporal information in the original fMRI data and the predictions of a formal model provides a powerful framework for future studies. However, as we will discuss below, there are other promising ways to link fMRI responses with theoretical models.

4. Memory: the context maintenance and retrieval model

Theoretical modeling has played a central role in the study of memory (see Raaijmakers & Shiffrin, 2002 for a review). At the same time, human neuroimaging has been used successfully to examine various mnemonic processes including long-term memory (see Henson, 2005; Wagner et al., 1998), free recall (Polyn, Natu, Cohen, & Norman, 2005) and visual working memory (Ester, Serences, & Awh, 2009; Harrison & Tong, 2009; Marois, 2016; Sprague, Ester, & Serences, 2014; Xu & Chun, 2006). However, bridging these approaches by applying formal memory models to neural data has been largely absent. One reason may be that, unlike the relatively simple normalization and perceptual template models considered above, memory models are often extremely complex. For example, neural network models of memory contain multiple layers, many units and numerous weighted connections, such that its performance cannot be expressed in a closed-form equation. As a consequence, their model predictions can only be approximated via simulation, and their parameters must be

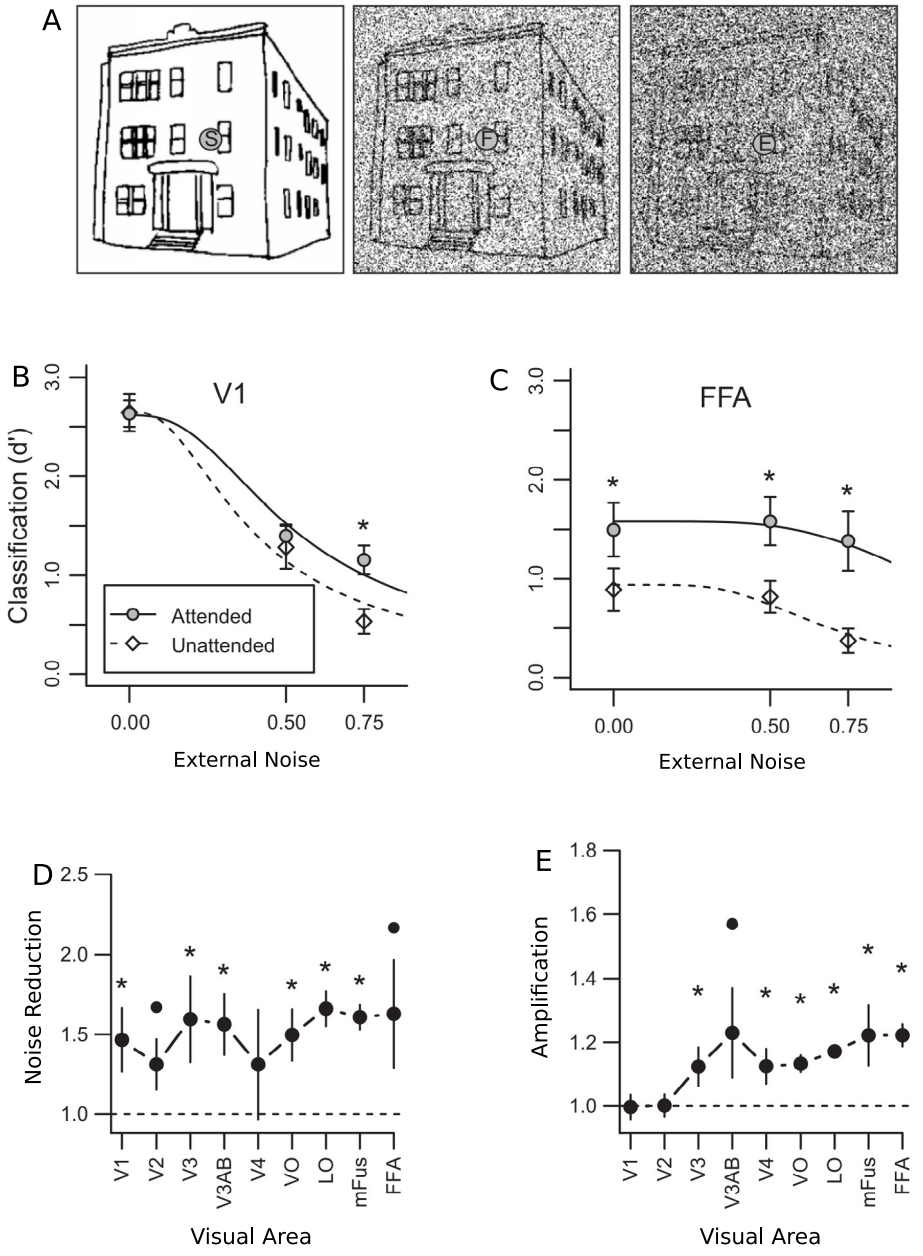

Fig. 3. Perceptual Template Model of Attention. (A) Example of stimuli used in Pratte et al. (2013), where here a house image is embedded in various levels of external noise. (B) Accuracy of a pattern classifier trained on V1 responses to identify the category of a viewed object (face, house, chair or shoe) plotted as a function of external noise level, separately for when the objects were attended (filled points) or un-attended (open points). Lines denote the fitted perceptual template model, and stars indicate significant attention effects at each noise level. (C) Classification accuracy for activity patterns in the FFA. (D) Estimates of noise reduction effects of attention in various visual areas obtained by fitting the PTM model to classification data from various regions of interest, arranged loosely from early visual areas to higher-level visual areas. (E) Estimates of amplification effects of attention.
Source: Figures adapted from Pratte et al. (2013).

fitted with advanced minimization routines such as differential evolution, or after extensive training of the network. With such complexity, applying these models even to simple behavioral data can be challenging, such that the task of linking model predictions to neural data is not straightforward.

Recently however, Kragel et al. (2015) successfully established links between a well-studied model of free recall and fMRI measures of brain activity, providing insights into neural mechanisms underlying memory processing that would not have been possible otherwise. A challenge in understanding human memory is the organization and structure that people naturally impose upon any set of learned items. This structure is particularly evident in studies that allow for free recall, in which participants are presented with a list of items to study and are then asked to recall as many words as they can in whatever order they choose. Several characteristics of free recall data have been pivotal for theory development, such as the prominence of primacy and recency effects. One particularly interesting feature of free recall is the strong influence of temporal context: upon recalling a word at test, people are more likely to recall words that were nearby in the study list (with the subsequent word more likely to be reported than the preceding word). This pattern of behavioral data has been interpreted to suggest that during the study of a word, the characteristics of the word are stored along with the temporal context in which the word was presented. At retrieval, the activation of a memory trace for a particular word will also activate the context with which that item was stored, and

this heightened activity helps bring to mind other studied items that occurred nearby in time.

This notion of temporal reinstatement has been formalized in the Context Maintenance and Retrieval Model (CMR, Polyn et al., 2009), a member of a large class of temporal context models (Howard & Kahana, 2002). According to the CMR model, a studied item can be represented by a feature vector, as is standard in many models of memory. Additionally, a separate vector represents the context at the time of study for a given item. The context surrounding a word at study includes the previously studied words, and will be more greatly influenced by more recent words. Consequently, this context vector changes gradually over the course of the study epoch, and serves to link the memory representations of temporally proximal words. At test, recalling an item will re-activate the context state associated with that item, thus increasing the probability of recalling words that were studied nearby.

Kragel et al. (2015) posited that changes in brain activity during the recall period may reflect these dynamics, driven by two different processes: (1) More or less brain activity in an area might correspond to how well a word is being recalled, termed *retrieval success*, and (2) Activity may be related not to the item strength, but rather to the degree with which the context surrounding the item is being activated, termed *temporal reinstatement*.

There is nothing inherent in the experimental design of a free recall task that can be used to decompose the neural signal into these two separate processes of temporal reinstatement and general memory performance. Instead, Kragel et al. devised an approach of first fitting a standard CMR model to the behavioral data (i.e. the recall events) collected during fMRI scanning, in which parameters were found by maximizing the likelihood of the behavioral data given the model. They then fitted enhanced versions of the model in which the fMRI signal from a given voxel was allowed to modulate a specific parameter of the CMR model over the time course of the recall period (see Fig. 1(C)). A voxel's average response to each recall event could either modulate the overall success rate of memory performance during the course of recall, or it could modify the strength of temporal reinstatement during recall. If the resulting model, informed by this modulatory source of fMRI signal, was more likely to have produced a participant's particular behavioral data than the model without fMRI modulation included, then the voxel was deemed to be related to the particular process (accounting for model complexity using AIC). By allowing the response of each voxel to affect one, both or neither model parameter, the researchers could identify voxels that were involved in memory performance, temporal reinstatement, or both.

The results of this analysis provided novel evidence to implicate distinct regions of the medial temporal lobe in temporal reinstatement (Fig. 4(A)) and retrieval success (Fig. 4(B)). A spatial gradient can be observed along the posterior–anterior axis, whereby more anterior voxels are involved in retrieval success and more posterior voxels co-vary with the level of temporal reinstatement. The authors used a statistical model comparison approach to assess the merit of various models, to determine which regions of interest were related to retrieval success (RS), temporal reinstatement (TR), or both processes. The model comparison results further supported the notion of a posterior–anterior gradient both in the hippocampus (Fig. 4(C)) and the medial temporal cortex (Fig. 4(D)). Taken together, these results suggest that when posterior regions showed increased neural activity, there was an increased likelihood that the next recalled item would be one that was studied nearby in time. In contrast, activity in anterior regions did not relate to the temporal organization of recalled items; instead, lower activity indicated a failure in retrieval success.

The ability to tease apart these cognitive processes relied on incorporating fMRI activity within a theoretical model of contextual reinstatement. The result of doing so provides new evidence for the role of the medial temporal lobe in episodic memory and free recall, by linking BOLD activity in these regions to formalized computational mechanisms of temporal context. These results also provide support for the CMR model: If the model were poorly aligned with cognitive/neural processes, then it is unlikely that its parameters would map onto brain responses in such a structured and coherent manner. Other temporal context models may fair better or worse at linking the BOLD signal to behavior, and the modeling approach developed by Kragel et al. provides a powerful framework for addressing such model-based questions.

The approach taken by Kragel et al. to establish links between a theoretical model and neural data is quite different than those discussed previously, where a model was fitted to amplitudes of fMRI population responses or to the accuracy of fMRI decoding. Instead, Kragel et al. incorporated fMRI signals to modulate latent processes within a model of behavioral data, and identified voxels for which doing so led to better behavioral predictions. This approach serves as a promising framework for future studies, and provides a compelling demonstration that there are many ways to accomplish a model-based approach to cognitive neuroscience.

5. Categorization: exemplar vs. prototype models

In the studies considered so far, a specific model was applied to some combination of behavioral and neural data, and the resulting fits and parameter estimates were used to make inferences about the underlying mechanisms. However, formal theoretical modeling provides for a more powerful approach of fitting multiple models to the data, such that the merits of different cognitive theories can be compared quantitatively. Admittedly, the task of developing multiple models to link to neural data presents a greater challenge, especially to ensure judicious and equitable implementations of each model. However, a recent study by Mack et al. (2013) demonstrates how doing so can lead to powerful ways to test cognitive theories.

People are highly adept at learning categorical boundaries in multidimensional spaces, and can accurately categorize new instances according to these learned categories. There are two prominent classes of models that describe how people might do this, namely prototype and exemplar-based models (see Ashby & Maddox, 2005 for a review). Prototype theories posit that as we encounter multiple instances from different categories, we construct a *prototype* for each category. For example, the prototype may be constructed as the average of all stimuli experienced in that category. Upon seeing a new exemplar of a category, the values of its features are then compared with the prototypes for each category. Greater similarity (or proximity) to one of the prototypes then provides the basis for categorization of the novel exemplar. In contrast, exemplar-based theories posit that people store particular instances from each category as they are experienced. Upon seeing a new instance that is to be categorized, the values of its features are compared with all previous exemplars stored in memory for each category, and the current instance is categorized based on its similarity with all exemplars stored in memory.

Mack et al. created stimuli that took on one of two values of color, shape, size, and spatial location. The set of possible stimuli were divided into two categories, such that the average values of each stimulus dimension varied across categories, and provided reasonable prototypes of each category. Participants were trained to categorize a subset of the stimuli, and were then tested with new instances from each category while undergoing fMRI scanning. According to prototype theories, the degree to which a new

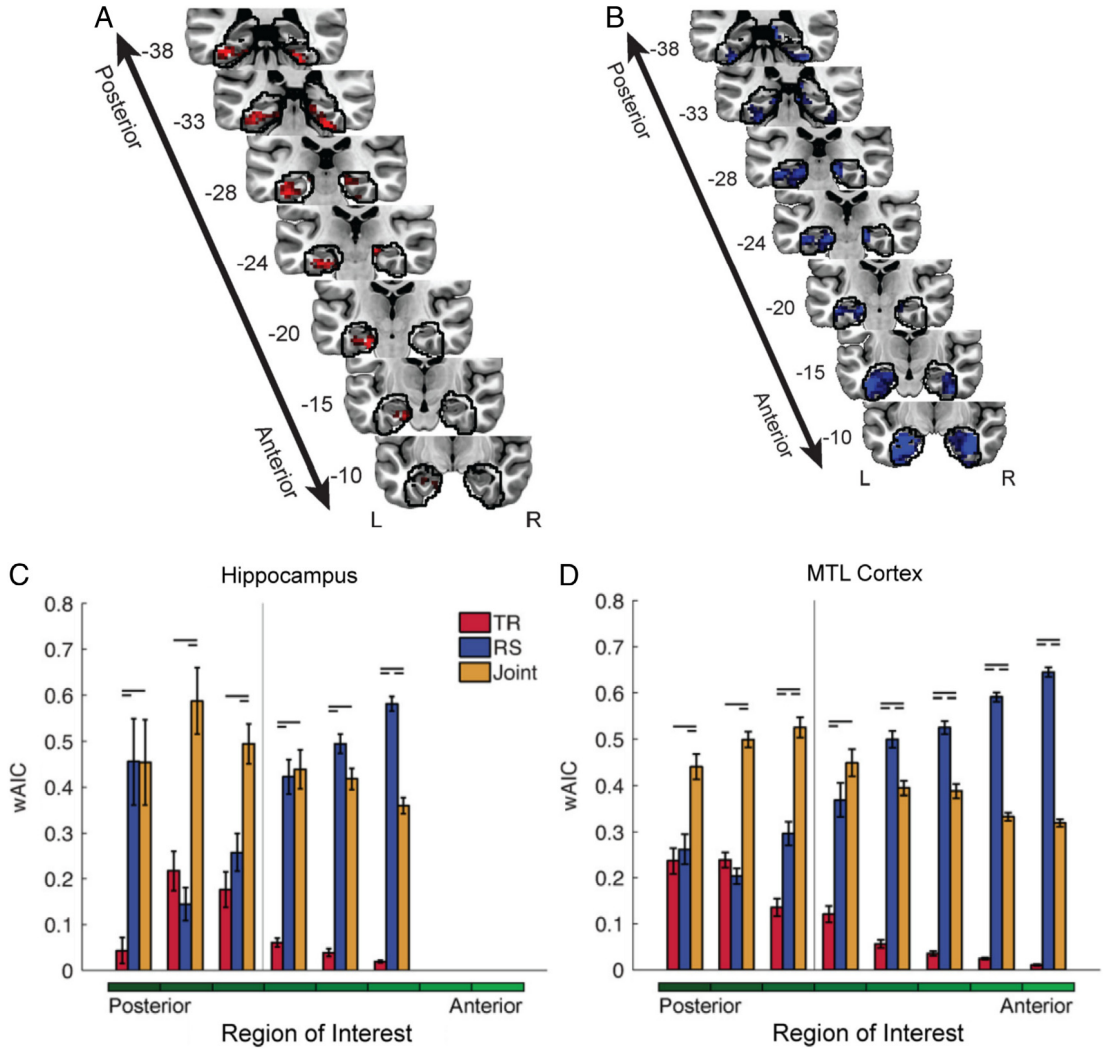

Fig. 4. Context Maintenance and Retrieval Model of Memory. (A) Voxels that lead to a better-fitting behavioral model when their signals modulate the temporal reinstatement parameter of the model. (B) Voxels that produce a better behavioral model when they modulate the *retrieval success* parameter. (C) Results of an ROI analysis in which the ROIs are ordered from their posterior-to-anterior position within the Hippocampus. Bars denote the relative evidence (weighted AIC) for each of the three fitted models: (1) a model in which the ROI signal modulates the *temporal reinstatement* parameters (TR, red), (2) the ROI signal modulates the *retrieval success* parameters (RS, blue), or (3) the ROI signal modulates both parameters (Joint, orange). (D) Same ROI analysis as in (C) but for ROIs within the medial temporal lobe. (Color figures available in the online version of this article.)
Source: Figures adapted from Kragel et al. (2015).

instance belongs to each category is based on how similar the new instance is to the average, or prototype, stored in memory for each category. If x_{km} denotes these averages for the kth category and mth feature dimension, and the feature values of a new item at test are y_m for the mth feature dimension, then the similarity between the item and the kth category prototype is

$$s_k = \exp\left(-c\sum_{m=1}^{4} w_m|y_m - x_{km}|\right)$$

where parameter c controls how quickly similarity decreases as the distance between a prototype and a new item increases, and parameters w_m allow for particular dimensions to be differentially weighted. For example, if more attention were paid to one feature dimension over the others, leading to greater weighting of that feature, then that feature would have a greater influence on the similarity evaluation (s_k) between an item (y_m) and the prototypes (x_{km}) for each category. The exponential function leads to a similarity value of 1.0 when the instance matches the prototype exactly, and the exponential decay of this similarity metric provides for a generally good characterization of human similarity judgments (Shepard, 1987).

The exemplar-based model follows a similar structure, however similarity is based on the summed distance between a new instance and all previously stored items in a category, rather than distance to the average item. If x_{jm} are the mth feature values for the jth previously studied item, then the total similarity between a new item y_m and the Kth category can be specified by

$$s_k = \exp\left(-c\sum_{j\in K}\sum_{m=1}^{4} w_m|y_m - x_{jm}|\right)$$

where parameters and w_m govern overall similarity and feature weighting, respectively.

For each item to be categorized at test, these equations describe the model predictions regarding the similarity (or *representational match*) between the representation of that item and that of each category. These similarity values can be used to compute a predicted probability that an item is placed within each category, and these predictions can be compared with observed categorization performance to evaluate the predictive power of the models. In their experiment, Mack et al. (2013) found that both models fitted the participants' behavioral performance equally

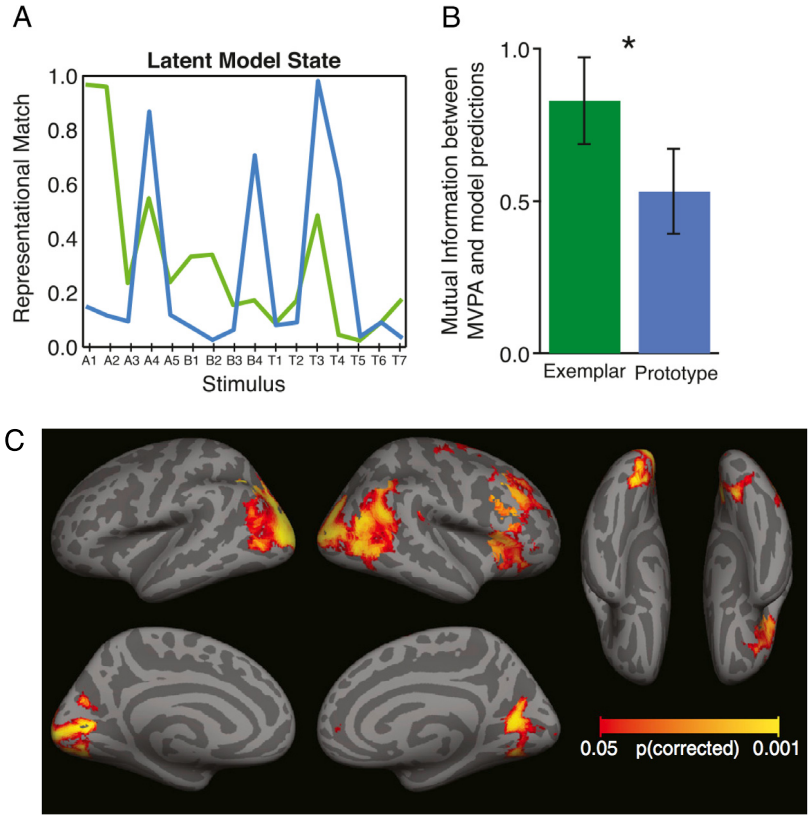

Fig. 5. Exemplar and Prototype Models of Categorization. (A) Categorization models predict a certain amount of representational match between an internal representation and a particular item from a learned category. Here the match for the exemplar (green) and prototype (blue) models are shown for various items used by Mack et al. (2013). (B) The level of agreement (measured as mutual information) between the representational match as predicted by each model, and the level of representational match as measured in the fMRI signal. Values of zero would indicate no correspondence between a model's predictions and the data; larger values indicate a higher correspondence. (C) Searchlight analysis to identify areas where the differences in brain activity to the different items at test were well predicted by the exemplar model. (Color figures available in the online version of this article.)
Source: Figures adapted from Mack et al. (2013).

well, as is often the case (Rouder & Ratcliff, 2004). Although the prototype and exemplar models predict different representational match values for each item (Fig. 5(A)), the coarseness of the behavioral measure (category A or category B) appeared to lack the sensitivity to discriminate between the more subtle differences between the models in their representational match predictions.

Mack et al. (2013) evaluated whether neural activity as measured with fMRI multivariate pattern analysis might provide a more sensitive measure of variations in representational match in this categorization task. To do so, they first fit the prototype model to a participant's behavioral data, and used the model to label each test trial by its predicted representational match value (Fig. 5(A)). In a cross-validation procedure, they applied a regression analysis to a subset of the data, mapping the weighted sum of voxel activity onto the model-defined representational match value for each trial (using 1000 voxels chosen from the entire brain using a localizer analysis). This fitted linear model was then used to predict the representational match values of trials not included in the training data set, based on the corresponding voxel activity patterns. This procedure was repeated such that representational match values were predicted for every trial, based on the neural signal. The accuracy of these predictions served as an index of how closely the neural activity patterns matched the prototype model's definition of representational match. Prediction accuracy above chance would indicate that the neural signal carries category information as defined by the prototype model. Critically, the same procedure was conducted for the exemplar model, providing prediction accuracy for representational match values as defined by distance from the exemplars. Comparing how well the neural data could be used to predict match values as defined by the two models provides a way to determine which model more closely matches the neural representation (see Fig. 1(D)).

The correspondence between model-based representational match values and those predicted from brain activity was captured by a measure of mutual information between predicted and observed values. This mutual information metric was significantly greater than zero for both models (Fig. 5(B)), suggesting that brain activity (from voxels chosen across the entire brain) conveys some information about the similarity between a presented item and the previously learned category representations as defined by either model. Critically however, this measure was higher for the exemplar model than for the prototype model, implying that the exemplar model provides a better description of category representations in the human brain. A follow up searchlight analysis was conducted to identify brain areas that showed activation differences to different items that were in line with item differences predicted by the exemplar model (Fig. 5(C)). The results indicated that regions of the lateral occipital complex, inferior parietal lobe, as well as the right inferior frontal lobe, exhibited responses consistent with the predictions of the exemplar-based model.

These results underscore the potential power of combining theoretical models with neuroimaging techniques. Although these models make unique predictions about behavioral data, the resolution of such data is not always sufficient to provide convincing evidence in favor of one model or another. By using multivariate pattern analysis to construct an appropriate link between the neural data and model parameters, the cortical

activity patterns observed in this study provided a measurement that could be used to adjudicate between the categorization models. The ability to make a coherent link between theories and brain imaging data underlies the success of this and other model-based neuroscience approaches. Mack et al. (2013) demonstrate that doing so can provide insights into processing that would not have otherwise been possible.

6. Error monitoring in the stop signal task

A variety of cognitive functions have been ascribed to the dorsal anterior cingulate cortex (dACC), based on studies conducted using fMRI, EEG, and other methods. Prominent theories have proposed roles for the dACC in error detection, conflict monitoring, error anticipation, volatility monitoring, action–outcome learning, reward, and the coding of errors in predicting future outcomes, to name a few (see e.g. Botvinick et al., 2004; Hayden, Heilbronner, Pearson, & Platt, 2011; Rushworth, Walton, Kennerley, & Bannerman, 2004). Teasing apart which of these roles the dACC is primarily responsible for during a task is extremely challenging, as most tasks that are designed to tap into one of these processes may also evoke some combination of the others. For example, tasks designed to elicit errors often do so by incorporating rare events within the experimental design, such that events that produce behavioral errors are also rare. Consequently, brain responses to these rare events might reflect the fact that they are unexpected, or alternatively, because they tend to elicit error responses. Fortunately, formal cognitive models have been devised that can provide separate measures of these various processes, based on the structure of the task and the resulting behavioral data. The key to determining how these processes map on to particular brain regions, such as the dACC, is to somehow leverage the models in order to isolate the neural activity associated with specific components of these overlapping mental processes.

Ide et al. (2013) set out to isolate these often co-occurring mental processes, such as error detection and conflict monitoring, in order to determine what brain areas are involved in which particular functions. They utilized a classic cognitive control task, known as the stop-signal task (Logan, Cowan, & Davis, 1984; see Verbruggen & Logan, 2008), in which participants make a button press in response to a "go" signal, such as a green dot. On a minority of trials, however, a "stop" signal (e.g. a red dot) appears some time following the go signal, instructing participants to refrain from making a response to the preceding go stimulus. Replicating previous results, Ide et al. (2013) found that the dACC as well as many surrounding regions showed a stronger BOLD response on stop trials than on go trials (yellow/orange colors in Fig. 6). However, this heightened activity might reflect any combination of many underlying cognitive processes. For example, higher activity on stop trials might reflect the processing of conflict between the prepotent "go" response and the required stopping behavior, or it might reflect the experience of *surprise* associated with the rare event of a stop trial.

The authors applied a Bayesian rational decision-making model to account for behavioral performance in the stop-signal task, with dynamic updating based on trial-by-trial learning (Shenoy, Angela, & Rao, 2010; Shenoy & Yu, 2011). The degree to which a participant expects a stop signal to occur on a given trial can be specified by a latent variable (P_{stop}) that varies from trial to trial depending on the stimulus history and the participant's past behavior. For example, following a long sequence of go trials, people tend to make progressively faster responses. This behavior is thought to reflect a lowered expectation of an impending stop trial after each successive go trial, such as by the buildup of an automatic, prepotent response. In the model this behavior is instantiated as a relaxing of the decision criterion to make a "go" response, and this

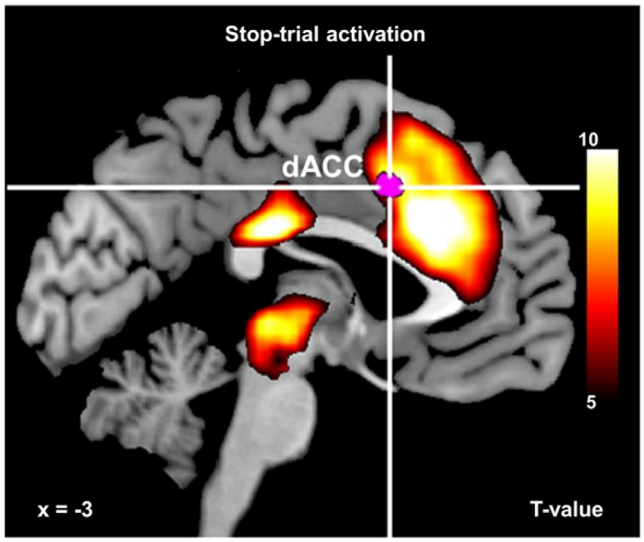

Fig. 6. Bayesian Model of Stop Signal Task. Contrasting the BOLD signal to identify regions of greater activity for stop than for go trials (red/orange) reveals extensive activation in both cortical and sub-cortical regions. The more nuanced model-based analysis, however, reveals that only a small region (magenta color) within the dorsal anterior cingulate cortex (indexed by cross hairs, labeled dACC) is related to the process of comparing expected outcomes with actual outcomes, termed *surprise*. (Color figures available in the online version of this article.)
Source: Figure adapted from Ide et al. (2013).

dynamic latent process is measured by parameter P_{stop}. Another parameter, P_{error}, reflects the likelihood that a participant will make an error on the upcoming trial, dependent upon the recent stimulus history and the participant's past behavior. By specifying how these and several other latent processes are linked to the behavioral data, the model is able to predict the engagement of these processes for each trial across an experiment.

Ide et al. (2013) fit this model to behavioral data collected while participants underwent fMRI scanning. The resulting fitted model provided trial-by-trial predictions regarding the levels of multiple latent constructs, such as P_{stop} and P_{error}. These values were then used as independent variables in a multiple regression model, in which the BOLD signal from individual voxels served as the dependent variable. The results of the regression analysis could then be used to identify voxels that co-varied with each particular latent process, such as the participant's expectations of a stop trial (see Fig. 1(E)). One critical analysis involved comparing the participant's presumed expectations about a trial, as specified by the model, with the actual outcome of the trial. In particular, the difference between the latent parameter P_{stop} and the actual stimulus condition (stop vs. go) on each trial provided a measure of the degree of mismatch between participant's expectations and the actual stimulus outcome, termed *surprise*. The magenta color in Fig. 6 shows voxels that were modulated specifically by this measure of *surprise*, and the results of this whole-brain analysis revealed that voxels sensitive to *surprise* are located squarely within the dACC (marked by cross hairs). This analysis suggested that the dACC's sensitivity to *surprise* on stop trials is independent of other cognitive processes associated with stop trials, such as cognitive control and error monitoring. In fact, the researchers performed several additional analyses to determine what other factors may co-vary with dACC activity, but found none.

The results of this fMRI modeling work suggest that the dACC is involved in comparing our expectations about the environment with actual outcomes, and that accounting for this cognitive process may preclude the need to invoke additional cognitive processes such as error monitoring or response conflict. Because these processes are so commingled in typical tasks,

contrasts between conditions in previous studies likely reflect a conglomeration of cognitive processes. However, Ide et al.'s use of a theoretical model in conjunction with fMRI provided a more nuanced and powerful approach for ascribing particular functions to the anterior cingulate.

7. Discussion

Human cognitive neuroscience and theoretical modeling of cognitive processes have led to remarkable advances in our understanding of mental function, but largely as isolated endeavors. The integration of these approaches is a natural step forward, and the recent proliferation of review articles (Forstmann, Wagenmakers, Eichele, Brown, & Serences, 2011), books (Forstmann & Wagenmakers, 2015), and special journal issues speaks to the growing interest in model-based cognitive neuroscience. Establishing a link between a model's constructs and non-invasive measures of human brain activity presents many challenges. The studies we have reviewed here highlight the diversity of recent approaches to meet these challenges, and we take their success as evidence that there are many promising paradigms for building a model-based cognitive neuroscience.

The studies reviewed here differ not only in the particular methods used, but differ in more fundamental ways with respect to how the models and neural data were combined (see Fig. 1). For example, some studies fitted a theoretical model directly to the neural data, and used the resulting parameter estimates and model fits to learn about the underlying psychological and neural processes (Brouwer & Heeger, 2011; Pratte et al., 2013). Other studies first fitted a model to the behavioral data and then used the resulting fitted model to make inferences about the neural signals (Ide et al., 2013). The opposite approach has also proven effective, in which fMRI signals from a voxel or brain region of interest are incorporated within a model in order to make better predictions of behavioral performance (Kragel et al., 2015). Finally, Mack et al. (2013) fitted multiple models to the behavioral data, and determined which model produced the highest correspondence with patterns of brain activity. The approaches taken in these studies is not an exhaustive list, but the diversity clearly shows that there are many possible paths for integrating theoretical modeling and cognitive neuroscience. Going forward, the development of new ways to link models and neural data has great potential for advancing our theoretical and empirical understanding of brain function. Although we are hesitant to make strong prescriptions for how one ought to approach model-based neuroscience, it may be helpful to consider the similarities of an approach with previously successful efforts. Although no two studies reviewed here completely converged in their methods, there are some commonalities among them that we highlight below as promising avenues for future work.

One approach for establishing links between high-dimensional fMRI data and parameters of interest in a cognitive model is to construct a lower-dimensional measure of the information contained in fMRI activity patterns. This approach was taken in the studies of the normalization model of perception (Brouwer & Heeger, 2011) and the perceptual template model of attention (Pratte et al., 2013). Both studies utilized multivariate pattern analyses as an intermediary step to link the multivariate BOLD responses with the theoretical model. For example, Pratte et al. (2013) applied pattern classification to determine how well the activity patterns in individual visual areas could predict which of four possible object categories the participant was viewing on individual stimulus blocks. The resulting prediction accuracy provided a measure of how robustly an object was represented within a particular visual area, which is precisely the targeted construct for the theoretical model. Adopting a multivariate approach in this way necessitates several considerations. First, the stimulus, experimental paradigm, and task may require some modifications to be well-suited for multivariate pattern classification (see Tong & Pratte, 2012 for details). In addition, special care must be taken to ensure that the results of such a complex intermediary analysis serve to measure the intended construct. In the studies considered here that leveraged multivariate pattern analysis (Pratte et al., 2013), multivariate regression (Mack et al., 2013) or a multivariate forward-modeling approach (Brouwer & Heeger, 2011), such considerations were built into the design and analysis of the experiment. However, demonstrating that these complex analysis methods are measuring the neural information that they are intended to extract will be critical for future studies that utilize such techniques. Nonetheless, we believe that multivariate approaches such as pattern classification analyses (Kamitani & Tong, 2005; Norman et al., 2006; Tong & Pratte, 2012) and forward modeling approaches (Brouwer & Heeger, 2009; Kay et al., 2008) will provide for powerful ways to link high-dimensional neuroimaging data with model parameters that isolate important cognitive constructs.

Another commonality among recent model-based neuroscience work is the use of temporally dynamic models. For example, in the work of Kragel et al. (2015), the Context Maintenance and Retrieval model specifies latent constructs such as the strength of temporal reinstatement over the course of free recall. By incorporating the amplitude of the fMRI signal on each trial into specific parameters of the model, they could determine whether the neurally informed model provided a better prediction of free recall performance. Models such as this, in which processes are proposed to vary over the course of an experiment, provide excellent candidates for a model-based neuroscience. For such models, the behavioral data alone is often not sufficient to identify how processes might vary over time. However, the corresponding neural measurements provide potentially informative information about the temporal dynamics of such latent processes. Although some dynamic cognitive phenomena occur on slow time scales that can be measured with fMRI (e.g., Davis, Love, & Preston, 2012; O'Doherty et al., 2003; Serences & Saproo, 2009), such as variations in performance across trials (e.g., Ide et al., 2013), the sluggish nature of the hemodynamic BOLD response may limit the ability to study fast cognitive processes that typically occur over a timeframe of a few hundred milliseconds. However, other methods that measure neural signals in real time, such as EEG and MEG, provide a promising avenue to investigate faster dynamic processes in the same manner.

Formal models have provided many insights into cognitive processes, but progress can sometimes stall following the development of more advanced and nuanced models, when the behavioral data fails to clearly distinguish between competing models. The comparison of prototype and exemplar models of categorization is a striking example, as these models aim to explain the same phenomena, and differ only subtly in the nature of their underlying latent representations. This feature has made the models particularly difficult to differentiate using behavioral measures alone, as the latent constructs produce similar behavior. However, by using the fMRI signal to tap into the underlying latent representation, Mack et al. (2013) were able to compare the models with a granularity not possible by studying behavior alone. We doubt that this is a special case, but rather, we suspect that neural measurements might be informative to the study of a variety of proposed latent constructs that are difficult to measure accurately with behavior alone. The challenge will be to design paradigms and analytic approaches that link neural signals to the proposed latent process in a testable manner. The studies reviewed herein provide evidence that doing so is highly feasible.

It is often said that no model is correct, but that some models are better than others, and among the better ones, some may be

considered adequate with respect to their predictive power and the plausibility of their assumptions. Each of the studies reviewed here included various checks to ensure that the utilized model was at least *adequate*, such that combining the model with neural data allowed for clearly interpretable inferences to be drawn. These inferences would not have been possible with a poorly specified model, and the often-demonstrated match between predicted and observed neural data speaks to the potential power of these models. Looking forward, we believe that one of the most promising avenues for a model-based neuroscience will lie in using cognitive neuroscience data in order to test, constrain, and compare different models, with the eventual goal of developing the next generation of better theoretical models.

Our review was selective, aiming to demonstrate the breadth of topics for which a model-based cognitive neuroscience has proven successful, and the diversity of approaches taken. Other notable achievements in combining theoretical models with functional imaging have been made in recent years, with respect to the study of information accumulation (Turner, van Maanen, & Forstmann, 2015), reinforcement learning (Badre & Frank, 2012; O'Doherty et al., 2007), and with complex cognitive architectures such as ACT-R (Borst & Anderson, 2014). In this review, we chose to focus on fMRI approaches to neuroscience, and on theoretical models of perception and cognition, but it is worth noting that the more general ideas underlying model-based neuroscience have the potential to impact all areas of neuroscience and psychology.

We anticipate that future advances in theoretical understanding will be bolstered by advances in technology. Developments in the methods used to estimate model parameters, such as Bayesian estimation (e.g., Turner et al., 2013), will allow researchers to test more sophisticated models than are currently possible. Likewise, advances in neuroimaging technologies, such as high-field fMRI (Duyn, 2012), will allow for more fine-grained measures of human brain activity. Spurred by progress on multiple fronts, the future of model-based neuroscience holds great promise for linking theoretical processes with their underlying neural instantiations.

Acknowledgments

This research was supported by National Science Foundation grant BCS-1228526 to FT and National Institutes of Health Fellowship F32-EY022569 to MP, and was facilitated by National Institutes of Health center grant P30-EY-008126 to the Vanderbilt Vision Research Center.

References

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*(3), 372–400.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*, 149–178.

Badre, D., & Frank, M. J. (2012). Mechanisms of hierarchical reinforcement learning in cortico-striatal circuits 2: evidence from fMRI. *Cerebral Cortex*, *22*(3), 527–536.

Bonin, V., Mante, V., & Carandini, M. (2005). The suppressive field of neurons in lateral geniculate nucleus. *Journal of Neuroscience*, *25*(47), 10844–10856.

Borst, J. P., & Anderson, J. R. (2014). Using the ACT-R cognitive architecture in combination with fMRI data. In B. U. Forstmann, & E. J. Wagenmakers (Eds.), *An Introduction to Model-Based Cognitive Neuroscience*. New York: Springer.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652.

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: an update. *Trends in Cognitive Sciences*, *8*(12), 539–546.

Boynton, G. M., Engel, S. A., Glover, G. H., & Heeger, D. J. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. *Journal of Neuroscience*, *16*(13), 4207–4221.

Brouwer, G. J., & Heeger, D. J. (2009). Decoding and reconstructing color from responses in human visual cortex. *Journal of Neuroscience*, *29*(44), 13992–14003.

Brouwer, G. J., & Heeger, D. J. (2011). Cross-orientation suppression in human visual cortex. *Journal of Neurophysiology*, *106*(5), 2108–2119.

Buckner, R. L., Bandettini, P. A., O'Craven, K. M., Savoy, R. L., Petersen, S. E., Raichle, M. E., et al. (1996). Detection of cortical activation during averaged single trials of a cognitive task using functional magnetic resonance imaging. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(25), 14878–14883.

Bullmore, E., & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, *10*(3), 186–198.

Buracas, G. T., & Boynton, G. M. (2007). The effect of spatial attention on contrast response functions in human visual cortex. *Journal of Neuroscience*, *27*(1), 93–97.

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, *13*(1), 51–62.

Carrasco, M. (2011). Visual attention: the past 25 years. *Vision Research*, *51*(13), 1484–1525.

Davis, T., Love, B. C., & Preston, A. R. (2012). Learning the exception to the rule: model-based FMRI reveals specialized representations for surprising category members. *Cerebral Cortex*, *22*(2), 260–273.

Duyn, J. H. (2012). The future of ultra-high field MRI and fMRI for study of the human brain. *Neuroimage*, *62*(2), 1241–1248.

Engel, S. A. (2012). The development and use of phase-encoded functional MRI designs. *Neuroimage*, *62*(2), 1195–1200.

Engel, S. A., Rumelhart, D. E., Wandell, B. A., Lee, A. T., Glover, G. H., Chichilnisky, E. J., et al. (1994). fMRI of human visual cortex. *Nature*, *369*(6481), 525.

Ester, E. F., Serences, J. T., & Awh, E. (2009). Spatially global representations in human primary visual cortex during working memory maintenance. *Journal of Neuroscience*, *29*(48), 15258–15265.

Forstmann, B. U., & Wagenmakers, E. J. (2015). *An Introduction to Model-Based Cognitive Neuroscience*. New York: Springer.

Forstmann, B. U., Wagenmakers, E. J., Eichele, T., Brown, S., & Serences, J. T. (2011). Reciprocal relations between cognitive neuroscience and formal cognitive models: opposites attract? *Trends in Cognitive Sciences*, *15*(6), 272–279.

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., & Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(27), 9673–9678.

Glover, G. H. (1999). Deconvolution of impulse response in event-related BOLD fMRI. *Neuroimage*, *9*(4), 416–429.

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.

Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, *458*(7238), 632–635.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430.

Hayden, B. Y., Heilbronner, S. R., Pearson, J. M., & Platt, M. L. (2011). Surprise signals in anterior cingulate cortex: neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. *Journal of Neuroscience*, *31*(11), 4178–4187.

Haynes, J. D., & Rees, G. (2005). Predicting the stream of consciousness from activity in human visual cortex. *Current Biology*, *15*(14), 1301–1307.

Henson, R. (2005). A mini-review of fMRI studies of human medial temporal lobe activity associated with recognition memory. *Quarterly Journal of Experimental Psychology Section B. Comparative and Physiological Psychology*, *58*(3–4), 340–360.

Herrmann, K., Heeger, D. J., & Carrasco, M. (2012). Feature-based attention enhances performance by increasing response gain. *Vision Research*, *74*, 10–20.

Honey, C. J., Sporns, O., Cammoun, L., Gigandet, X., Thiran, J. P., Meuli, R., et al. (2009). Predicting human resting-state functional connectivity from structural connectivity. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(6), 2035–2040.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*, 269–299.

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, *76*(6), 1210–1224.

Ide, J. S., Shenoy, P., Yu, A. J., & Li, C. S. (2013). Bayesian prediction and evaluation in the anterior cingulate cortex. *Journal of Neuroscience*, *33*(5), 2039–2047.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*(5), 679–685.

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*(11), 4302–4311.

Kastner, S., Pinsk, M. A., De Weerd, P., Desimone, R., & Ungerleider, L. G. (1999). Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron*, *22*(4), 751–761.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–355.

Kintsch, W. (1967). Memory and decision aspects of recognition learning. *Psychological Review*, *74*(6), 496–504.

Kragel, J. E., Morton, N. W., & Polyn, S. M. (2015). Neural activity in the medial temporal lobe reveals the fidelity of mental time travel. *Journal of Neuroscience*, *35*(7), 2914–2926.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., et al. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–1141.
Logan, G. D., Cowan, W. B., & Davis, K. A. (1984). On the ability to inhibit simple and choice reaction time responses: a model and a method. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(2), 276–291.
Love, B. C. (2015). The algorithmic level is the bridge between computation and brain. *Topics in Cognitive Science*, *7*(2), 230–242.
Lu, Z. L., & Dosher, B. A. (1998). External noise distinguishes attention mechanisms. *Vision Research*, *38*(9), 1183–1198.
Lu, Z. L., Li, X., Tjan, B. S., Dosher, B. A., & Chu, W. (2011). Attention extracts signal in external noise: a BOLD fMRI study. *Journal of Cognitive Neuroscience*, *23*(5), 1148–1159.
Mack, M. L., Preston, A. R., & Love, B. C. (2013). Decoding the brain's algorithm for categorization from its neural implementation. *Current Biology*, *23*(20), 2023–2027.
Marois, R. (2016). The brain mechanisms of working memory: An evolving story. In P. Jolicoeur, C. Lefebvre, & J. Martinez-Trujillo (Eds.), *Mechanisms of Sensory Working Memory: Attention and Perfomance (XXV)*. Elsevier.
Marr, D. (1982). *Vision. A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Freeman.
McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86.
McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, *4*(4), 310–322.
Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K. M., Malave, V. L., Mason, R. A., et al. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, *320*(5880), 1191–1195.
Moradi, F., & Heeger, D. J. (2009). Inter-ocular contrast normalization in human visual cortex. *Journal of Vision*, *9*(3), 11–22. 13.
Nichols, T., & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical Methods in Medical Research*, *12*(5), 419–446.
Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430.
Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*(2), 266–300.
O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*(2), 329–337.
O'Doherty, J. P., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, *304*(5669), 452–454.
O'Doherty, J. P., Hampton, A., & Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences*, *1104*, 35–53.
Polyn, S. M., Natu, V. S., Cohen, J. D., & Norman, K. A. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science*, *310*(5756), 1963–1966.
Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, *116*(1), 129–156.
Pratte, M. S., Ling, S., Swisher, J. D., & Tong, F. (2013). How attention extracts objects from noise. *Journal of Neurophysiology*, *110*(6), 1346–1356.
Raaijmakers, J. G. W., & Shiffrin, R. M. (2002). Models of memory. In H. Pashler, & D. Medin (Eds.), *Memory and Cognitive Processes*: *Vol. 2. Stevens' Handbook of Experimental Psychology* (third ed.). New York: John Wiley & Sons, Inc.
Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356.
Rouder, J. N., & Ratcliff, R. (2004). Comparing categorization models. *Journal of Experimental Psychology: General*, *133*(1), 63–82.
Rushworth, M. F., Walton, M. E., Kennerley, S. W., & Bannerman, D. M. (2004). Action sets and decisions in the medial frontal cortex. *Trends in Cognitive Sciences*, *8*(9), 410–417.
Serences, J. T., & Boynton, G. M. (2007). Feature-based attentional modulations in the absence of direct visual stimulation. *Neuron*, *55*(2), 301–312.
Serences, J. T., & Saproo, S. (2009). Population response profiles in early visual cortex are biased in favor of more valuable stimuli. *Journal of Neurophysiology*, *104*(1), 76–87.
Sereno, M. I., Dale, A. M., Reppas, J. B., Kwong, K. K., Belliveau, J. W., Brady, T. J., et al. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science*, *268*(5212), 889–893.
Shenoy, P., Angela, J.Y., & Rao, R.P. 2010. A rational decision making framework for inhibitory control. Paper presented at the Advances in neural information processing systems.
Shenoy, P., & Yu, A. J. (2011). Rational decision-making in inhibitory control. *Frontiers in Human Neuroscience*, *5*, 48.
Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.
Sprague, T. C., Ester, E. F., & Serences, J. T. (2014). Reconstructions of information in visual spatial working memory degrade with memory load. *Current Biology*, *24*(18), 2174–2180.
Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, *25*(3), 251–260.
Tavor, I., Parker Jones, O., Mars, R. B., Smith, S. M., Behrens, T. E., & Jbabdi, S. (2016). Task-free MRI predicts individual differences in brain activity during task performance. *Science*, *352*(6282), 216–220.
Tong, F., & Pratte, M. S. (2012). Decoding patterns of human brain activity. *Annual Review of Psychology*, *63*, 483–509.
Turner, B. M., Forstmann, B. U., Wagenmakers, E. J., Brown, S. D., Sederberg, P. B., & Steyvers, M. (2013). A Bayesian framework for simultaneously modeling neural and behavioral data. *Neuroimage*, *72*, 193–206.
Turner, B. M., van Maanen, L., & Forstmann, B. U. (2015). Informing cognitive abstractions through neuroimaging: the neural drift diffusion model. *Psychological Review*, *122*(2), 312–336.
Verbruggen, F., & Logan, G. D. (2008). Response inhibition in the stop-signal paradigm. *Trends in Cognitive Sciences*, *12*(11), 418–424.
Wagner, A. D., Schacter, D. L., Rotte, M., Koutstaal, W., Maril, A., Dale, A. M., et al. (1998). Building memories: remembering and forgetting of verbal experiences as predicted by brain activity. *Science*, *281*(5380), 1188–1191.
Xu, Y., & Chun, M. M. (2006). Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature*, *440*(7080), 91–95.