Novel Tools and Methods

# Inverted Encoding Models Reconstruct an Arbitrary Model Response, Not the Stimulus

Justin L. Gardner[1] and Taosheng Liu[2]

https://doi.org/10.1523/ENEURO.0363-18.2019

[1]Department of Psychology, Stanford University, Stanford, CA 94305, and [2]Department of Psychology, Michigan State University, East Lansing, MI 48824

## Abstract

Probing how large populations of neurons represent stimuli is key to understanding sensory representations as many stimulus characteristics can only be discerned from population activity and not from individual single-units. Recently, inverted encoding models have been used to produce channel response functions from large spatial-scale measurements of human brain activity that are reminiscent of single-unit tuning functions and have been proposed to assay "population-level stimulus representations" (Sprague et al., 2018a). However, these channel response functions do not assay population tuning. We show by derivation that the channel response function is only determined up to an invertible linear transform. Thus, these channel response functions are arbitrary, one of an infinite family and therefore not a unique description of population representation. Indeed, simulations demonstrate that bimodal, even random, channel basis functions can account perfectly well for population responses without any underlying neural response units that are so tuned. However, the approach can be salvaged by extending it to reconstruct the stimulus, not the assumed model. We show that when this is done, even using bimodal and random channel basis functions, a unimodal function peaking at the appropriate value of the stimulus is recovered which can be interpreted as a measure of population selectivity. More precisely, the recovered function signifies how likely any value of the stimulus is, given the observed population response. Whether an analysis is recovering the hypothetical responses of an arbitrary model rather than assessing the selectivity of population representations is not an issue unique to the inverted encoding model and human neuroscience, but a general problem that must be confronted as more complex analyses intervene between measurement of population activity and presentation of data.

*Key words:* computation; feature; fMRI; representation; tuning; vision

## Significance Statement

We recently showed that inverted encoding models conflate signal-to-noise ratio with neural tuning width. Sprague and colleagues argued that despite this short falling, inverted encoding models "assay population-level stimulus representations." However, we show that inverted encoding models reconstruct the model responses, not the stimulus. This is problematic because the model, as we derive here, is only determined up to a linear transform and thus the recovered model responses are only one of an infinite family of equivalent solutions. The approach thus fails to provide a unique assay of population representation. This problem can be circumvented by extending the approach to estimate the probability of different values of the stimulus, thus resulting in an interpretable assay of population representation.

Received September 18, 2018; accepted March 3, 2019; First published March 3, 2019.

The authors declare no competing financial interests.

Author contributions: J.L.G. and T.L. designed research; J.L.G. and T.L. performed research; J.L.G. and T.L. contributed unpublished reagents/analytic tools; J.L.G. and T.L. analyzed data; J.L.G. and T.L. wrote the paper.

There is no cone type in the human retina that responds selectively and uniquely to the color chartreuse. Nor is there a cone type for fuchsia, indigo, ebony, crimson, azure, or cerulean. Not even for the three color primaries: red, green, and blue. Rather, the relative activity of just three different receptor types was hypothesized (Young, 1802), and later validated through color-matching experiments (Helmholtz, 1867), to give rise to the multitude of color sensations. This population code for color contrasts with a pure labeled line hypothesis in which each color sensation would be due to a single class of uniquely devoted neurons (Doetsch, 2000). Even for sensory structures like the olfactory system that maintain strictly segregated connectivity from odorant receptor types in the olfactory epithelium to glomeruli in the olfactory bulb, individual odorants can activate numerous different odorant receptors leading to combinatorial possibilities that allow discrimination of many tens of thousands of different compounds despite there being only a few hundred distinct odorant receptors in humans (Buck, 2004). These key findings in sensory physiology firmly place population coding, that is, the idea that for each distinct sensory percept there is some invariant spatiotemporal pattern of activity that can only be discerned from a population rather than a single neuron, as a fundamental concept of sensory representation.

Recently, it has been proposed that an inverted encoding model approach to analysis of functional imaging data from human cortex can assay such “population-level stimulus representations” (Sprague et al., 2018a). However, here, we show that it is the model assumed in the analysis that is reconstructed, not the stimulus. Moreover, the model is arbitrary in that it is only specified to within a linear transform and thus unsuitable for assaying population representation. Typically, encoding models (Naselaris et al., 2011; Serences and Saproo, 2012) are used as lower-dimensional representations of complex sensory stimuli whose responses are then used as linear predictors of cortical responses. For example, a channel encoding model (Brouwer and Heeger, 2009) is one in which a continuous variable like color (Brouwer and Heeger, 2009, 2013; Yu and Shim, 2017), orientation (Brouwer and Heeger, 2011; Ho et al., 2012; Scolari et al., 2012; Ester et al., 2013, 2015, 2016; Garcia et al., 2013; Byers and Serences, 2014; Chong et al., 2016; Bullock et al., 2017; Yu and Shim, 2017; Liu et al., 2018; Lorenc et al., 2018), direction of motion (Saproo and Serences, 2014; Chen et al., 2015), or spatial location (Sprague and Serences, 2013; Sprague et al., 2014, 2016, 2018b; Samaha et al., 2016; Vo et al., 2017) is conceived of exciting several channels with different selectivity for the variable. To take a specific example, hypothetical orientation channels (channel basis functions) with different preferred orientations but identical bandwidths (typically a sinusoidal function raised to an exponent) are created (Fig. 1). The selectivity of the orientation channels are meant to mimic the known selectivity of individual primary visual cortex neurons (Campbell et al., 1968; Rose and Blakemore, 1974; Watkins and Berkley, 1974; Gardner et al., 1999; Ringach et al., 2002; Finn et al., 2007). For each oriented stimulus that is presented, one can calculate how the hypothetical channels would respond. Across many presentations of different stimuli, a matrix of channel responses is constructed and regression coefficients (weights) can be calculated that best predict each voxels’ response in a functional magnetic resonance imaging experiment. After fitting these regression coefficients on a training dataset, predicted channel responses can be computed by inverting the procedure for some left-out dataset, by multiplying the pseudo-inverse of the voxel regression coefficients with the observed voxel responses. If there is reliable selectivity in the population response for the stimulus variable, the resulting predicted channel responses will exhibit a tuned profile that approximates the channel basis functions built into the analysis.

This approach has been called an inverted encoding model (Sprague et al., 2018a) to emphasize that it is an extension to the more typical approach which uses an encoding model to predict BOLD responses (Dumoulin and Wandell, 2008; Kay et al., 2008; Brouwer and Heeger, 2009) without then inverting the procedure to estimate the model responses. The tuned profiles that inverted encoding models produce have been used to characterize population stimulus representations across different task contexts such as during working memory (Ester et al., 2013, 2015; Foster et al., 2016; Sprague et al., 2016; Yu and Shim, 2017; Lorenc et al., 2018) or comparisons across different allocations of attention (Scolari et al., 2012; Garcia et al., 2013; Sprague and Serences, 2013; Ester et al., 2016). Simulations show that these predicted channel responses can index neural tuning in that the widths of the functions change with the width of the underlying selectivity of neurons in the population. However, the predicted channel response functions also change width as a function of the overall signal-to-noise ratio of the measurement, thus conflating neural selectivity with noise (Liu et al., 2018; Sprague et al., 2018a).

If these predicted model responses are to be taken as measures of population stimulus representations, it raises the question as to what exactly a “stimulus representation” is. A long tradition in physiology has measured neural responses as sensory stimuli are systematically varied to assess the relationship between neural response and stimulus properties. Perhaps the most fundamental relationship is that of the receptive field (Hartline, 1938), which is now commonly used in a stimulus space-referred (rather than the original sensory-organ referred) fashion,

This work was supported by the National Institutes of Health Grant R01 EY022727 (T.L.) and a Low Vision Research Award from Research to Prevent Blindness and Lions Clubs International Foundation (J.L.G.). We thank Eli Merriam for comments on an earlier version of the manuscript and Tony Norcia, Guillaume Riesen, Akshay Jagadeesh, Minyoung Lee, Shaw Hsu, Shihwei Wu, Dylan Cable and the Vision Brunch community at Stanford for helpful discussions.

Correspondence should be addressed to Justin L. Gardner at jlg@stanford.edu. or Taosheng Liu at tsliu@msu.edu

https://doi.org/10.1523/ENEURO.0363-18.2019

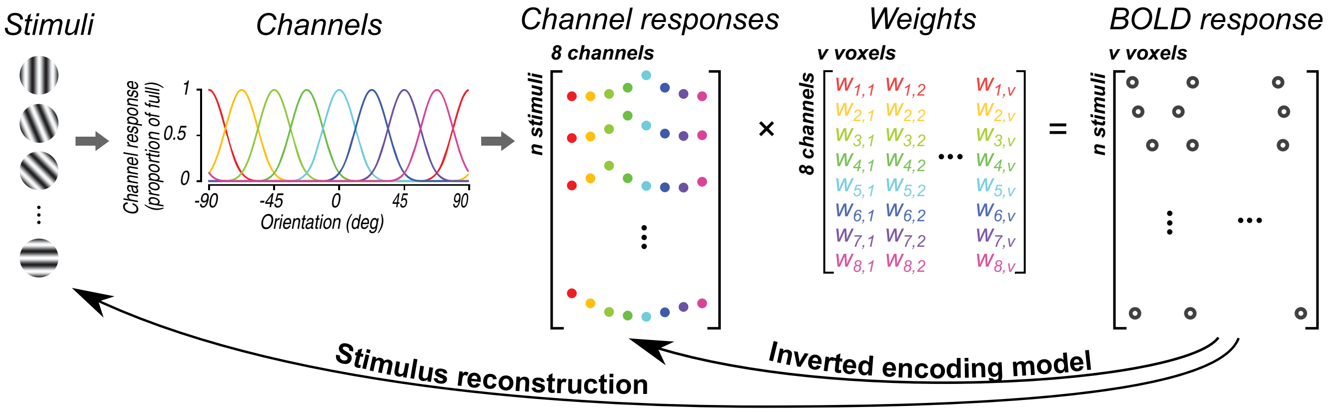
Stimuli
Channels
Channel response (proportion of full)
1
0.5
0
-90
-45
0
45
90
Orientation (deg)
Channel responses
8 channels
n stimuli
×
Weights
v voxels
8 channels
$w_{1,1}$ $w_{1,2}$ $w_{1,v}$
$w_{2,1}$ $w_{2,2}$ $w_{2,v}$
$w_{3,1}$ $w_{3,2}$ $w_{3,v}$
$w_{4,1}$ $w_{4,2}$ $w_{4,v}$
$w_{5,1}$ $w_{5,2}$ $w_{5,v}$
$w_{6,1}$ $w_{6,2}$ $w_{6,v}$
$w_{7,1}$ $w_{7,2}$ $w_{7,v}$
$w_{8,1}$ $w_{8,2}$ $w_{8,v}$
=
BOLD response
v voxels
n stimuli
Inverted encoding model
Stimulus reconstruction

**Figure 1.** Overall schematic of the channel encoding model and its applications. A number of stimuli varying along a dimension of interest (in this case, orientation) are presented ("stimuli") and neural responses are measured. The measured neural responses are assumed to reflect summed activity from a set of underlying mechanisms ("channels"), which are characterized by basis functions that resemble tuning curves of sensory neurons. Each channel's response to each stimulus can be calculated based on the channel's basis function ("channel responses"). These channel responses are multiplied by a weight matrix ("weights") that reflects the relative contribution of each channel in each voxel (i.e., $w_{i,j}$ is the contribution of $i^{th}$ channel in $j^{th}$ voxel). The weighted sum of the channel responses produces the measured neural response ("BOLD response"). By calculating the weights and inverting the model on independent datasets, the inverted encoding model recovers a set of channel responses, whereas by taking into account the structure of the model, one can also reconstruct the stimuli that most likely generated the measured neural responses. To facilitate visualization, each channel and its associated responses and weights are depicted in a different color.

as when it describes the location within the visual field from which a response can be elicited. As physiologists discovered more complex response properties of single neurons to stimulus features such as orientation (Hubel and Wiesel, 1959, 1962), it became common to characterize neural tuning functions. That is, the response as measured as a function of parametric variation of a stimulus, such as orientation (Campbell et al., 1968; Rose and Blakemore, 1974; Watkins and Berkley, 1974; Gardner et al., 1999; Ringach et al., 2002; Finn et al., 2007). Tuning functions have been used to characterize the stimulus representation not only by the firing rate of single-units, but also by other neural measures such as membrane potentials (Finn et al., 2007; Priebe and Ferster, 2012), EEG potentials (Maffei and Campbell, 1970; Regan and Regan, 1987; Baker et al., 2011), reflectance changes from intrinsic signals (Grinvald et al., 1986; Swindale et al., 2003), fluorescence signals from voltage-sensitive dyes (Benucci et al., 2007; Chen et al., 2012), and calcium-imaging measurements (Ohki et al., 2005). Even for BOLD activity averaged across a visual area, parametric sensitivity to the strength of a visual stimulus can be assessed by plotting response magnitude as a function of stimulus properties like contrast (Tootell and Taylor, 1995; Boynton et al., 1996, 1999; Tootell et al., 1998; Logothetis et al., 2001; Avidan et al., 2002; Olman et al., 2004; Gardner et al., 2005; Pestilli et al., 2011) or motion coherence (Rees, 2000; Braddick et al., 2001; Costagli et al., 2014; Birman and Gardner, 2018), which are expected to result in monotonic increases in response of all neurons in a population. Typical for all of these characterizations of stimulus representation is that they report a measurement of neural activity as a stimulus property is systematically varied. Some tuning functions may be derived through a number of analytic steps, such as when computing a tuning function (DeAngelis et al., 1993; Gardner et al., 1999) from a reverse-correlation mapped receptive field profile (Jones and Palmer, 1987) or when Fourier components are computed in a frequency-tagged EEG measurement (Regan and Regan, 1987; Baker et al., 2011; Tsai et al., 2012; Verghese et al., 2012). Nonetheless, the interpretation is straight-forward: the representation characterizes neural response as a function of stimulus variation.

While inverted encoding models can generate a predicted channel response function visually similar to these classically measured tuning functions, the ordinate of the graph is no longer a direct measurement of neural activity. Indeed, a rather odd feature of the literature using inverted encoding model is that there is a lack of consensus over what units to label the ordinate with. It has been alternately labeled as arbitrary units (Brouwer and Heeger, 2011; Ho et al., 2012; Ester et al., 2013; Garcia et al., 2013; Byers and Serences, 2014; Chong et al., 2016), without any specified units (Sprague and Serences, 2013), normalized units (Saproo and Serences, 2014) or in the units of the measurement, for example, as the percentage signal change of BOLD response (Brouwer et al., 2015), or the power of an EEG measurement (Samaha et al., 2016; Bullock et al., 2017), or normalized BOLD (Chen et al., 2015) or BOLD $z$ score (Sprague et al., 2014, 2016, 2018b; Ester et al., 2015; Vo et al., 2017), or relative magnitude (Scolari et al., 2012; Chong et al., 2016; Yu and Shim, 2017). The units of the ordinate are arbitrary in the sense that they can be manipulated by simply changing the maximum response of the modeled channels. Typically set to a unit value, if instead, the maximum channel response is set to two, in the ideal case of no noise in response or measurement, the inverted encoding model will produce predicted channel response functions with doubled height. Making the channel response functions to have a maximum response of forty-two will produce predicted values that will scale accordingly, without any change in the underlying measured responses. Thus, despite being linearly weighted responses, because the maximum channel response can be arbitrarily scaled, the predicted channel response no longer reflects the units of the measurement. Instead, this arbitrary scaling of the ordinate with model assumptions can be avoided by simply plotting the ordinate in proportion or percentage of the full model response (Liu et al., 2018). Because the inverted encoding model is simply a linear regression that attempts to predict channel responses from BOLD responses (Fig. 1), in the limit of no noise, the predicted channel response functions should approach the full amplitude of the model basis functions. Put another way, imagine an encoding model in which one predicts BOLD response magnitude from the age of the subject. If one were to invert this encoding model, then BOLD responses would be used to predict age, and the ordinate would be in units of what is being predicted, years of age, rather than in the units of the predictor, percentage signal change. Viewed as producing proportion of the full model response, the predicted channel response function lies in stark contrast to other tuning functions in which the ordinate is a measurement of neural activity. Thus, the output of the inverted encoding model, i.e., the channel response function, is not a measured response against different stimulus values. Instead, it is the predicted response of a hypothetical modeled channel.

To better explicate the distinction between a classical tuning function and the predicted channel response function, it is instructive to consider a, seemingly, extreme case of poor model specification. We therefore built and tested a channel encoding model on a synthetic data set using published techniques (Liu et al., 2018), except that we changed the channel basis function to have a bimodal shape (Fig. 2*A*). We ran the channel encoding model on simulated data, using procedures identical to those previously reported (Liu et al., 2018). Briefly, the model contained 100 voxels, where each voxel was assumed to contain a random proportion of neurons sampled from a bank of identical, orientation tuned neurons with uniformly distributed orientation preference. Neural tuning functions were circular Gaussians as implemented by von Mises functions. The random proportions in each voxel constitute a weight vector that specifies the contribution of each neuron to the voxel's response. When presented with a stimulus, the response of each neuron was calculated using its neural tuning function, and the response of each voxel was calculated as a weighted sum of the neuronal response according to the voxel's weight vector. Independent Gaussian noise with standard deviation systematically varied to simulate different amounts of noise was added to this response to yield a final response of each voxel. We then simulated an experiment in which eight evenly spaced orientation stimuli were each presented 27 times (Liu et al., 2018) to generate BOLD responses for each trial.

Despite the fact that simulated responses were generated by neurons with unimodal tuning functions, the inverted encoding model with bimodal channels can produce a bimodal channel response function. For example, with a unimodal neural tuning width of 40° (half-width at half-height of the von Mises) and at low noise level (high r$^2$), channel response function had a bimodal shape (Fig 2*B*, right panel), which is expected given that we have shown that the predicted channel response function converges to the channel basis function at low noise level (Liu et al., 2018). We also note that at a higher noise level (low r$^2$), the channel response appeared unimodal (Fig. 2*B*, left panel). Critically, the predicted channel response function does not reflect the underlying neural tuning of the simulated data. The bimodal shape of the predicted channel response function is entirely a consequence of the choice of encoding model basis functions, not of any particular consequence of the modeled responses. This is troubling for an interpretation of the channel response function as a measure of population stimulus representation, because it simply recapitulates the model assumptions, in this case of bimodality, rather than any intrinsic property of the simulated data. While the simulations show that a bimodal

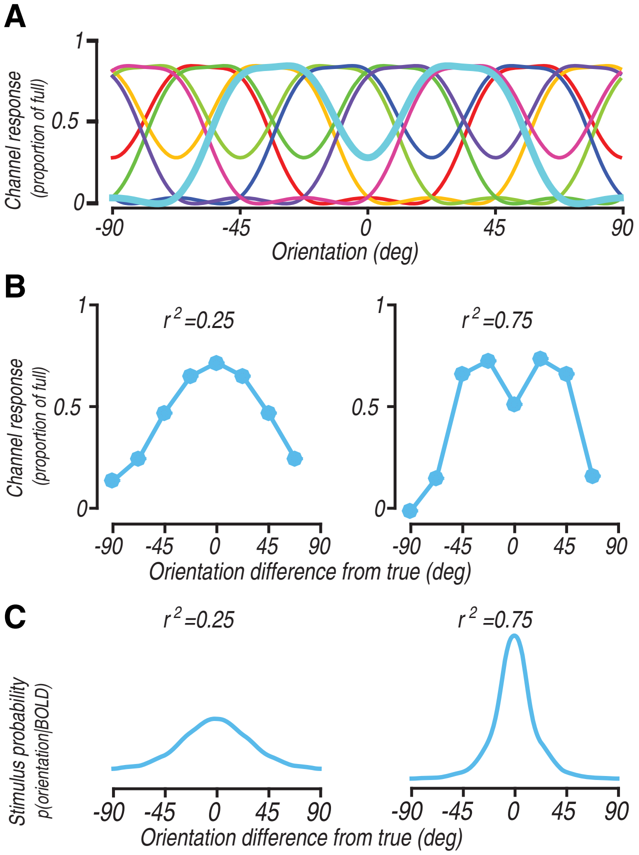

**Figure 2.** Simulation results with a bimodal basis function. ***A***, Depiction of eight channel basis functions, each one with two peaks positioned ~67° apart. To facilitate visualization, the center channel (cyan) is plotted in a thicker line. The channels are obtained by multiplying the original channels (Fig. 1) with a matrix that transforms the unimodal to a bimodal shape. ***B***, Channel response functions derived by the inverted encoding model at high noise (left panel) and low noise (right panel) levels. ***C***, Posterior probability of the stimulus derived by the Bayesian approach at high noise (left panel) and low noise (right panel) levels.

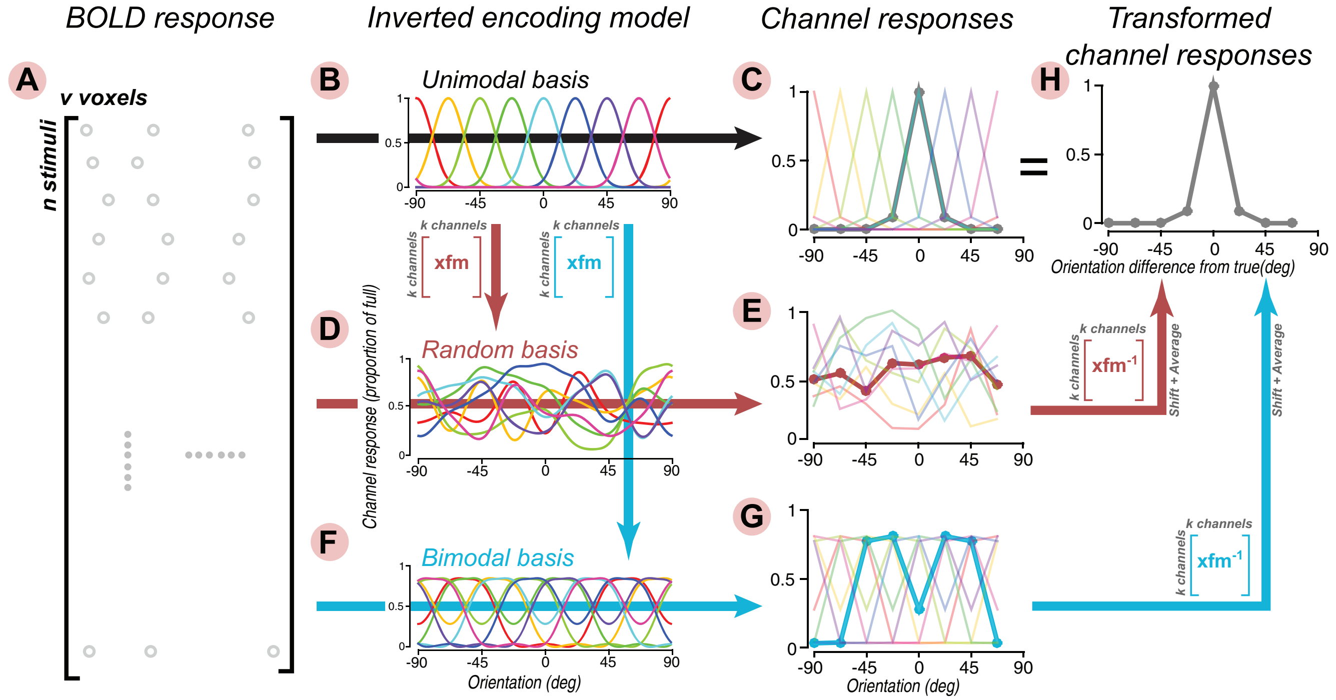

**Figure 3.** Illustration of the behavior of the inverted encoding model under transformed channel basis functions. The simulated BOLD responses (***A***) are generated as before, assuming a set of unimodal neuronal tuning functions. In the first case, standard channel basis functions (depicted in ***B***) are used to estimate the weights and invert the model (depicted by the horizontal arrow), which gives rise to a set of channel response functions (***C***). Here, we depicted both individual channel responses (colored lines) and the shifted and averaged channel response (thick gray line); the latter is typically reported in the literature, and duplicated in ***H***. In the second case, the standard channel basis functions are multiplied by a transformation matrix filled with random numbers (depicted in the red matrix) to generate a set of new basis functions (***D***). After model inversion, individual and averaged channel responses are seemingly random (***E***). In the third case, a set of bimodal basis functions (***F***; same as Fig. 2*A*) were obtained by multiplying the standard basis functions with an appropriate transform (depicted in the blue matrix), which yielded bimodal channel response functions after model inversion (***G***). When the individual channel responses in ***E***, ***G*** are multiplied by the inverse of their respective transforms, shifted, and averaged, an identical channel response is obtained as in the standard unimodal case (***H***). To facilitate visualization, these simulations were conducted assuming zero noise. The same results also hold under non-zero noise conditions.

channel response function emerges as noise is reduced, it would clearly be a mistake to use this analysis and conclude that the population stimulus representation has changed from a unimodal to a bimodal function across these two simulated conditions.

While one might think that the issue is one of poor model specification that could be resolved through appropriate usage of model comparison statistics, it is not. In fact, the amount of variance accounted for by the encoding model using the typical unimodal functions (Fig. 1) and the bimodal functions is identical. Indeed, the bimodal encoding model, though obviously "wrong," was constructed as a linear transform of the "right" unimodal model and thus is mathematically interchangeable (Fig. 3*B*,*F*). More specifically, the unimodal and bimodal channel basis functions were defined as follows:

$$R_1 = SC_1 \tag{1}$$

and

$$R_2 = SC_2$$

where

$$C_2 = C_1P$$

Where the Rs are n × k (n = number of trials, k = number of channels) matrices of channel response functions (Fig. 1, channel responses). The stimuli S are projected onto the channel basis functions C. S is a n x s (s = number of different stimulus types) stimulus matrix with zeros everywhere except for a one in each row at the appropriate column to indicate which stimulus type was presented during that trial. The Cs are s × k matrices which contain channel basis functions in the columns evaluated at each of the stimulus values. The subscripts indicate the unimodal (1) and bimodal (2) channel basis functions. *P* is an invertible channel conversion matrix (k × k) which we have designed to convert the unimodal channel basis functions into bimodal functions. Thus, the channel response matrices for the unimodal and bimodal basis functions are related as follows:

$$R_2 = SC_2 = SC_1P = R_1P \tag{2}$$

By construction then the unimodal and bimodal channel basis functions span the same linear subspace and therefore both encoding models account for the same amount of variance. In fact, the weight matrices for the two models are related by a linear transform. To see this, consider the equations for how the encoding model accounts for

BOLD responses (Brouwer and Heeger, 2009; Serences and Saproo, 2012; Liu et al., 2018):

$$B = R_1W_1 + \eta \quad \text{and} \quad B = R_2W_2 + \eta \tag{3}$$

Where B is a n × v (v = number of voxels) matrix of BOLD responses for all trials, the Ws are k × v weight matrices and $\eta$ is zero mean Gaussian noise. The weight matrices can be estimated using least squares estimation from a training set of BOLD data $B_T$ (Brouwer and Heeger, 2009; Serences and Saproo, 2012; Liu et al., 2018):

$$\hat{W}_1 = (R_1^TR_1)^{-1}R_1^TB_T \quad \text{and} \quad \hat{W}_2 = (R_2^TR_2)^{-1}R_2^TB_T \tag{4}$$

Where the superscript T indicates transpose and –1 indicates inverse. The relationship between the estimated weights for the model with the bimodal basis functions, $\hat{W}_2$, and the unimodal functions, $\hat{W}_1$ can be derived as follows:

$$\hat{W}_2 = (R_2^TR_2)^{-1}R_2^TB_T$$

$$\hat{W}_2 = ((R_1P)^TR_1P)^{-1}(R_1P)^TB_T$$
by substitution of Equation 2

$$\hat{W}_2 = (P^TR_1^TR_1P)^{-1}P^TR_1^TB_T$$
by expansion of transpose

$$\hat{W}_2 = P^{-1}(R_1^TR_1)^{-1}P^{T-1}P^TR_1^TB_T$$
by expansion of inverse

$$\hat{W}_2 = P^{-1}(R_1^TR_1)^{-1}R_1^TB_T$$
multiplication by inverse is identity

$$\hat{W}_2 = P^{-1}\hat{W}_1 \tag{5}$$
by substitution of Equation 4

Thus, in sum, the unimodal and bimodal channel basis functions span the same subspace, account for the same amount of variance in the encoding model, and the estimated weight matrices are related by a linear transform.

In fact, both models will produce identical predictions for stimulus test values that were never even used to train the models. Let $\hat{B}_{1,LO}$ and $\hat{B}_{2,LO}$ be the predicted BOLD responses for the unimodal and bimodal models, respectively, for test stimuli $S_{LO}$ that were left out of the training set. Note that $S_{LO}$will have dimensions $n_{lo}$ × $s_{lo}$ for the number of left out stimuli and the number of types of left out stimuli. The channel basis functions $C_{1,LO}$and $C_{2,LO}$ will have dimensions $s_{lo}$ × k because they are evaluated at each of the $s_{lo}$ left out stimulus values. By Equations 1, 3, the predicted BOLD responses for the unimodal and bimodal models are as follows:

$$\hat{B}_{1,LO} = S_{LO}C_{1,LO}\hat{W}_1 \quad \text{and} \quad \hat{B}_{2,LO} = S_{LO}C_{2,LO}\hat{W}_2 \tag{6}$$

We can show that $\hat{B}_{1,LO}$ and $\hat{B}_{2,LO}$ are equal as follows:

$$\hat{B}_{2,LO} = S_{LO}C_{2,LO}\hat{W}_2$$

$$\hat{B}_{2,LO} = S_{LO}C_{1,LO}P\hat{W}_2$$
by substitution of Equation 1

$$\hat{B}_{2,LO} = S_{LO}C_{1,LO}PP^{-1}\hat{W}_1$$
by substitution of Equation 5

$$\hat{B}_{2,LO} = S_{LO}C_{1,LO}\hat{W}_1$$
multiplication by inverse is identity

$$\hat{B}_{2,LO} = \hat{B}_{1,LO}$$
by substitution of Equation 6

Thus, both encoding models produce exactly the same predictions for BOLD responses even for stimulus test values for which the models were not trained on.

Not only are the unimodal and bimodal encoding models interchangeable and produce identical predictions, the inverted encoding models result in estimated channel response functions that are a linear transform of each other. Consider the way in which channel response functions are estimated from a held-out validation BOLD data set, $B_V$ (Brouwer and Heeger, 2009; Serences and Saproo, 2012; Liu et al., 2018):

$$\hat{R}_1 = B_V\hat{W}_1^T(\hat{W}_1\hat{W}_1^T)^{-1} \quad \text{and} \quad \hat{R}_2 = B_V\hat{W}_2^T(\hat{W}_2\hat{W}_2^T)^{-1} \tag{7}$$

The relationship between the estimated channel response functions using the inverted encoding model with unimodal, $\hat{R}_1$, and bimodal, $\hat{R}_2$, channel basis functions can be derived as follows:

$$\hat{R}_2 = B_V\hat{W}_2^T(\hat{W}_2\hat{W}_2^T)^{-1}$$

$$\hat{R}_2 = B_V(P^{-1}\hat{W}_1)^T(P^{-1}\hat{W}_1(P^{-1}\hat{W}_1)^T)^{-1}$$
by substitution of Equation 5

$$\hat{R}_2 = B_V\hat{W}_1^TP^{-1T}(P^{-1}\hat{W}_1\hat{W}_1^TP^{-1T})^{-1}$$
by expansion of transpose

$$\hat{R}_2 = B_V\hat{W}_1^TP^{T-1}(P^{-1}\hat{W}_1\hat{W}_1^TP^{T-1})^{-1}$$
Interchange transpose and inverse

$$\hat{R}_2 = B_V\hat{W}_1^TP^{T-1}P^T(\hat{W}_1\hat{W}_1^T)^{-1}P$$
by expansion of inverse

$$\hat{R}_2 = B_V\hat{W}_1^T(\hat{W}_1\hat{W}_1^T)^{-1}P$$
multiplication by inverse is identity

$$\hat{R}_2 = \hat{R}_1 P \tag{8}$$

by substitution of Equation 7

Thus, one can take the reconstructed bimodal channel response functions from the inverted encoding model analysis and turn them back into unimodal channel response functions by multiplying them by the inverse of the linear transform used to create the bimodal channel basis functions (Fig. 3*G*,*H*).

As the recovered channel response functions from the inverted encoding model are only constrained up to an invertible linear transformation, the channel response functions can even be converted randomly. As long as the transformation to the random channel basis functions is an invertible transformation, the analysis will result in estimated channel response functions that can be converted through a linear transform back into the unimodal functions (Fig. 3*D*,*E*). Indeed, the channel response functions can be converted between any of the infinitely many equivalent channel response functions related by invertible transforms. In this sense, the particular choice of channel basis functions to display within these infinite possibilities is a completely arbitrary assumption of the analysis and cannot be interpreted as uniquely indicative of the population representation.

This problem of recapitulating the arbitrary model assumptions with an inverted encoding model can be circumvented by using a related Bayesian approach (van Bergen et al., 2015; van Bergen and Jehee, 2018) which computes the posterior probability of the stimulus given the measured responses. The Bayesian approach follows the same structure as an inverted encoding model analysis, but characterizes the residual variance as due to independent, identically distributed noise from the channels and independent and correlated components of voxel noise (for our voxel model we did not simulate correlated voxel noise so we did not fit this component). Having fit both the channel model and the noise, the probability of producing any particular response given a stimulus can be computed. Using Bayes' rule and a uniform prior, the posterior probability of any stimulus given a particular response can then be computed. Using this approach with the exact same simulated data and bimodal encoding model, we found a posterior always centered at the actual stimulus orientation, with its spread reflecting the uncertainty (Fig. 2*C*). Similar behavior was observed over a range of combinations of parameters. This approach highlights a useful interpretation of these model responses. The posterior function represents what probability one could guess the stimulus orientation after having observed a BOLD response. The wider the function, the more uncertain the stimulus orientation is. Notably, the approach yields a unimodal posterior function regardless of whether the channel basis functions are unimodal (van Bergen et al., 2015; Liu et al., 2018) or bimodal as simulated here. This is a sensible outcome as it shows the peak probability at the actual stimulus orientation which decays uniformly around that orientation.

The reason for this striking difference in which the Bayesian approach produces a unimodal posterior and the inverted encoding model yields a bimodal channel response function is simply because the Bayesian approach aims at stimulus reconstruction rather than model reconstruction (Fig. 1). Given a neural response and a model for how that response could be generated, stimulus reconstruction attempts to determine what stimulus occurred (Stanley et al., 1999). To simplify the task, identification of the most likely stimulus among a finite number of possibilities (Kay et al., 2008) or classification into a number of discrete categories (Haxby et al., 2001; Kamitani and Tong, 2005) and/or the use of more simplified stimuli (Miyawaki et al., 2008) have all been used. There can be no claim about whether that representation of the stimulus is used in the brain, only that information is available in the measured responses that can be used to recreate the stimulus. Reconstruction, identification and classification have been used in many experiments to compare sensory responses under different cognitive states like attention (Kamitani and Tong, 2005, 2006; Jehee et al., 2011; Dobs et al., 2018) or working memory (Harrison and Tong, 2009), examine the influence of priors and expectancy (Kok et al., 2012, 2013; Vintch and Gardner, 2014) and a wide variety of other purposes. Channel encoding models have also been fruitfully used for stimulus reconstruction, for example by reconstructing color values that the model was never trained on (Brouwer and Heeger, 2009).

The inverted encoding model approach does not aim to reconstruct the stimulus, but rather aims to reconstruct an intermediate step of the analysis: the encoding model's representation of the stimulus. The parameters of the tuning functions of different channels in the encoding model are often taken to mimic the selectivity of neurons or groups of neurons, yet the reconstructed channel response functions do not unambiguously reflect the tuning properties of these neurons (Liu et al., 2018). Therefore, the predicted channel response that the analysis recreates exists only as a theoretic construct; it is neither inherent in the stimulus nor in the population representation. As demonstrated above, a bimodal channel response can be reconstructed from a population representation that was built from unimodal representations of the stimulus. However, the Bayesian analysis, despite using the same bimodal encoding model, recovers a unimodal posterior because it aims to reconstruct the stimulus rather than the model. While channels for basic stimulus properties like color, orientation and spatial frequency can be informed by existing physiologic literature, model specification is less well constrained for more complex stimulus properties and the possibility of poor model specification giving rise to misleading results becomes more likely. To be clear, building encoding models based on well-understood tuning functions even with the ambiguities described here is not necessarily problematic as it can be a useful way to reduce the dimensionality of the stimulus space in a principled way. However, inverting the encoding model even for these cases where the single-unit tuning functions are well known, simply recapitulates the assumptions about the channel basis functions, such as their tuning width, and therefore does not provide a useful assay of population tuning. Thus, inverted encoding models produce a result that is not interpretable as a population nor a neural tuning function, but instead is an estimate of the arbitrary model basis function.

Rather than inverting the encoding model to display the fit to the intermediate model assumptions, examining the weights that are needed to explain population responses can be informative about the population representation. That is, encoding models without inversion, have often been used to understand population representations. For example, a Gabor wavelet model can be used to encode visual stimuli into spatially local filters with different orientation and spatial frequency selectivity meant to mimic the selectivity of primary visual cortex neurons (Kay et al., 2008). After fitting such a model, the location, orientation and spatial frequency selectivity can be determined for each voxel, allowing for retinotopic mapping of visual cortex and evaluation of the amount of orientation and scale information available in voxel representations. Similarly, a population receptive field model which encodes visual stimuli like high contrast bars into Gaussian receptive fields (Dumoulin and Wandell, 2008) with an exponential non-linearity (Kay et al., 2013) is routinely used to define retinotopic field maps (Benson et al., 2018). More complex encoding models of semantic category of visual objects (Naselaris et al., 2009; Huth et al., 2012) or language (Huth et al., 2016) have also been fit to voxel responses and examination along which dimensions of the model space the fitted weights vary the most can be used to understand the nature of what is represented.

That inverted encoding models recover the model responses, not the stimulus, is not to say that they have no useful purpose. Inverted encoding models have been fruitfully used to tease apart responses to different aspects of a compound stimulus into target and mask responses to evaluate predictions of normalization models (Brouwer and Heeger, 2011; Brouwer et al., 2015). Reconstructing model responses might be particularly important in a brain machine interface, where the model might include, for example, the response of different actuators for a robotic arm. Inverting a channel encoding model also allows for reconstruction of stimuli for which the model has never been trained, by comparing the recovered channel responses to those that would be elicited by untrained stimuli and selecting the stimulus whose channel response is most correlated with the one recovered by the inverted model (Brouwer and Heeger, 2009; Lorenc et al., 2018). Summing model receptive fields weighted by the recovered channel responses (Sprague and Serences, 2013; Sprague et al., 2014, 2016, 2018b; Samaha et al., 2016; Vo et al., 2017) is a computation similar in spirit to a vector-average read-out (Georgopoulos et al., 1986; Lee et al., 1988; Gardner et al., 2004) in that it allows each channel to “vote” for its preferred spatial location according to its reconstructed response. Thus, this approach can be viewed as a further elaboration of the inverted encoding model as it aims to determine the expected population read-out of a stimulus compatible with the measured response, rather than a model reconstruction. However, unlike the Bayesian approach (van Bergen et al., 2015; van Bergen and Jehee, 2018), it does not provide an estimate of how likely any stimulus is given the measured response. Despite these valuable usages of inverted encoding models, when the model inversion recovers theoretical channel responses such as orientation tuned channels, the properties of those channel responses should be considered a property of the model and the estimation process and not as a measurement of underlying selectivity of the hypothetical neural tuning functions (Liu et al., 2018) or the population. As a specific example, the tuning width of the channel responses should not be taken as a measure of population selectivity as it will depend on the tuning width of the particular (and arbitrary) channel basis functions used.

Our results here show that channel basis functions are only determined up to an invertible linear transform, but this does not preclude comparison of encoding models whose basis functions are not related by an invertible linear transform. In such cases, standard statistical model comparisons that take into account the number of parameters and the goodness-of-fit can be used to select the best fitting model. Because these non-linearly-relatable models make different predictions, one can also compare model predictions to other behavioral and neural measures of perceptual space to select models. As a concrete example, Brouwer and Heeger (2009) compared a six-channel hue tuning model with a four-channel cone opponency tuning model and concluded that the former was more consistent with the data in hV4. This is possible because these two models are not related by an invertible transform.

Proper inferences from computational modeling of data can only be achieved if the limits imposed by these techniques are explored and recognized by the communities that use them. Our results can be considered an example of this principle. Another analogous example to the issue that we describe here can be found in the theory and experiments of population coding of color. Indeed, the trichromatic color theory developed from the work of Young and Helmholtz (Young, 1802; Helmholtz, 1867), can only establish color matching functions up to a linear transform because they depend on the spectral power distribution of the three primary lights used in the matching experiment (Wandell, 1995). However, because the linear assumptions of color matching theory were known for over a century (Grassmann, 1854), experimenters were able to make the correct inference that the cone sensitivities in the primate retina would only need to match up to a linear transform (Baylor et al., 1987) to the color matching functions measured perceptually. Thus, the linking hypothesis between population coding in the retina and perception of colors was validated only because there was clear understanding of the limits imposed by the underlying theory.

While sophisticated new computational techniques such as inverted encoding models offer the possibility of new discovery from large and complicated datasets, they also intervene many layers of mathematical analysis between measurement and data presentation, thus creating interpretational challenges. This is not a challenge unique to human imaging, but shared with other analyses of population activity measures including electrophysiologically or through calcium imaging. Whether a computational analysis is discovering structure within data or imposing it can at times be difficult to adjudicate. For example, dimensionality reduction techniques have been used to uncover rotational dynamics in motor preparatory population activity (Churchland et al., 2012), but it could be that the computational techniques are able to extract dimensions of rotational dynamics whether or not they are in the data. One possible way to address this question is by the use of carefully designed surrogate data sets which have various components of population activity removed, to understand where effects are coming from (Elsayed and Cunningham, 2017). The larger question in assessing population stimulus representations remains as to what information is carried in a population that is not inherent in the single-unit representation. Indeed, even theoretic notions that try to decompose information into components that are represented by individual neurons and ones that are synergistically represented have difficulty in formally defining what is meant by synergistic information that arises from the population but is not in the individual units (Lizier et al., 2018). Moving forward, our analyses and understanding of population stimulus representations will need to derive from agreed on definitions for what is meant by population representations and from considerations of how much analyses impose on structure versus how much they reveal.

## References

Avidan G, Harel M, Hendler T, Ben-Bashat D, Zohary E, Malach R (2002) Contrast sensitivity in human visual areas and its relationship to object recognition. J Neurophysiol 87:3102–3116.

Baker TJ, Norcia AM, Candy TR (2011) Orientation tuning in the visual cortex of 3-month-old human infants. Vision Res 51:470–478.

Baylor DA, Nunn BJ, Schnapf JL (1987) Spectral sensitivity of cones of the monkey *Macaca fascicularis*. J Physiol 390:145–160.

Benson NC, Jamison KW, Arcaro MJ, Vu A, Glasser MF, Coalson TS, Van Essen DC, Yacoub E, Ugurbil K, Winawer J, Kay K (2018) The HCP 7T retinotopy dataset: description and pRF analysis. bioRxiv 308247.

Benucci A, Frazor RA, Carandini M (2007) Standing waves and traveling waves distinguish two circuits in visual cortex. Neuron 55:103–117.

Birman D, Gardner JL (2018) A quantitative framework for motion visibility in human cortex. J Neurophysiol 120:1824–1839.

Boynton GM, Engel SA, Glover GH, Heeger DJ (1996) Linear systems analysis of functional magnetic resonance imaging in human V1. J Neurosci 16:4207–4221.

Boynton GM, Demb JB, Glover GH, Heeger DJ (1999) Neuronal basis of contrast discrimination. Vision Res 39:257–269.

Braddick OJ, O'Brien JM, Wattam-Bell J, Atkinson J, Hartley T, Turner R (2001) Brain areas sensitive to coherent visual motion. Perception 30:61–72.

Brouwer GJ, Heeger DJ (2009) Decoding and reconstructing color from responses in human visual cortex. J Neurosci 29:13992–14003.

Brouwer GJ, Heeger DJ (2011) Cross-orientation suppression in human visual cortex. J Neurophysiol 106:2108–2119.

Brouwer GJ, Heeger DJ (2013) Categorical clustering of the neural representation of color. J Neurosci 33:15454–15465.

Brouwer GJ, Arnedo V, Offen S, Heeger DJ, Grant AC (2015) Normalization in human somatosensory cortex. J Neurophysiol 114:2588–2599.

Buck LB (2004) Olfactory receptors and odor coding in mammals. Nutr Rev 62:S184–S188.

Bullock T, Elliott JC, Serences JT, Giesbrecht B (2017) Acute exercise modulates feature-selective responses in human cortex. J Cogn Neurosci 29:605–618.

Byers A, Serences JT (2014) Enhanced attentional gain as a mechanism for generalized perceptual learning in human visual cortex. J Neurophysiol 112:1217–1227.

Campbell FW, Cleland BG, Cooper GF, Enroth-Cugell C (1968) The angular selectivity of visual cortical cells to moving gratings. J Physiol 198:237–250.

Chen N, Bi T, Zhou T, Li S, Liu Z, Fang F (2015) Sharpened cortical tuning and enhanced cortico-cortical communication contribute to the long-term neural mechanisms of visual motion perceptual learning. Neuroimage 115:17–29.

Chen Y, Palmer CR, Seidemann E (2012) The relationship between voltage-sensitive dye imaging signals and spiking activity of neural populations in primate V1. J Neurophysiol 107:3281–3295.

Chong E, Familiar AM, Shim WM (2016) Reconstructing representations of dynamic visual objects in early visual cortex. Proc Natl Acad Sci USA 113:1453–1458.

Churchland MM, Cunningham JP, Kaufman MT, Foster JD, Nuyujukian P, Ryu SI, Shenoy KV (2012) Neural population dynamics during reaching. Nature 487:51–56.

Costagli M, Ueno K, Sun P, Gardner JL, Wan X, Ricciardi E, Pietrini P, Tanaka K, Cheng K (2014) Functional signalers of changes in visual stimuli: cortical responses to increments and decrements in motion coherence. Cereb Cortex 24:110–118.

DeAngelis GC, Ohzawa I, Freeman RD (1993) Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. II. Linearity of temporal and spatial summation. J Neurophysiol 69:1118–1135.

Dobs K, Schultz J, Bülthoff I, Gardner JL (2018) Task-dependent enhancement of facial expression and identity representations in human cortex. Neuroimage 172:689–702.

Doetsch GS (2000) Patterns in the brain. Neuronal population coding in the somatosensory system. Physiol Behav 69:187–201.

Dumoulin SO, Wandell BA (2008) Population receptive field estimates in human visual cortex. Neuroimage 39:647–660.

Elsayed GF, Cunningham JP (2017) Structure in neural population recordings: an expected byproduct of simpler phenomena? Nat Neurosci 20:1310–1318.

Ester EF, Anderson DE, Serences JT, Awh E (2013) A neural measure of precision in visual working memory. J Cogn Neurosci 25:754–761.

Ester EF, Sprague TC, Serences JT (2015) Parietal and frontal cortex encode stimulus-specific mnemonic representations during visual working memory. Neuron 87:893–905.

Ester EF, Sutterer DW, Serences JT, Awh E (2016) Feature-selective attentional modulations in human frontoparietal cortex. J Neurosci 36:8188–8199.

Finn IM, Priebe NJ, Ferster D (2007) The emergence of contrast-invariant orientation tuning in simple cells of cat visual cortex. Neuron 54:137–152.

Foster JJ, Sutterer DW, Serences JT, Vogel EK, Awh E (2016) The topography of alpha-band activity tracks the content of spatial working memory. J Neurophysiol 115:168–177.

Garcia JO, Srinivasan R, Serences JT (2013) Near-real-time feature-selective modulations in human cortex. Curr Biol 23:515–522.

Gardner JL, Anzai A, Ohzawa I, Freeman RD (1999) Linear and nonlinear contributions to orientation tuning of simple cells in the cat's striate cortex. Vis Neurosci 16:1115–1121.

Gardner JL, Tokiyama SN, Lisberger SG (2004) A population decoding framework for motion aftereffects on smooth pursuit eye movements. J Neurosci 24:9035–9048.

Gardner JL, Sun P, Waggoner RA, Ueno K, Tanaka K, Cheng K (2005) Contrast adaptation and representation in human early visual cortex. Neuron 47:607–620.

Georgopoulos AP, Schwartz AB, Kettner RE (1986) Neuronal population coding of movement direction. Science 233:1416–1419.

Grassmann P (1854) XXXVII. On the theory of compound colours. Lond Edinb Dublin Philos Mag J Sci 7:254–264.

Grinvald A, Lieke E, Frostig RD, Gilbert CD, Wiesel TN (1986) Functional architecture of cortex revealed by optical imaging of intrinsic signals. Nature 324:361–364.

Harrison SA, Tong F (2009) Decoding reveals the contents of visual working memory in early visual areas. Nature 458:632–635.

Hartline HK (1938) The response of single optic nerve fibers of the vertebrate eye to illumination of the retina. Am J Physiol 121:400–415.

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293:2425–2430.

Helmholtz HV (1867) Helmholtz's treatise on physiological optics. In: The perceptions of vision, Vol 3 (Southall JPC, ed), 1924 ed. Washington, DC: The Optical Society of America.

Ho T, Brown S, van Maanen L, Forstmann BU, Wagenmakers E-J, Serences JT (2012) The optimality of sensory processing during the speed–accuracy tradeoff. J Neurosci 32:7992–8003.

Hubel DH, Wiesel TN (1959) Receptive fields of single neurones in the cat's striate cortex. J Physiol 148:574–591.

Hubel DH, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol 160:106–154.

Huth AG, Nishimoto S, Vu AT, Gallant JL (2012) A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76:1210–1224.

Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL (2016) Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532:453–458.

Jehee JFM, Brady DK, Tong F (2011) Attention improves encoding of task-relevant features in the human visual cortex. J Neurosci 31:8210–8219.

Jones JP, Palmer LA (1987) The two-dimensional spatial structure of simple receptive fields in cat striate cortex. J Neurophysiol 58:1187–1211.

Kamitani Y, Tong F (2005) Decoding the visual and subjective contents of the human brain. Nat Neurosci 8:679–685.

Kamitani Y, Tong F (2006) Decoding seen and attended motion directions from activity in the human visual cortex. Curr Biol 16:1096–1102.

Kay KN, Naselaris T, Prenger RJ, Gallant JL (2008) Identifying natural images from human brain activity. Nature 452:352–355.

Kay KN, Winawer J, Mezer A, Wandell BA (2013) Compressive spatial summation in human visual cortex. J Neurophysiol 110:481–494.

Kok P, Jehee JFM, de Lange FP (2012) Less is more: expectation sharpens representations in the primary visual cortex. Neuron 75:265–270.

Kok P, Brouwer GJ, van Gerven MAJ, de Lange FP (2013) Prior expectations bias sensory representations in visual cortex. J Neurosci 33:16275–16284.

Lee C, Rohrer WH, Sparks DL (1988) Population coding of saccadic eye movements by neurons in the superior colliculus. Nature 332:357–360.

Liu T, Cable D, Gardner JL (2018) Inverted encoding models of human population response conflate noise and neural tuning width. J Neurosci 38:398–408.

Lizier J, Bertschinger N, Jost J, Wibral M (2018) Information decomposition of target effects from multi-source interactions: perspectives on previous, current and future work. Entropy 20:307.

Logothetis NK, Pauls J, Augath M, Trinath T, Oeltermann A (2001) Neurophysiological investigation of the basis of the fMRI signal. Nature 412:150–157.

Lorenc ES, Sreenivasan KK, Nee DE, Vandenbroucke ARE, D'Esposito M (2018) Flexible coding of visual working memory representations during distraction. J Neurosci 38:5267–5276.

Maffei L, Campbell FW (1970) Neurophysiological localization of the vertical and horizontal visual coordinates in man. Science 167:386–387.

Miyawaki Y, Uchida H, Yamashita O, Sato M-A, Morito Y, Tanabe HC, Sadato N, Kamitani Y (2008) Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. Neuron 60:915–929.

Naselaris T, Prenger RJ, Kay KN, Oliver M, Gallant JL (2009) Bayesian reconstruction of natural images from human brain activity. Neuron 63:902–915.

Naselaris T, Kay KN, Nishimoto S, Gallant JL (2011) Encoding and decoding in fMRI. Neuroimage 56:400–410.

Ohki K, Chung S, Ch'ng YH, Kara P, Reid RC (2005) Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex. Nature 433:597–603.

Olman CA, Ugurbil K, Schrater P, Kersten D (2004) BOLD fMRI and psychophysical measurements of contrast response to broadband images. Vision Res 44:669–683.

Pestilli F, Carrasco M, Heeger DJ, Gardner JL (2011) Attentional enhancement via selection and pooling of early sensory responses in human visual cortex. Neuron 72:832–846.

Priebe NJ, Ferster D (2012) Mechanisms of neuronal computation in mammalian visual cortex. Neuron 75:194–208.

Rees G (2000) A direct quantitative relationship between the functional properties of human and macaque V5. Nat Neurosci 3:716–723.

Regan D, Regan MP (1987) Nonlinearity in human visual responses to two-dimensional patterns, and a limitation of fourier methods. Vision Res 27:2181–2183.

Ringach DL, Shapley RM, Hawken MJ (2002) Orientation selectivity in macaque V1: diversity and laminar dependence. J Neurosci 22:5639–5651.

Rose D, Blakemore C (1974) An analysis of orientation selectivity in the cat's visual cortex. Exp Brain Res 20:1–17.

Samaha J, Sprague TC, Postle BR (2016) Decoding and reconstructing the focus of spatial attention from the topography of alpha-band oscillations. J Cogn Neurosci 28:1090–1097.

Saproo S, Serences JT (2014) Attention improves transfer of motion information between V1 and MT. J Neurosci 34:3586–3596.

Scolari M, Byers A, Serences JT (2012) Optimal deployment of attentional gain during fine discriminations. J Neurosci 32:7723–7733.

Serences JT, Saproo S (2012) Computational advances towards linking BOLD and behavior. Neuropsychologia 50:435–446.

Sprague TC, Serences JT (2013) Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. Nat Neurosci 16:1879–1887.

Sprague TC, Ester EF, Serences JT (2014) Reconstructions of information in visual spatial working memory degrade with memory load. Curr Biol 24:2174–2180.

Sprague TC, Ester EF, Serences JT (2016) Restoring latent visual working memory representations in human cortex. Neuron 91:694–707.

Sprague TC, Adam KCS, Foster JJ, Rahmati M, Sutterer DW, Vo VA (2018a) Inverted encoding models assay population-level stimulus representations, not single-unit neural tuning. eNeuro 5:ENEURO.0098–18.2018.

Sprague TC, Itthipuripat S, Vo VA, Serences JT (2018b) Dissociable signatures of visual salience and behavioral relevance across attentional priority maps in human cortex. J Neurophysiol 119:2153–2165.

Stanley GB, Li FF, Dan Y (1999) Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. J Neurosci 19:8036–8042.

Swindale NV, Grinvald A, Shmuel A (2003) The spatial pattern of response magnitude and selectivity for orientation and direction in cat visual cortex. Cereb Cortex 13:225–238.

Tootell RB, Taylor JB (1995) Anatomical evidence for MT and additional cortical visual areas in humans. Cereb Cortex 5:39–55.

Tootell RBH, Hadjikhani NK, Vanduffel W, Liu AK, Mendola JD, Sereno MI, Dale AM (1998) Functional analysis of primary visual cortex (V1) in humans. Proc Natl Acad Sci USA 95:811–817.

Tsai JJ, Wade AR, Norcia AM (2012) Dynamics of normalization underlying masking in human visual cortex. J Neurosci 32:2783–2789.

van Bergen RS, Jehee JFM (2018) Modeling correlated noise is necessary to decode uncertainty. Neuroimage 180:78–87.

van Bergen RS, Ji Ma W, Pratte MS, Jehee JFM (2015) Sensory uncertainty decoded from visual cortex predicts behavior. Nat Neurosci 18:1728–1730.

Verghese P, Kim Y-J, Wade AR (2012) Attention selects informative neural populations in human V1. J Neurosci 32:16379–16390.

Vintch B, Gardner JL (2014) Cortical correlates of human motion perception biases. J Neurosci 34:2592–2604.

Vo VA, Sprague TC, Serences JT (2017) Spatial tuning shifts increase the discriminability and fidelity of population codes in visual cortex. J Neurosci 37:3386–3401.

Wandell BA (1995) Foundations of vision. Sunderland, MA: Sinauer Associates.

Watkins D, Berkley M (1974) The orientation selectivity of single neurons in cat striate cortex. Exp Brain Res 19:433–446.

Young T (1802) The Bakerian lecture: on the theory of light and colours. Philos Trans R S Lond 92:12–48.

Yu Q, Shim WM (2017) Occipital, parietal, and frontal cortices selectively maintain task-relevant features of multi-feature objects in visual working memory. Neuroimage 157:97–107.