Seeing White: Qualia in the Context of Decoding Population Codes

Sidney R. Lehky
Cognitive Brain Mapping Laboratory, Brain Science Institute, Institute of Physical and Chemical Research (RIKEN), Wako-shi, Saitama 351-0198, Japan

Terrence J. Sejnowski
Howard Hughes Medical Institute, Computational Neuroscience Laboratory, The Salk Institute, La Jolla, CA 92037, U.S.A., and Department of Biology, University of California, San Diego, La Jolla, CA 92093, U.S.A.

When the nervous system is presented with multiple simultaneous inputs of some variable, such as wavelength or disparity, they can be combined to give rise to qualitatively new percepts that cannot be produced by any single input value. For example, there is no single wavelength that appears white. Many models of decoding neural population codes have problems handling multiple inputs, either attempting to extract a single value of the input parameter or, in some cases, registering the presence of multiple inputs without synthesizing them into something new. These examples raise a more general issue regarding the interpretation of population codes. We propose that population decoding involves not the extraction of specific values of the physical inputs, but rather a transformation from the input space to some abstract representational space that is not simply related to physical parameters. As a specific example, a four-layer network is presented that implements a transformation from wavelength to a high-level hue-saturation color space.

1 Introduction

Population coding is the notion that a perceptual or motor variable is represented in the nervous system by the pattern of activity in a population of neurons, each coarsely tuned to a different but overlapping range of the parameter in question. The response of a single neuron, having a roughly bell-shaped tuning curve (not necessarily gaussian), is ambiguous, but the joint activity of all neurons in the population is not (see Figure 1). An alternative to population coding is rate encoding, in which the parameter is indicated by the activity of a single neuron whose firing rate increases

Neural Computation **11, 1261–1280** (1999) © 1999 Massachusetts Institute of Technology

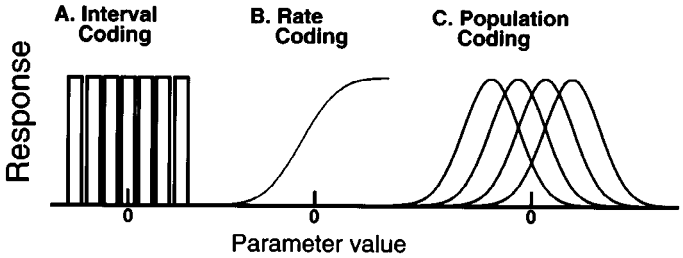


Figure 1: Three methods for encoding a physical variable such as orientation of a line or disparity between two eyes. (A) Interval encoding: A separate unit is dedicated for each narrow range of values. (B) Rate encoding: The firing rate is monotonically related to the value of the physical variable. (C) Population encoding: The pattern of activity in a population of neurons with broad overlapping tuning curves represents the value. (Adapted from Figure 1 in Lehky & Sejnowski, 1990a.)

monotonically as the parameter changes. Another alternative is interval encoding, in which, again, the activity of a single neuron indicates the parameter value, this time by firing only when the parameter falls in some small interval (i.e., the neuron is "labeled" for that interval). As the parameter value changes, a different neuron fires. The resolution of an interval code (discrimination threshold) depends on the width of the tuning curve, unlike a population code, where it is a function of tuning curve slope (Lehky & Sejnowski, 1990a).

The first population code proposed, and almost certainly the best known, is the trichromatic theory of color vision (Young, 1802). This holds that perceived color is due to the relative activities in three broadly tuned color channels in the visual system. Given that color was the first population code devised, it is not surprising that the first model for interpreting a population code was also developed in the context of color vision. This was the line-element model of Helmholtz (1909/1962) (later modified by the physicist Schrödinger among others; see Wyszecki & Stiles, 1982, for a review of line element models). Roughly speaking, it treated the activities of the three elements of the color-coding population as components of a vector and proposed that two colors become discriminable when the vector difference reaches some threshold value.

Over the past century there has been an extensive psychophysical literature on line element models, in part because this is an instance where a good model for deciphering a neural population code is of some commercial importance. Manufacturers would like to predict how much variability in a production process can be tolerated before colors appear nonuniform. In other words, at what point does color appearance change when there are small changes in the activities of the channels of the color code? This is a problem of population code interpretation.

1.1 Different Approaches to Decoding. In recent years there has been an expanded interest in neural population codes and models for decoding them (including work by Chen & Wise, 1997; Lee, Rohrer, & Sparks, 1988; Lehky & Sejnowski, 1990a; Paradiso, 1988; Pouget & Thorpe, 1991; Pouget, Zhang, Deneve, & Lathan, 1998; Salinas & Abbott, 1994, 1995; Sanger, 1996; Seung & Sompolinsky, 1993; Snippe, 1996; Vogels, 1990; Wilson & Gelb, 1984; Wilson & McNaughton, 1993; Young & Yamani, 1992; Zhang, Ginsburg, McNaughton, & Sejnowski, 1998; Zohary, 1992). One influential approach has been vector-averaging models, developed by Georgopolous, Schwartz, and Kettner (1986) in the context of predicting the direction of arm movements from a population of direction-tuned motor cells. In these models, each unit in the population is represented by a vector pointing in the direction of the peak of that unit's tuning curve and whose length is proportional to the unit's activity. The parameter value represented by the population as a whole is given by the vector average of these components.

This "Georgopolous type" vector model and the "Helmholtz type" line element model differ in purpose. The Helmholtz model cannot give the parameter value but seeks only to determine the smallest discriminable change, while the Georgopolous model seeks to determine the actual value of the parameter. It is significant to note that the Georgopolous vector model fails completely when applied to predicting color appearance from wavelength tuning curves, for it can never predict the appearance of "white." The model would take a weighed average of the peak wavelengths of the tuning curves, producing the value of some other wavelength, and there is no single wavelength that corresponds to "white." This example is a problem not only for vector models of population decoding, but for others as well, as will be outlined below.

Shadlen, Britten, Newsome, and Movshon (1996) use a population decoding method, which can be considered part of the same general class as vector averaging. They had two pools of visual neurons tuned to stimulus motion in 180-degree opposite directions (a pool is multiple copies of neurons with the same tuning and partially correlated noise). The represented motion direction was indicated by whichever of the two pools had the greater average activity. This population decoding method is less sophisticated than vector averaging, for it just looks at which vector is the largest (i.e., it implements peak detection among activities of members in the encoding population, counting each "pool" as one member of a population code). This "biggest vector" method is limited because N vectors (N different tuning curves) can represent only N discrete parameter values, whereas vector averaging can represent a continuum. In this case, two neural pools worked because the model was given the constraint that motion could occur in only exactly two directions. Three allowable stimulus directions would have required three pools. Allowing a continuous range of inputs is problematic and would require an enormous number of pools whose tuning peaks differed by about the value of the just-noticeable difference in motion direction. Although presented as a population code, the Shadlen et al. (1996) method in practice seems to operate more like interval coding.

A population code, unlike an interval code, can also be used for fine parameter discrimination of a parameter (hyperacuity) with a relatively small number of tuning curves. This is because when there is a small increment in parameter value, the resulting change in activity in each tuning curve of the population can be pooled to produce a total change in population activity that is significant relative to noise. In one implementation of this approach (Lehky & Sejnowski, 1990a), the probability of a single tuning curve's detecting a parameter increment is given by:

$$p_i = \sqrt{\frac{2}{\pi}} \int_{-\infty}^{d'/\sqrt{2}} e^{-x^2/2} dx - 1 \qquad (p_1 = 0 \text{ for } d' < 0)$$

where d' is the ratio of change in activity in a given tuning curve to the noise.[1] The total probabilities of N statistically independent tuning curves could be pooled as $p = 1 - \prod_{i=1}^{N} (1 - p_i)$. The threshold for detecting a change in parameter (disparity in this particular case) occurs when the total probability p reaches a criterion value.

Returning to parameter estimation, a different approach to decoding a population code besides vector models is one based on probabilistic techniques (Paradiso, 1988; Pouget et al., 1998; Sanger, 1996; Seung & Sompolinsky, 1993; Snippe, 1996; Zhang et al., 1998, reviewed by Oram, Földiák, Perrett, & Sengpiel, 1998). Given a set of responses from a noisy population, the problem is to estimate the most probable stimulus that may have caused it. Two major classes of probabilistic models exist: those based on Bayesian estimation and those based on maximum likelihood calculations. Although they are more cumbersome to calculate, probabilistic models generally give more accurate interpretations of noisy encoding populations than the vector models do. A factor contributing to this superior performance is that the probabilistic models assume knowledge of the shapes of all the tuning curves in the population. In the Georgopolous-style vector model, tuning curves are always assumed to be cosine shaped, regardless of the actual situation. The superior performance of maximum likelihood and Bayesian models contributes to their popularity among theoreticians, and the simplicity of the vector models, as well as the relative straightforwardness of constructing neural implementations for them, contributes to their popularity among experimentalists. Other well-known models are special cases

[1] This corrects an erratum in the original presentation.

of those described so far. For example, it is possible to restate the Shadlen et al. (1996) "biggest vector" model in terms of a probabilistic, maximum likelihood formalism (A. Pouget, personal communication), although the restriction of allowing only two discrete output values still renders this model atypical of population coding models.

Probabilistic models can suffer from the same defect mentioned previously in connection with vector models: the inability to predict "white" from wavelength tuning curves. This happens when the statistical algorithm is set up so that it must interpret the population as representing one particular value of the parameter in question. For example, Bayesian estimation models output a probability distribution of possible values of the stimulus being represented by the population. The population is then often interpreted as representing whatever parameter value occurs at the peak of that distribution (for example, see Sanger, 1996; Zhang et al., 1998). In the case of color, this peak in the distribution is at some particular wavelength, which of course cannot represent "white."

Probabilistic models are not restricted to having single-valued outputs, and it is not difficult to conceive of more sophisticated variations. Steps in this direction have been taken by Anderson (1994) and Zemel, Dayan, and Pouget (1998), which seek to estimate the entire probability distribution of a parameter rather than just the peak of the distribution. This would allow the use of a multimodal distribution to represent multiple parameter values simultaneously.

1.2 Problems Caused by Mixtures of Stimuli. Being able to represent multiple values simultaneously is still not enough. Another aspect to the problem is synthesizing these multiple values to form something qualitatively different from any of the components. In addition to the problem of predicting "white" from wavelength tuning curves, a second example of such a "multiple-input" problem is transparent surfaces, where the cues indicating transparency can be either different disparities (Lehky & Sejnowski, 1990a) or different motions (Adelson & Movshon, 1982; Qian et al., 1994; Stoner & Albright, 1992). An interesting aspect of such stimulus mixtures, or "complex stimuli," whether involving motion, disparity, or color, is that the percept of the mixture can produce something that is qualitatively different from that produced by any "simple stimulus." There is no single wavelength that produces the percept of white; there is no single motion or disparity that produces the percept of transparency.

What we see in complex stimulus composed of x_1 and x_2 is not some sort of averaging process $(x_1 + x_2)/2$, which is more or less what the various vector and statistical models that output a single value are doing. White is not produced by averaging the wavelengths of blue and yellow. Nor is what we see (x_1 AND x_2), as might be the output of models with a multimodal probability distribution (Anderson, 1994; Nowlan & Sejnowski, 1995; Zemel et al., 1998). When blue and yellow are mixed, we see not blue and yellow but a unitary percept, which is different from either. Moreover, there are an infinite number of such wavelength mixtures (metamers) which give rise to an identical percept of white. What is missing in the multimodal models is a synthetic process combining the different components to form something new. What the existence of "white" is telling us is that the process of population decoding maps inputs onto a new representation space that may not correspond simply to any physical variable

Extracting a single physical parameter (or even two or three discrete numbers) from a population code is convenient for an outside observer, or perhaps even a homunculus inside the brain, but may not fit well with how population codes might be used internally by the brain. Rather, if one network feeds into another, and again into another, in a series of vector-vector transforms, then there is never any need to make information contained in the population code explicit. The final output, whether into a distributed motor output program, memory storage scheme, or subjective percepts (qualia), would still be some pattern of activity in a population, albeit in a different representation space.

Under this form of organization, the interpretation placed on the pattern of activity in a population depends on the characteristics of the network it feeds into. We shall call this the *decoding network*. As a special case, there may be decoding networks that try to extract the value of the physical parameter underlying the population activity, but generally this need not be true. It is also the decoding network that provides the synthetic capability of mixing multiple inputs into something qualitatively different from any of the components. This repeats a point we have made earlier (Lehky & Sejnowski, 1988) that meaning is determined to a major extent by the projective fields (decoding networks), and not solely by the receptive fields (tuning curves) of a population of neurons.

For any stimulus, the input population will form some pattern of activity, and the decoder network will act as a pattern recognition device and assign an output to that pattern. In a certain sense this can be thought of as a template matching operation, with the decoding network implicitly containing a set of templates. A pattern of input activity that matches something the decoder network is "looking for" will trigger a particular response. The use of the term *template matching* here does not imply a commitment to any particular mathematical formulation, and as is often the case with neural networks, the process may be difficult to express in closed mathematical form. (For previous applications of template matching to population decoding see Buchsbaum & Goldstein, 1979; Lehky & Sejnowski, 1990b,[2] and Pouget et al., 1998.)

[2] In the template matching method used by Lehky and Sejnowski (1990a), in the matrix of equation 9, row 3, column 3 should read +1 rather than −1.

This approach therefore interprets the problem of population decoding as a problem in pattern recognition, which neural networks are good at. Although the language used here implies a degree of discreteness in the process, there is no reason that smooth changes in inputs cannot map into smooth changes in outputs, although there is nothing that requires smooth mappings. As an added feature, in a noisy system, template matching (if it happens to be implemented as a least squares fit) can represent a maximum likelihood estimate of a pattern, assuming gaussian noise (Pouget et al., 1998; Salinas & Abbott, 1994), so the process can be statistically efficient. (On the other hand, whether actual biological decoding processes are statistically efficient is an empirical question, so statistical efficiency is not an automatic virtue in decoding models. If data show that a biological process is not efficient, then one may not want an efficient model, except as a benchmark for evaluating biological performance.)

The role of the decoding network may be less apparent in motor systems than perceptual systems, because motor systems deal with the position and movement of physical objects, and may be more constrained by the physics of the situation than perceptual systems are constrained in assigning qualia to their inputs. This does not mean that population decoding proceeds in fundamentally different ways in perceptual and motor systems, but rather that one may be more likely in motor systems to hit on a situation where a simpler model (vector averaging, for example) is "good enough."

The significance attached to the pattern of activity in a population depends on the nature of the decoding network into which it feeds. The relationship between syntax (pattern of activity in a population) and semantics (meaning of the pattern) is arbitrary. This arbitrary connection between physical manifestation and meaning is a characteristic of a symbolic process. In essence, what was argued above is that there is an aspect to population decoding that resembles symbolic processing, though not quite the same because the element of discreteness is not present. This is to say, for example, that the percept of white can be thought of as an arbitrary symbol or label marking the presence of a certain combination of color-tuning curve activities (and not a simple index of physical wavelength, which it obviously is not). It may be a limitation of some current models of population decoding that they treat the process as a purely physical analog one rather than one that has quasi-symbolic aspects with only an arbitrary connection to the physical world.

2 Example: Creating a Population Code for "White"

Having said all this, let us be more specific about what a model for decoding population codes might look like. What we envisage is a network acting as a transformation between physical input space and a nonphysical perceptual space (where such things as "white" reside). Recording from monkeys, Komatsu, Ideura, Kaji, and Yamane (1992; see also Komatsu, 1998) have measured color responses of inferotemporal neurons as a function of position in Commission Internationale de l'Eclairage 1931 (CIE) color space rather than wavelength. (CIE color space is essentially a two-dimensional hue-saturation space, with white toward the center and different hues arranged around the rim, as in Figure 3b). They found most inferotemporal units were responsive to local patches within this color space, which represented a complex transform of cone responses, but a simple arrangement of grouping similar colors in our perceptual space. In contrast, color properties at the early stages of the visual system, such as striate cortex or lateral geniculate nucleus (LGN), are linear transforms of cone responses. The sort of properties seen in inferotemporal cortex, going from a more physical space to a more psychological space, is an example of what we have in mind during the process of decoding a population whose input activities are tied to some physical parameter. It is straightforward to create something like this transformation with a neural network. The point here is not that some novel modeling technique is required, for that is not the case. Rather, what is important is that we have changed the question asked of population decoding away from specifying a physical parameter to specifying some region in an abstract psychological space.

As a specific example of how such a high-level color representation might be implemented, we created a four-layer, feedforward, fully connected network using the backpropagation algorithm (Rumelhart, Hinton, & Williams, 1986) (see Figure 2). The input (layer 1) consisted of the three cone types, and the output (layer 4) formed a population of units having overlapping, two-dimensional gaussian receptive fields in CIE color space. Layer 2 units were a set of linear color opponent channels plus a luminance channel, and layer 3 units had properties that developed in the course of training the network to have the desired input-output relationship. Previous color models of related interest include De Valois and De Valois (1993), Usui, Nakauchi, and Miyake (1994), and Wray and Edelman (1996).

2.1 Defining the Network. The wavelength responses for the three cone inputs were defined by the following equation, which is equation 2′ from Lamb (1995):

$$S(\lambda) = \frac{1}{\exp a(A - \lambda_{\max}/\lambda) + \exp b(B - \lambda_{\max}/\lambda) + \exp c(C - \lambda_{\max}/\lambda) + D}, \quad (2.1)$$

where $a = 70$, $b = 28.5$, $c = -14.1$, $A = 0.880$, $B = 0.924$, $C = 1.104$. and $D = 0.655$. The values of $\lambda_{\max}$ for the $R(\lambda)$, $G(\lambda)$, and $B(\lambda)$ cones were 560 nm, 540 nm, and 440 nm, respectively. (We shall refer to cones using RGB rather than LMS notation.) These curves are plotted in Figure 3a.

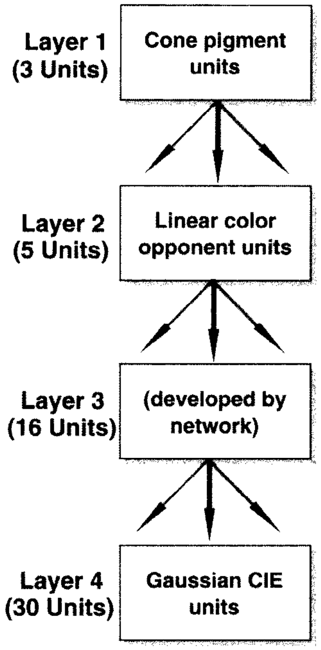


Figure 2: Block diagram showing the four layers of the network model. Each unit in a layer is connected with every unit in the subsequent layer. There were no feedback connections or lateral connections within a layer.

The transformation from wavelength λ to a point in CIE xyz space is given by:

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 1.6452 & -1.3074 & 0.4851 \\ 0.4633 & 0.2882 & -0.0057 \\ 0.0132 & -0.0177 & 2.2468 \end{pmatrix} \begin{pmatrix} R(\lambda) \\ G(\lambda) \\ B(\lambda) \end{pmatrix}. \tag{2.2a}$$

$$\begin{aligned} x &= X/(X+Y+Z) \\ y &= Y/(X+Y+Z) \\ z &= Z/(X+Y+Z). \end{aligned} \tag{2.2b}$$

Since $z = 1-(x+y)$, a two-dimensional plot of x and y, as in Figure 3b, is sufficient to represent this system. The coefficient matrix in Equation 2.2a is a curve-fitting approximation and does not produce exact values of standard CIE tables. Wavelength mixtures are handled by assuming that the total response of a cone reflects a linear integration of responses to components.

This transformation maps any single wavelength to the upper, parabola-like boundary of the chromaticity diagram in Figure 3b, and all mixtures of wavelengths to the interior of the diagram. The mapping is many-to-one, since different mixtures of wavelengths, called metamers in the psychophysical literature, can map to the same point in CIE space (i.e., can produce identical subjective colors or qualia). Since the transformation groups together stimuli that are perceptually similar rather than physically similar, it can be thought of as mapping stimuli into a "qualia space." Because the transformation is many-to-one (a consequence of the integration of the wavelength distribution by the cone tuning curves, as well as the normalization in equation 2.2b), it forms a difficult inverse problem, a problem we have no intention of solving since the model does not aim to recover physical parameters.

Equations 2.1 and 2.2 define an input-output relation that can be used to train a backpropagation network. The output representation chosen for the network was a population of 30 units having overlapping gaussian tuning curves in CIE space, indicated by the circles in Figure 3a. Any mixture of wavelengths defines a point in CIE space, and the responses of all units in the output population at this point can be calculated.

A point in CIE space is enough to define two variables of color appearance, hue and saturation, but a third variable needs to be considered: brightness. White and gray map to the same CIE coordinates, but differ in brightness. The model represented brightness by uniformly scaling the activities of all output units by a multiplicative factor proportional to the luminance of the stimulus, where luminance (L) is defined in equation 2.3. Therefore, CIE coordinates (hue and saturation) were indicated by the relative activities of units in the output population and brightness by the absolute levels of activity.

This model does not handle color constancy (insensitivity of color appearance to changes in the wavelength spectrum incident on a surface) (Land, 1986; Lucassen & Walraven, 1996; Zeki, 1983) or the related phenomenon of simultaneous color contrast (change in color appearance depending on the color properties of spatially adjacent areas) (Brown & MacLeod, 1998; Zaidi, Billibon, Flanigan, & Canova, 1992). Both of these seem to involve long-range lateral spatial interactions within a network (see models by Courtney, Finkel, & Buchsbaum, 1995; Hurlbert & Poggio, 1983; and Wray & Edelman, 1996), which were not included in this model. In a more realistic model, the use of luminance here would be replaced by some sort of ”lightness” computation using lateral connections. The existences of color constancy and simultaneous contrast effects are further examples of the complex and indirect relationship between the physical parameters of a stimulus and the qualia being extracted by neural networks in the brain that has been emphasized above.

The training set for the model consisted of randomly generated wavelength mixtures $\lambda = \{I_1\lambda_1, I_2\lambda_2, \ldots, I_n\lambda_n\}$, with n randomly selected from 1 to 3. The total light intensity (quantal flux) of the wavelength mixture was always constant, $I_{\text{tot}} = 1.0$, but this constant quantal flux was randomly partioned among the different wavelengths so that $I_{\text{tot}} = I_1 + I_2 + \cdots + I_n$. Distribution of the input wavelengths and intensities was nonuniform in such a manner that the sampling of the output CIE space was approximately uniform (for example, on average, light intensities for wavelengths at the blue end of the spectrum were much lower than elsewhere).

The responses of the three cone units of the model layer 1 were defined by equation 2.1 and shown in Figure 5a. The five units of layer 2 were formed by linear combinations of the cone units to form various color opponent units, plus a luminance channel, as in Figure 5b. This was motivated by standard descriptions of color organization in the early visual pathways (Kaiser & Boynton, 1996). These linear transforms were hard-wired in the network as follows:

$$\begin{pmatrix} +r\ -g \\ +g\ -r \\ +y\ -b \\ +b\ -y \\ L \end{pmatrix} = \begin{pmatrix} 1.4840 & -1.4153 & 0.0000 \\ -1.1444 & 1.4673 & 0.0000 \\ 0.3412 & 0.1706 & -0.2983 \\ -0.1712 & -0.0856 & 0.5273 \\ 0.6814 & 0.3407 & 0.0000 \end{pmatrix} \begin{pmatrix} R(\lambda) \\ G(\lambda) \\ B(\lambda) \end{pmatrix} + \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.0 \end{pmatrix}. \tag{2.3}$$

The labels for the color opponent units, such as "$+r-g$," give an indication of the excitatory and inhibitory influences on those units. Layer 3 had 16 units, with initially random characteristics that developed under the influence of the backpropagation algorithm. Finally, the output layer (layer 4), described above, had 30 units whose target properties were defined as a set of overlapping gaussian tuning curves in CIE space (see Figure 6b). Layers 3 and 4 were additionally constrained to have low levels of spontaneous activities.

2.2 Results. The network was readily able to learn the desired input-output transformation, with a diagram of the weights shown in Figure 4. Figure 6b shows the response properties of three output units after training, indicating that they did acquire reasonable approximations to the desired circularly symmetric gaussian receptive field profiles in CIE space. Since wavelength maps onto the upper rim of the CIE chart, the wavelength tuning of these output units can be predicted by observing where they intersect the rim. Units with their centers located near the rim (such as Figure 6b, center) will respond well to a narrow range of wavelengths (average bandwidth: 0.14 nm, half-width at half-height). Those located far from any rim will not respond well to any narrow-band wavelength stimulus. Some units will respond well to white light, and others will not, depending on where they are located relative to the CIE coordinates for "white" (roughly {0.32, 0.32}). Some of these color units will not respond well to either narrow-band wavelength stimuli or white light, but only to a certain range of pastel colors, for example.

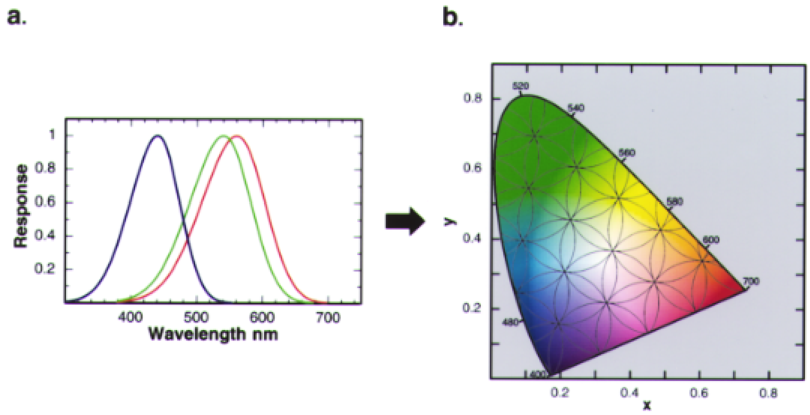
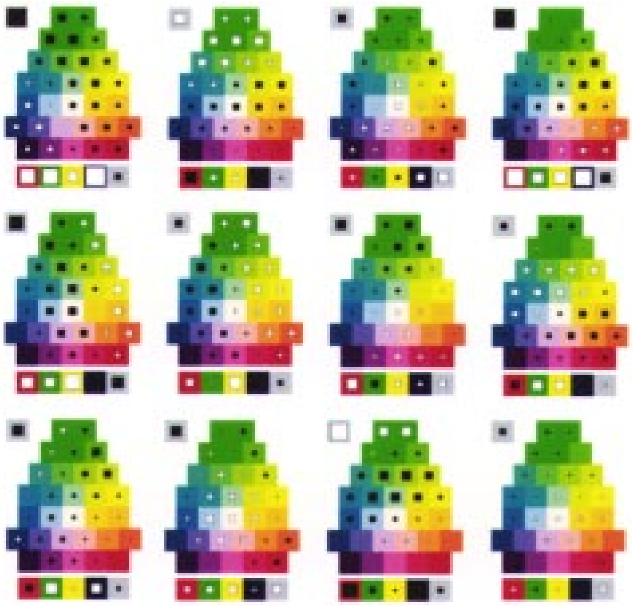


Figure 3: The network model presented here creates a transformation between (a) “red,” “blue,” and “green” cones, with their respective wavelength tuning curves, and (b) a population of 30 overlapping gaussian tuning curves in CIE 1931 color space. Colors in the diagram approximate perceived color at a particular CIE coordinate. Single wavelengths map to the upper rim of the diagram and wavelength mixtures to the interior.



Three examples of units in layer 3 are shown in Figure 6a. The units in this layer were noncolor opponent and divided the CIE color space into two regions, responsive and unresponsive, whose border in the color space was a diagonal line at various positions and angles for different units. Only 12 of the 16 units in layer 3 developed significant weights. The other 4 units were completely unresponsive to any stimuli.

It would be natural to make a comparison between layer 3 units in the model and V4 cells, given that layer's intermediate location between the "striate" and "inferotemporal" cortices of the model (layers 2 and 4). The color properties of V4 cells have been studied by Schein and Desimone (1990). The layer 3 units resemble V4 in several ways. They are noncolor opponent. They have their tuning peaks spread out over many wavelengths. Their wavelength-tuning curves could have one or two peaks, but never more than two (2 out of the 12 units had double peaks). There were differences as well. The ratio of responsivity to white light and optimal colored light was on average 0.26, lower than 0.58 seen in V4 cells. Tuning bandwidths were broader (average: 0.42, half-width at half-height), compared to 0.27 in the data. Overall, the experimental V4 cells have properties that are intermediate to our layer 3 and 4 units. Possibly this may reflect laminar heterogeneity in V4, so that V4 output layers have properties closer to inferotemporal units than do intermediate and input layers, skewing the population statistics collected over all layers.

3 Discussion

Layers 2 through 4 of this model can be viewed as a decoding network for the population of wavelength tuning curves (cones) at the input stage.

Figure 4: *Facing page.* Diagram of weights in the network. There are 12 icons, representing 12 units (out of 16) in layer 3. The other 4 units in layer 3 failed to develop significant weights. The white and black squares in each icon represent the size of excitatory and inhibitory weights between a particular unit in layer 3 and units in both adjacent layers (layers 2 and 4). The 5 squares at the bottom of each icon represent the weights from the five units (4 color opponent and 1 luminance unit) in layer 2 that feed onto a particular unit in layer 3. Among these five units, the gray square represents the luminance unit, and the background colors of the other 4 units indicate the peak sensitivities of the units (red = $+r$ $-g$, green = $+g$ $-r$, yellow = $+y$ $-b$, and blue = $+b$ $-y$). The 30 squares in a roughly triangular array are the weights from a unit in layer 3 to the 30 output units in layer 4. The background color for each weight indicates the color that best excites whatever layer 4 unit that weight connects to. The isolated gray square at the top left of each icon is an imaginary "true unit," which acts as a bias on the activity of the unit it connects to, and influences spontaneous activities. The weights from layer 1 to layer 2 are not shown because they were fixed, as given in equation 2.3.

These layers jointly behave as a family of implicit templates. They serve as a pattern recognition device for the activities in the input population, assigning each pattern to a point in a qualia space (which is encoded by another population). In the decoding process, no attempt was made to recover the physical parameters (wavelengths) that underlay the input population activity. Indeed, such an attempt might be misguided, for there is no behavioral or physiological evidence that such information is ever used or even available at the higher visual areas. For example, if we see white, we have no way of knowing if it was produced by a mixture of narrow-band blue and yellow lights or a continuous broad-band mixture of wavelengths.

If the input to the network were always just a single wavelength, then it might make sense to form an estimate of what that wavelength was. But as the wavelength mixture increases to two, three, or infinitely many (for continuous wavelength distributions, which would be the most typical case), it becomes more difficult to compute an accurate estimate of the physical stimulus, and a different strategy is needed. This strategy may simply be to attach a label to each pattern of activity in the population without worrying about the details of the physical cause of the pattern. A particular ratio of activities in the population of wavelength tuning curves is assigned the label "white," and the distribution of wavelengths that caused it does not matter. The set of all labels forms a qualia space. In this way the system avoids dealing with a difficult inverse problem and instead does something simple but perhaps behaviorally useful. Information is lost in this process, but the information that remains appears useful enough that there are evolutionary advantages to developing such systems.

Noise was not included in this model, but if there were noise, then the output layer 4 would have to form a statistical estimate of what the pattern of activity in layer 1 was (but not an estimate of wavelengths). For this purpose, probabilistic models developed for extracting physical parameters (for example, Pouget et al., 1998) could also be highly useful when transferred to operate with nonphysical parameters. However, there is a certain peculiarity in dealing with an arbitrary qualia space rather than a physical space. For a physical parameter there is an objective, correct answer which the statistical process is trying to estimate. For a qualia space there really is not a "correct answer." Distortions and biases in the decoding process would simply mean that a slightly different transformation was in effect, shifting all color appearances slightly—appearances that were arbitrary to begin with.

The lack of objective criteria in qualia spaces also leads to problems in examining the statistical efficiency of the decoding process. In statistical estimation theory, the Cramer-Rao bound is the theoretically smallest possible variance in the estimate of the "true" parameter value (the "true" color in this case), which can be determined from a given set of noisy input. Systems that produce this minimum-variance estimate (and are unbiased) are called "efficient." But if there is no "true" parameter value to estimate in a qualia

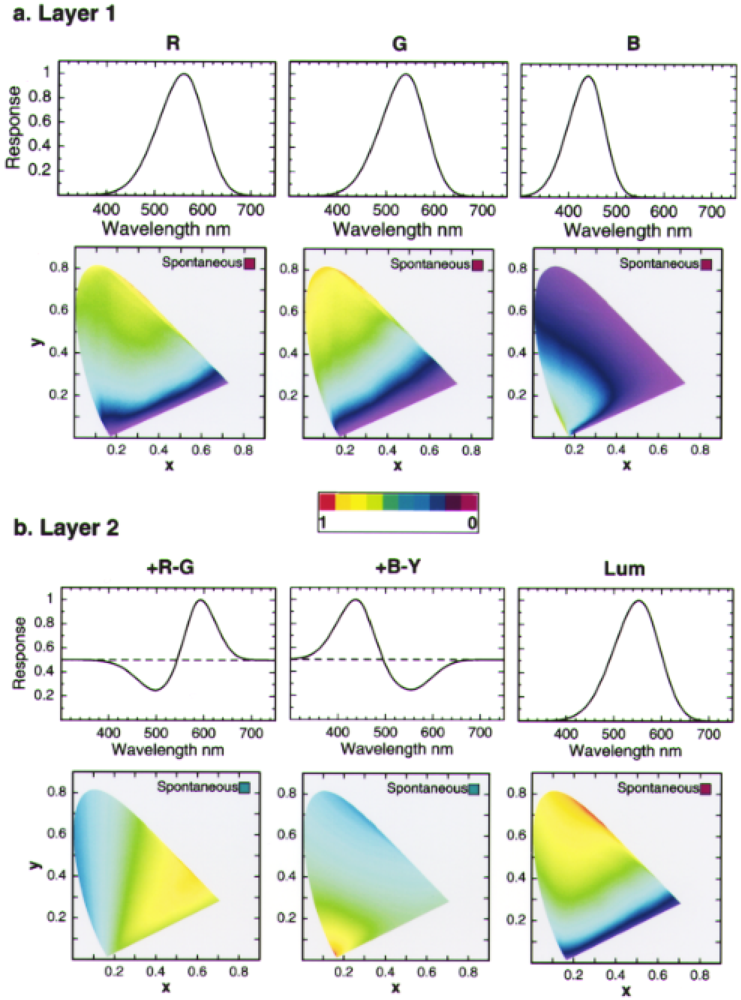


Figure 5: Response tuning properties of three example units from the first two layers of the four-layer network: (a) cone layer and (b) linear color opponent layer. The response tunings are shown as both a function of wavelength and a function of CIE coordinates. Since wavelength maps to the upper rim of the CIE chromaticity diagram, the wavelength response of a unit can also be seen by examining this rim. (In other words, the upper rim of the CIE chart is a distorted version of the *x*-axis of the wavelength tuning graph.) The same color code is used in all CIE charts to indicate responsiveness (red = maximum response, purple = minimum response).

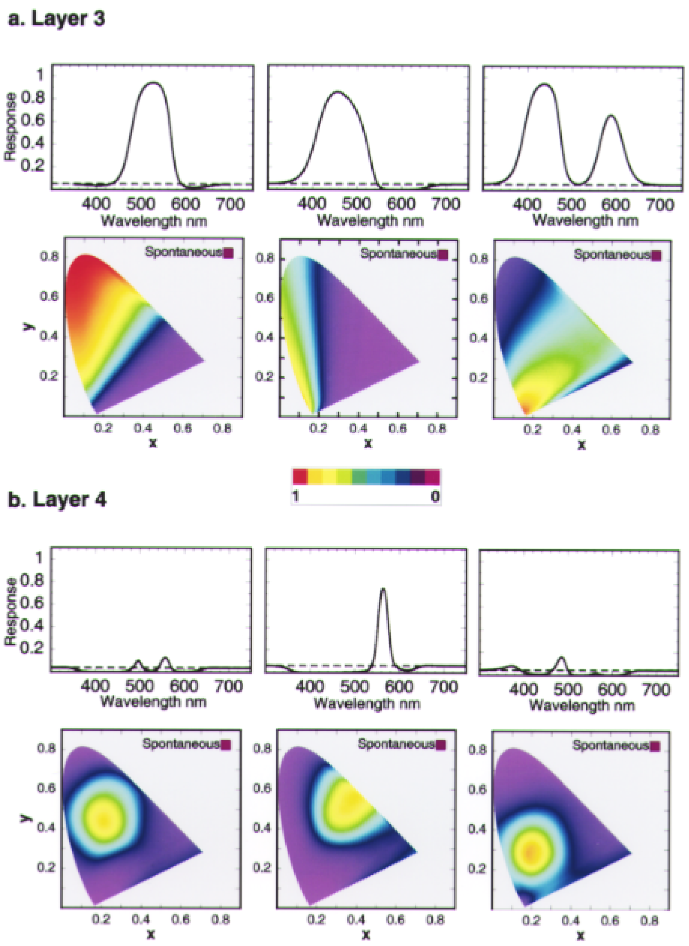


Figure 6: Response tuning properties of three example units from the third and fourth layers of the four-layer network: (a) hidden layer units developed by the network and (b) output gaussian CIE units. The description of the axes given in Figure 5 applies here too.

space, then the notion of statistical efficiency becomes problematic. Perhaps a way around this, if the system is time invariant, is to define the "true" parameter value as being whatever the long-term average output is for a particular input. The extent to which perceptual systems are actually time invariant is an empirical question. Shifts in our subjective color spaces over a period of hours, days, or years would be difficult to detect because there is no standard to compare them to other than fallible memory.

Returning to the color model, one might argue that we have not decoded the population at all but moved from a population code in one representational space to a population code in another representational space. How do you decode that new population? Isn't this the first step of an infinite regress? An answer is that at some point, one simply has to say that a certain pattern of activity is our percept and there is nothing simpler than this pattern to extract. The act of decoding implies something "looking at" the population. At the last link in the chain, one cannot decode or interpret without invoking a homunculus. It is at this point we come up against the classic mystery of how our subjective experiences can arise from brain activity that has bedeviled students of mind and brain for centuries. As for the question of why bother to change the representational space in the first place, it may be that certain representations will increase the salience of those aspects of the stimulus that are behaviorally meaningful, in comparison to other representations which are more simply linked to the physical input.

In redefining the question of population decoding away from extracting physical parameters, we move closer to more abstract and symbolic forms of representation. It would seem that the purpose of the visual system is not so much to transmit a complete set of information and reconstruct a faithful copy of the external world inside the head, but rather to extract and represent behaviorally relevant information (Churchland, Ramachandran, & Sejnowski, 1994). From an evolutionary perspective, all that is required is that this information be useful. There is no requirement that it be an analog representation of the physical world, and indeed internal representations may bear a very indirect and abstract relationship to the physical world. Thus, it may be that it is the properties of the decoding networks in the cortex and the transforms they define, rather than the patterns of neural activity per se, that will prove to be more central to our understanding of the neural substrate of qualia.

Acknowledgments

Portions of this article were presented at the European Conference on Visual Perception, Oxford, England, 1998. S. R. L. thanks Keiji Tanaka for his kind support.

References

Adelson, E., & Movshon, J. A. (1982). Phenomenal coherence of moving visual patterns. *Nature, 300*, 523–525.

Anderson, C. H. (1994). Basic elements of biological computational systems. *International Journal of Modern Physics C, 5*, 135–137.

Brown, R. O., & MacLeod, D. I. A. (1998). Color appearance depends on the variance of surround colors. *Current Biology, 7*, 844–849.

Buchsbaum, G., & Goldstein, J. L. (1979). Optimum probabilistic processing in colour perception. II. Colour vision as template matching. *Proceedings of the Royal Society of London B, 205*, 245–266.

Chen, L. L., & Wise, S. P. (1997). Conditional oculomotor learning: Population vectors in the supplementary eye field. *Journal of Neurophysiology, 78*, 1166–1169.

Churchland, P. S., Ramachandran, V. S., & Sejnowski, T. J. (1994). A critique of pure vision, In C. Koch & J. Davis (Eds.), *Large-scale neuronal theories of the brain* (pp. 23–60). Cambridge, MA: MIT Press.

Courtney, S. M., Finkel, L. H., & Buchsbaum, G. (1995). Network simulations of retinal and cortical contributions to color constancy. *Vision Research, 35*, 413–434.

De Valois, R. L., & De Valois, K. (1993). A multi-stage color model. *Vision Research, 33*, 1053–1065.

Georgopoulos, A. P., Schwartz, A., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science, 233*, 1416–1419.

Helmholtz, H. von. (1962). *Physiological optics*. New York: Dover, 1962; Reprinted of English translation by J. P. C. Southall for the Optical Society of America (1924) from the 3rd German edition of Handbuch der physiologischen Optik (Voss, Hamburg. 1909).

Hurlbert, A. C., & Poggio, T. A. (1983). Synthesizing a color algorithm from examples. *Science, 239*, 482–485.

Kaiser, K. P., & Boynton, R. M. (1996). *Human color vision*. (2nd ed.) Washington D.C.: Optical Society of America.

Komatsu, H. (1998). Mechanisms of central color vision. *Current Opinion in Neurobiology, 8*, 503–508.

Komatsu, H., Ideura, Y., Kaji, S., & Yamane, S. (1992). Color selectivity of neurons in the inferior temporal cortex of the awake macaque monkey. *Journal of Neuroscience, 12*, 408–424.

Lamb, T. D. (1995). Photoreceptor spectral sensitivities: Common shape in the long-wavelength region. *Vision Research, 35*, 3083–3091.

Land, E. H. (1986). Recent advances in retinex theory. *Vision Research, 26*, 7–21.

Lee, C., Rohrer, W. H., & Sparks, D. L. (1988). Population coding of saccadic eye movements by neurons of the superior colliculus. *Nature, 332*, 357–360.

Lehky, S. R., & Sejnowski, T. J. (1988). Network model of shape-from-shading: Neural function arises from both receptive and projective fields. *Nature, 333*, 452–454.

Lehky, S. R., & Sejnowski, T. J. (1990a). Neural model of stereoacuity and depth interpolation based on a distributed representation of stereo disparity. *Journal of Neuroscience, 10*, 2281–2299.

Lehky, S. R., & Sejnowski, T. J. (1990b). Neural network model of the visual cortex for determining surface curvature from images of shaded surfaces. *Proceedings of the Royal Society, London, B, 240*, 251–278.

Lucassen, M. P., & Walraven, J. (1996). Color constancy under natural and artificial illumination. *Vision Research, 36*, 2699–2711.

Nowlan, S. J., & Sejnowski, T. J. (1995). A selection model for motion processing in area MT of primates. *Journal of Neuroscience, 15*, 1195–1214.

Oram, M. W., Földiák, P., Perrett, D. I., & Sengpiel, F. (1998). The "ideal homunculus": Decoding neural population signals. *Trends in Neuroscience, 21*, 259–265

Paradiso, M. A. (1988). A theory for the use of visual orientation information which exploits the columnar structure of the striate cortex. *Biological Cybernetics, 58*, 35–49.

Pouget, A., & Thorpe, S. J. (1991). Connectionist models of orientation identification. *Connection Science, 3*, 127–142.

Pouget, A., Zhang, K., Deneve, S., & Latham, P. E. (1998). Statistically efficient estimations using population coding. *Neural Computation, 10*, 373–401.

Qian, N., Anderson, R. A., & Adelson, E. H. (1994). Transparent motion perception as detection of unbalanced motion signals. I. Psychophysics. *Journal of Neuroscience, 14*, 7357–7366.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1: foundations* (pp. 318–368). Cambridge: MIT Press.

Salinas, E., & Abbott, L. (1994). Vector reconstruction from firing rates. *Journal of Computational Neuroscience, 1*, 89–97.

Salinas, E., & Abbott, L. (1995). Transfer of coded information from sensory to motor networks. *Journal of Neuroscience, 15*, 6461–6471.

Sanger, T. D. (1996). Probability density estimation for the interpretation of neuronal population codes. *Journal of Neurophysiology, 76*, 2790–2793.

Schien, S. J., & Desimone, R. (1990). Spectral properties of V4 neurons in the macaque. *Journal of Neuroscience, 10*, 3369–3389.

Seung, H. S., & Sompolinsky, H. (1993). Simple models for reading neuronal population codes. *Proceedings of the National Academy of Sciences, USA, 90*, 10749–10753.

Shadlen, M. N., Britten, K. H., Newsome, W. T., & Movshon, J. A. (1996). A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *Journal of Neuroscience, 16*, 1486–1510.

Snippe, H. P. (1996). Parameter extraction from population codes: A critical assessment. *Neural Computation, 8*, 511–529.

Stoner, G. R., & Albright, T. D. (1992). Neural correlates of perceptual motion coherence. *Nature, 358*, 412–414.

Usui, S., Nakauchi, S., & Miyake, S. (1994). Acquisition of the color opponent representation by a three-layered neural network. *Biological Cybernetics, 72*, 35–41.

Vogels, R. (1990). Population coding of stimulus orientation by striate cortical cells. *Biological Cybernetics, 64*, 25–31.

Wilson, H. R., & Gelb, D. J. (1984). Modified line-element theory for spatial frequency and width discrimination. *Journal of the Optical Society of America A, 1*, 124–131.

Wilson, M. A., & McNaughton, B. L. (1993). Dynamics of the hippocampal ensemble code for space. *Nature, 261*, 1055–1058.

Wray, J., & Edelman, G. M. (1996) A model of color vision based on cortical reentry. *Cerebral Cortex, 6*, 701–716.

Wyszecki, G., & Stiles, W. S. (1982). *Color science: Concepts and methods, quantitative data and formulae.* (2nd ed.). New York: John Wiley.

Young, M. P., & Yamani, S. (1992). Sparse population encoding of faces in the inferotemporal cortex. *Nature, 256*, 1327–1331.

Young, T. (1802). II. The Bakerian Lecture. On the theory of light and colors.*Philosophical Transactions of the Royal Society of London, 91*, 12–48.

Zaidi, Q., Billibon, Y., Flanigan, N., & Canova, A. (1992). Lateral interactions within color mechanisms in simultaneous induced contrast. *Vision Research, 32*, 1695–1707.

Zeki, S. (1983). Colour coding in the cerebral cortex: the responses of wavelength-selective and colour-coded cells in monkey visual cortex to changes in wavelength composition. *Neuroscience, 9*, 767–781.

Zemel, R. S., Dayan, P., & Pouget, A. (1998). Probabilistic interpretation of population codes. *Neural Computation, 10*, 403–430.

Zhang, K., Ginsburg, I., McNaughton, B. L., & Sejnowski, T. J. (1998). Interpreting neuronal population activity by reconstruction: Unified framework with application to hippocampal place cells. *Journal of Neurophysiology, 79*, 1017–1044.

Zohary, E. (1992). Population codes of visual stimuli by cortical neurons tuned to more than one dimension. *Biological Cybernetics, 66*, 265–272.

Received February 18, 1997; accepted October 29, 1998.