The Constructive Nature of Color Vision and Its Neural Basis

Dissertation
zur Erlangung des Grades eines Doktors der Naturwissenschaften
der Mathematisch-Naturwissenschaftlichen Fakultät
und
der Medizinischen Fakultät
der Eberhard-Karls-Universität Tübingen

vorgelegt
von

Michael Mario Bannert
aus Engelskirchen, Deutschland
April 2017

Tag der mündlichen Prüfung	: 24. Oktober 2017
Dekan der Math.-Nat. Fakultät	: Prof. Dr. W. Rosenstiel
Dekan der Medizinischen Fakultät	: Prof. Dr. I. B. Autenrieth
1. Berichterstatter	: Prof. Dr. Andreas Bartels
2. Berichterstatter	: PD Dr. Axel Lindner
3. Berichterstatter	: Prof. Dr. Karl Gegenfurtner
Prüfungskommission:	Prof. Dr. Andreas Bartels
	Prof. Dr. Matthias Bethge
	Prof. Dr. Nikos Logothetis
	PD Dr. Axel Lindner

Ich erkläre, dass ich die zur Promotion eingereichte Arbeit mit dem Titel:

"The Constructive Nature of Color Vision and Its Neural Basis"

selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

Tübingen, ______________________ ______________________________

Datum Unterschrift

Meinem Vater Peter Bannert

The Constructive Nature of Color Vision and Its Neural Basis

Michael M. Bannert
Werner Reichardt Centre for Integrative Neuroscience, Eberhard Karls University
Bernstein Center for Computational Neuroscience
Max Planck Institute for Biological Cybernetics
Department of Psychology, Eberhard Karls University
Tübingen

Abstract

Our visual world is made up of colored surfaces. The color of a surface is physically determined by its reflectance, i.e., how much energy it reflects as a function of wavelength. Reflected light, however, provides only ambiguous information about the color of a surface as it depends on the spectral properties of both the surface and the illumination. Despite the confounding effects of illumination on the reflected light, the visual system is remarkably good at inferring the reflectance of a surface, enabling observers to perceive surface colors as stable across illumination changes. This capacity of the visual system is called color constancy and it highlights that color vision is a constructive process. The research presented here investigates the neural basis of some of the most relevant aspects of the constructive nature of human color vision using machine learning algorithms and functional neuroimaging. The experiments demonstrate that color-related prior knowledge influences neural signals already in the earliest area of visual processing in the cortex, area V1, whereas in object imagery, perceived color shared neural representations with the color of the imagined objects in human V4. A direct test for illumination-invariant surface color representation showed that neural coding in V1 as well as a region anterior to human V4 was robust against illumination changes. In sum, the present research shows how different aspects of the constructive nature of color vision can be mapped to different regions in the ventral visual pathway.

Keywords: Color vision, human fMRI, pattern classification

Contents

1 INTRODUCTION 2

1.1 Visual Perception as Inference 2

1.2 Color Vision as a Model System for Perceptual Inference 3

1.3 Color Constancy and the Problem of Surface Color Perception 4

2 THE FUNCTIONAL NEUROANATOMY OF COLOR PROCESSING 9

2.1 Retina and Lateral Geniculate Nucleus 9

2.2 Primary Visual Cortex 12

2.3 Area V2 14

2.4 Area V4 16

2.5 Color Processing in Other Extrastriate Brain Regions 19

3 RESEARCH AIMS 20

4 Experiment I: Decoding the Yellow of a Gray Banana 22

5 Experiment II: Invariance of Surface Color Representations Across Illuminant Changes in the Human Cortex 38

6 Experiment III: Human V4 Encodes Object Color in Visual Imagery 67

7 SUMMARY AND CONCLUSION 82

8 ACKNOWLEDGMENTS 84

1 INTRODUCTION

1.1 Visual Perception as Inference

Since the cognitive revolution (Broadbent, 1958; Neisser, 1967), visual perception is frequently viewed as an information-processing task. It is thought to comprise both the *processes* of acquiring and transforming information about the visual environment but also *representations* for storing that information. The visual system detects visual features in the sensory input step by step to arrive at a mental representation of the physical properties in the world, which can then be used for reasoning and to guide behavior (Marr, 1982).

The conceptualization of vision as an information-processing task has stimulated a lot of research into the visual system and enjoys wide acceptance among visual neuroscientists to this day. Despite its success, the approach remains relatively mute on the question how visual percepts exactly relate to the physical world. This has been a central question in the vision sciences at last since the beginnings of its experimental study by Fechner (1860), who, at the time, believed that the sensory input completely specifies perceptual content.

More recently, this question experienced increased interest within a framework that regards visual perception as a Bayesian inference process (Knill & Richards, 1996; Yuille & Kersten, 2006). According to this view, the purpose of visual perception is to infer the physical causes of sensory stimulation – a challenge that is commonly referred to as "inverse optics problem". Visual percepts are not mental representations of the physical properties in the environment but rather correspond to inferences about probable causes in the physical world that explain the sensory input. Formally, this probability can be computed using Bayes theorem as the posterior probability of world states given the sensory input and prior knowledge about those states. This probability combines prior knowledge about world states with the causal structure between the stimulus and sensations. Perceptual systems thus embody generative models that capture the causal structure between world states and sensory inputs.

The perception-as-inference hypothesis has been applied to modern cognitive neuroscience as well. Accordingly, theoretical extensions to this hypothesis have been developed that regard the brain's architecture as an implementation of a generative model. The hypothesis has become known as the "Bayesian brain hypothesis" in the field (Friston, 2010; Knill & Pouget, 2004; Pouget, Beck, Ma, & Latham, 2013).

The idea of perception as Bayesian inference is frequently associated with Hermann von Helmholtz (Westheimer, 2008). He put forward the hypothesis that our perceptual responses to stimulation correspond to the inferred causes that, under normal viewing conditions, would give rise to the same sensory excitations registered by the visual system (Helmholtz, 1867). Perceptual systems rely on such inferences to link the only information that is available to them, namely, the patterns of excitations in the nervous system to the outside world, which, in contrast, remains necessarily inaccessible. Visual percepts thus are not passively induced by the sensory input but instead represent the outcome of an interaction between sensory input and the perceiving system itself.

1.2 Color Vision as a Model System for Perceptual Inference

It may not be surprising that Helmholtz dedicated a lot of his research effort to the study of color perception. It exemplifies, perhaps more than other domains in the vision sciences, the inferential nature of perception. Consequently, as he wrote in the second part of his *Handbuch der Physiologischen Optik* (Helmholtz, 1867, pp. 406–408) and as quoted by Zeki (1990), the perception of color depends crucially on unconscious inference ("Urtheil") rather than on mere sensation ("Empfindung"). After all he regarded color mainly as a meaningful visual cue to the extent that it indicates a physical property of surrounding objects, i.e., the proximal stimulus, which eludes direct access of the sensory system. Since different colors can simultaneously be present in the same visual field location (e.g., consider a surface color viewed under a certain illumination through a pair of sunglasses)[1], he suggested that the illuminant must somehow be "discounted" to infer object color.

Hence, when one considers what it means to perceive color, the constructive nature of color vision and the inferences that it involves become very apparent: just as visual perception generally consists in determining *what* objects exist in our visual field and *where* they are located (Marr, 1982, p. 3), color vision in particular can analogously be construed as being about determining *what color* these objects are.

It is important to note how this conceptualization of color vision differs from another common definition according to which color vision refers to the ability to discriminate two isoluminant lights based on their spectral composition (e.g., Solomon & Lennie, 2007). Such discriminatory skill is of course closely linked to the assignment of colors to surfaces because, without this ability, it would at most be possible to assign different lightness levels to surfaces. It is obvious that the richness of color experience goes beyond the mere perception of how much light a surface reflects. But it is insufficient for perceiving what color a given surface is because differences in surface color may not translate into differences with respect to the spectral composition of the light impinging on the retina.

[1] Note that this peculiarity distinguishes color from other visual attributes such as shape because no two shapes can occupy the same visual field location.

Most importantly, this definition misses the environmental and behavioral context in which color vision occurs. For in order to understand color vision, it is essential to consider what the visual system does in perceiving color. The following section will explicate some of the computational challenges the visual system has to solve.

1.3 Color Constancy and the Problem of Surface Color Perception

The physical stimulus underlying all vision is light. Light consists of electromagnetic waves that is characterized by a spectral power distribution (SPD). The SPD is a description of how much radiant energy the light contains at each wavelength along the electromagnetic spectrum. The spectrum of light that is visible to the human eye lies between approximately 400 nm and 700 nm. This means that the sensory system responds to light that falls within this range.

However, the relationship between the light reaching the eye and surface color is far from trivial. Consider for instance the light reflected off an object viewed in daylight on a cloudless morning. It contains relatively more radiant energy in the short-wavelength range of the visible spectrum whereas it will contain more power in the long-wavelength spectrum when viewed in candlelight in the evening. Despite such considerable variability in the wavelength composition of the light falling into the eye that is brought about by illumination changes, the color of surfaces appears rather stable in normal human vision. The phenomenon that surface color perceptions remain relatively stable across changes in illumination is referred to as “color constancy”.

It may seem natural to us that our perceptions of surface color do not change much when the illumination changes. However, despite the apparent ease with which human observers perceive the colors of objects, this impression stands in stark contrast with the computational problem posed by color constancy.

The physical property determining the color of an object surface is called “reflectance”. The reflectance of a surface specifies the proportion of incident light that it reflects at each wavelength when it is under illumination (Figure 1). The reflected light then travels from the surface to the viewer and eventually stimulates the receptors within the retina. The light incident on the retina hence is the proximal stimulus in color vision whereas surface reflectance represents the distal stimulus that needs to be inferred. It is given by:

$$C(\lambda) = S(\lambda) \cdot E(\lambda), \tag{1}$$

where λ is wavelength, $C(\lambda)$ is the SPD of the light reflected off the surface (i.e., the

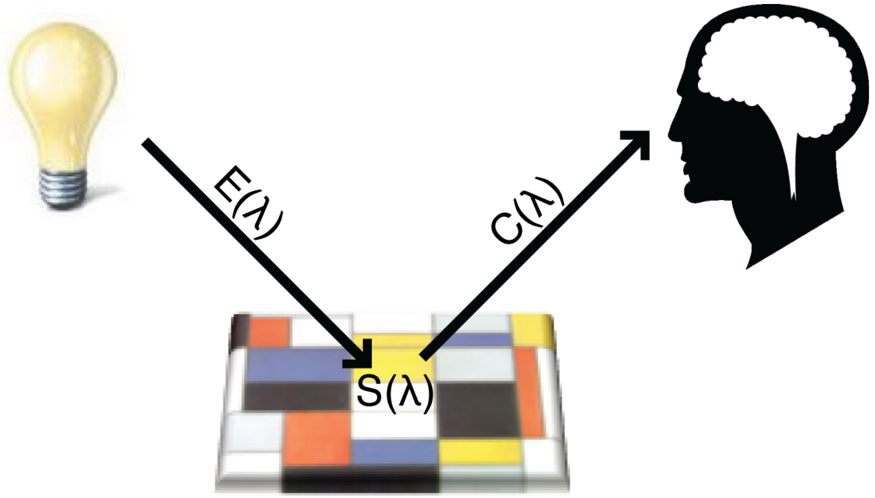

Figure 1. **Color constancy problem.** Light with wavelength distribution $E(\lambda)$ is reflected by a surface with reflectance $S(\lambda)$. The reflected light reaching the observer has a wavelength distribution that is the product of the two, $C(\lambda)$. The reflectance $S(\lambda)$ of the surface, and hence its color, therefore is not directly accessible to the visual system.

proximal stimulus), $S(\lambda)$ is the reflectance[2] defining the color of a surface (i.e., the distal stimulus), and $E(\lambda)$ is the illumination under which the surface is viewed.

What follows from Equation 1 is that changing the illumination while keeping surface reflectance identical results in a change in the proximal stimulus $C(\lambda)$. This means that a surface can be made to reflect lights that are physically very different. It also means that there is an infinite number of pairs of illuminations and surface reflectances that can all give rise to reflected lights with identical SPDs.

Thus it is obvious that the proximal stimulus cannot unambiguously indicate surface color as it confounds information about reflectance with illumination: in order to solve Equation 1 for surface reflectance $S(\lambda)$ and thus retrieve surface color from the proximal stimulus $C(\lambda)$, one needs to know illumination $E(\lambda)$; but in order to solve for the illumination, one needs to know the surface reflectance. Brainard and Maloney (2011) summarized these two properties as the "surface-illuminant duality".

The computational problem of surface color perception is further complicated by the fact that even the SPD of the proximal stimulus (reflected light) is not available to the visual

[2] A reflectance function of the form $S(\lambda)$ is of course an oversimplification of how surfaces in the real world reflect light. More complex reflectance models are conceivable that describe how the appearance of a surface depends on the angles at which it is illuminated and at which it is viewed, respectively (see e.g., Koenderink, 2010, pp. 668–672).

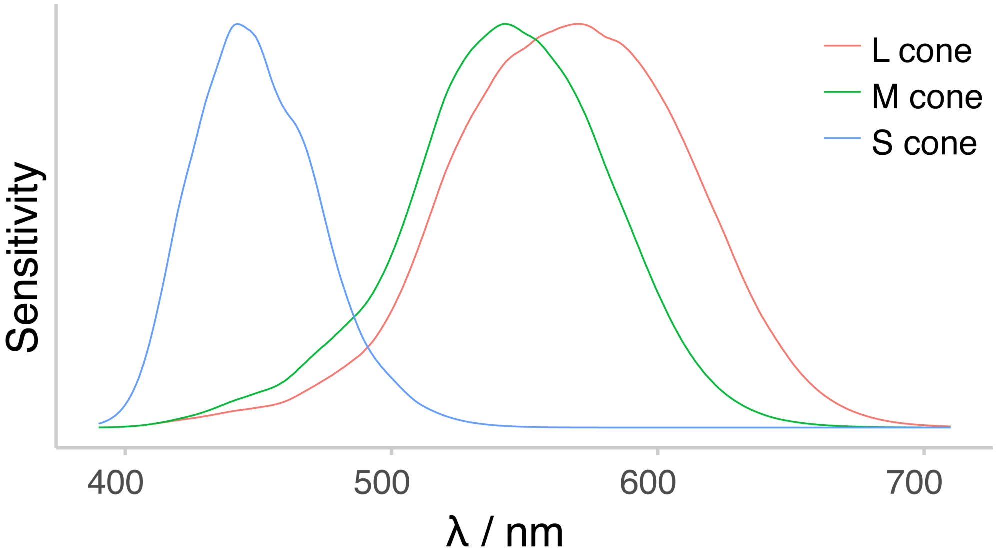

Figure 2. **Human cone spectral sensitivities.** Sensitivity profiles of the three cone classes with their peaks at short, middle, or long wavelengths (from Stockman & Sharpe, 2000).

system either. This is caused by the way the cone receptors transduce stimulus energy into neural signals. There are three different classes of cones with different sensitivities as a function of wavelength, which yields a three-dimensional representation of the SPD. The response ρ_k of a cone from class $k \in \{1, 2, 3\}$ depends on its corresponding sensitivity function $R_k(\lambda)$ and the proximal stimulus $C(\lambda)$ in the following way:

$$\rho_k = \int R_k(\lambda) \cdot C(\lambda)\,\mathrm{d}\lambda \tag{2}$$

The visible spectrum defines the integration interval. The three cone classes differ in that they are most sensitive at long, middle, or short wavelengths within the visible spectrum (Figure 2). They are therefore often referred to as L, M, and S cones. The responses of these three cone types form the starting point for neural processing in the visual system. Just as light with a given SPD entering the eye can arise from several different combinations of illuminations and reflecting surfaces, so may identical triplets of cone excitations be caused by a multitude of such combinations. Moreover, various spectrally different lights can elicit exactly the same triplets. This becomes obvious if one considers that the same cone response to a monochromatic light (i.e., a light with an extremely narrow, peaked SPD) presented at a given wavelength, can be obtained by another monochromatic light of (say) doubled intensity that is presented at another wavelength where cone sensitivity is only half as large.

The fact that different lights can elicit the same cone responses is called "metamerism" and the two lights are referred to as metamers of each other.

The estimation of surface color is hence an extremely ill-posed problem. However, the physics of the visual environment in which surface color perception takes place imposes a few constraints on the algorithms that may be used by the visual system (reviewed by Maloney, 1999): both illuminants and surfaces are often smooth and typically show statistical regularities that allow the dimensionality of SPDs and surface reflectances to be reduced. A projection of SPDs and reflectance functions to a lower-dimensional space makes it easier to obtain approximate solutions to the computational problem. Some algorithms try to estimate the illumination by making simplifying assumptions. The gray-world algorithm for example estimates the illumination by assuming that the average surface reflectances are gray. In scenes with a 3D structure, additional cues to the illuminant, among others, are specular highlights and mutual reflections.

Many of these algorithms can be implemented in Bayesian models as well (Maloney, 1999). Approaching the color constancy problem in terms of Bayesian inference makes it possible to formally integrate prior knowledge about reflectances and illuminations in the stimulus environment ($P(C)$) with the likelihood of the sensory input ($P(S|C)$) with respect to a generative model. This yields an estimate of the probable causes in the stimulus environment given the sensory input ($P(C|S)$), which in Helmholtz's view constitutes our perception. The central equation in these models is Bayes theorem, which is:

$$P(C|S) = \frac{P(S|C)}{P(S)}P(C) \tag{3}$$

$P(S)$ is the "evidence" which is the probability of the sensory input considering all possible causes. Such Bayesian approaches have been applied to color (Brainard & Freeman, 1997; Brainard et al., 2006) and lightness (Allred & Brainard, 2013; Olkkonen, Saarela, & Allred, 2016) perception. The significance of the prior probability becomes evident when considering how an observer's knowledge about objects may bias the estimation of surface reflectance. For example some so-called color-diagnostic objects are typically associated with a particular surface color (e.g., bananas are yellow) and may thus influence how color is processed. Hering (1920), who recognized this influence on color perception, referred to the prior knowledge about the typical color of an object as "memory color".

To conclude, the problem of surface color perception shows that it is impossible to deduce the color of an object directly from the light that is reflected from it. As Shevell and Kingdom (2008, p. 144) put it, "color is not in light" but is instead determined, in interaction with the viewing context, by the neural responses to it.

This raises the question how the visual system resolves the ambiguity in the light signal stimulating its sensors. What other sources of information may it use in order to achieve color constancy? The research presented in this dissertation focuses on the question how the brain represents color that is not linked to the immediate chromatic input. In other words, the question becomes: if color is not in light, where else can it be found?

2 THE FUNCTIONAL NEUROANATOMY OF COLOR PROCESSING

This section will review the current knowledge of color processing in the central nervous system starting at the cones and covering the cortical areas of the ventral visual pathway V1, V2, and V4 that are involved in color vision (e.g., Conway et al., 2010; Solomon & Lennie, 2007). Although the current understanding of the relationship between color vision and its neural architecture is still poor, there are a few instances in which it is possible to establish meaningful links between perception and its underlying brain mechanisms.

The behavioral and perceptual phenomena adduced to explain how the nervous system gives rise to color experience are quite diverse ranging from color-matching experiments and the color of after-images to the preservation of perceptual similarity spaces and the effect of chromatic context on local color experience. Ideally, however, a complete theory of color vision needs to account for the whole range of psychophysical and neurobiological phenomena.

2.1 Retina and Lateral Geniculate Nucleus

There are two types of photoreceptors in the human retina: cones and rods. Both are sensitive to light because they contain photopigments (opsin and rhodopsin, respectively) that undergo a chemical change when struck by a photon. Rod receptors are most sensitive at low light intensities. Visual processing is therefore based on rods in dark viewing conditions. Cone receptors on the other hand require relatively higher light intensities and thus contribute to visual processing under normal conditions at daytime. As already mentioned in the previous section, one distinguishes between L (long-wavelength), M (middle-wavelength), and S (short-wavelength) cones depending on the location of their sensitivity peaks in the visible spectrum.

A relevant property of chromatic analysis at the cone level is the "univariance principle". It states that the response of each of these photoreceptors elicited by a photon does not depend on the spectral properties or the intensity of the triggering stimulus. The strength of a photoreceptor response thus conflates the intensity and the spectral properties of the stimulus signal. The visual system requires a comparison between receptors with different wavelength sensitivities to distinguish a change in receptor outputs due to intensity from a change in the wavelength spectrum. Since rod receptors all have the same spectral sensitivity, rod-based vision does not support color vision.

The excitation patterns of cones therefore constitute the retinal basis for human color vision. As there are three classes of cones, human color vision is trichromatic, as has first

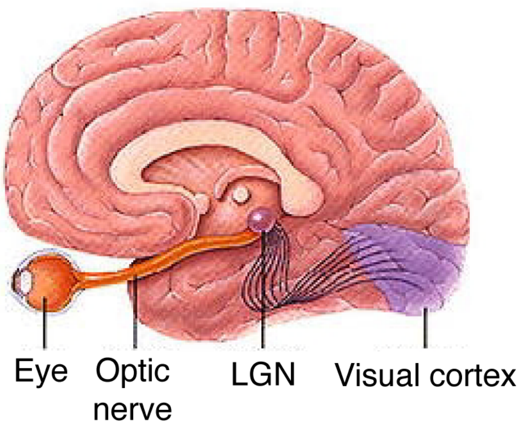

Figure 3. **Visual pathway from eye to cortex.** In the eye the physical light is transduced into neural signal, which is propagated via the lateral geniculate nucleus (LGN) to the visual cortex.

been hypothesized in the early 19th century by Thomas Young (1802). Young as well as Helmholtz carried out color-matching experiments in the 19th centuries that supported this view. These experiments showed that normal viewers required three spectrally independent single-wavelength lights ("primaries") to produce a perceptual match for any other single-wavelength light suggesting a three sensor mechanism underlying color vision. Viewers with only two cone classes hence need only two primaries to produce color matches. Their theory has been known as the "Young-Helmholtz theory of color vision". (The fact that any color can be matched with three primaries is exploited for instance in television and computer displays.)

At the the physiological level the neural signals generated in the photoreceptors are propagated through a network of amacrine, bipolar, and horizontal cells to the retinal ganglion cells (Field & Chichilnisky, 2007). Interestingly, Horiguchi, Winawer, Dougherty, and Wandell (2013) suggested that the light-sensitive protein melanopsin found in some retinal ganglion cells, previously thought to be involved only in non-visual functions like the control of circadian rhythm and pupil dilation, also contribute to human color perception albeit only in the peripheral visual field.

Most retinal ganglion cells (RGC) in primates project to the lateral geniculate nucleus (LGN, see Figure 3) in the thalamus while a minority of them project to the superior colliculi. A prominent distinction has been made between midget RGCs sending signals mainly to the parvocellular LGN layers (P pathway), parasol RGCs sending signals to magnocellular LGN layers (M pathway), and the small bistratified RGCs that project to the superior colliculus and koniocellular LGN layers (K pathway) (Dacey & Lee, 1994; Hendry & Reid, 2000; Leventhal, Rodieck, & Dreher, 1981; Rodieck, Binmoeller, & Dineen, 1985). Besides these well known pathways originating in the retina, a total of 11 pathways have been documented so far (Field & Chichilnisky, 2007).

RGCs combine inputs from cones with different spectral sensitivities. Parasol cells combine L and M cones in an additive way to represent a luminance signal. Midget cells on the other hand are sensitive to contrasts between L and M cones while the small bistratified cells respond to contrasts between S cones and a mixture of L and M cones. Such signal combinations were expected based on Ewald Hering's "opponent process-theory of color vision" (Hering, 1920), which postulated that color perception is mediated by comparisons between different color signals. This is exemplified for instance in the observation that adaptation to a color elicits an opponent afterimage: when adapting to yellow, a blue afterimage is experienced afterwards. Accordingly, a yellow afterimage is seen after adapting to blue. Note also the close relation to the univariance principle, which states that color vision presupposes the comparison of chromatic signals from at least two receptors to disentangle the intensity and the wavelength spectrum of light.

Experimental support for cone opponency (see Figure 4) in primate color vision was first found in the macaque LGN (De Valois & Jacobs, 1968). Opponent neuron responses reflect a combination of cone signals causing some of them to respond preferentially to "red-green" contrast and others to "blue-yellow" contrast. In the parvocellular layer, the majority of cells feature L-M cone opponency while activity in only a small fraction of cells is modulated by S cone input to produce S-(L+M) cone opponency (Derrington, Krauskopf, & Lennie, 1984). Together with the L+M luminance signal conveyed mainly by magnocellular neurons, these two color-opponent mechanisms form the so-called "cardinal axes" of the Derrington-Krauskopf-Lennie (DKL) color space. Modulation by S cone contrast only is commonly seen in neurons in the koniocellular LGN layers (Martin, White, Goodchild, Wilder, & Sefton, 1997; Szmajda, Buzás, Fitzgibbon, & Martin, 2006). However, S cone input has not been observed in the superior colliculus (Martin & Lee, 2014).

Non-invasive neuroimaging experiments have successfully disentangled visual responses in parvocellular and magnocellular layers of the human LGN as well (Denison, Vu, Yacoub, Feinberg, & Silver, 2014; P. Zhang, Zhou, Wen, & He, 2015). Response characterstics of P and M layers when stimulated with chromatic and achromatic stimuli at different spatial and temporal frequencies are in general agreement with electrophysiological evidence from nonhuman primates. Moreover, the effects of feature-based attention to motion versus color have also been mapped to ventromedial versus dorsolateral portions of

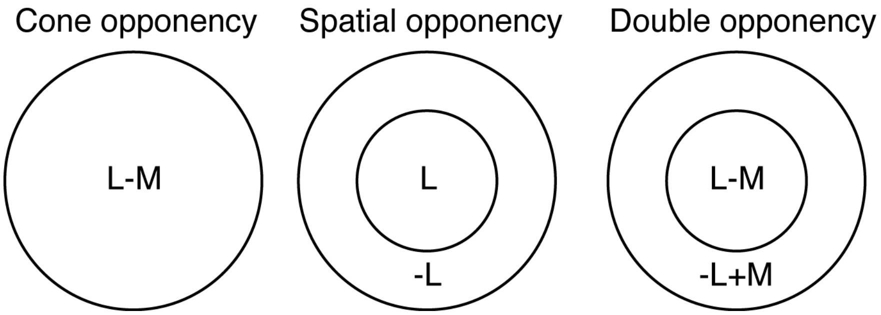

Figure 4. **Single and double opponency.** Cone opponency combines inputs with different signs across the whole receptive field. Spatial opponency combines signals from a single cone class with different signs in center and surround compartments. Double opponency responds to cone contrasts of opposite signs in center and surround compartments.

LGN, roughly corresponding to the locations of M and P layers, respectively (Schneider, 2011). The functional and structural anatomy of the subcortical visual system in primates thus exhibits a certain degree of specialization for color processing. M layers are thought to process motion and luminance signals while P layers are thought to carry chromatic and fine-grained spatial information (Schiller, Logothetis, & Charles, 1990).

2.2 Primary Visual Cortex

The large majority of visual input enters the cortex via connections from LGN to the primary visual cortex (V1). A comparison between response properties of V1 neurons and parvocellular LGN neurons provides a window to the signal transformations occurring at this processing step. Compared to their parvocellular counterparts in LGN, V1 neurons thus feature different cone weighting, more frequent nonlinear output expansion, and a more diverse distribution of hue preferences that does not align with the L-M and S-(L+M) cardinal axes (De Valois, Cottaris, Elfar, Mahon, & Wilson, 2000; Lennie, Krauskopf, & Sclar, 1990). The seeming deviation from (opponent) cone axes can be explained by the fact that a large fraction of V1 neurons combine cone responses nonlinearly (Horwitz & Hass, 2012). When quadratic models are fitted to their firing patterns, their tuning again matches perceptually relevant cone directions although they are only slightly more complex than the linear models.

The functional specialization for processing chromatic information primarily in parvocellular layers in LGN is to some degree also mirrored in primary visual cortex: V1 cells located in patches of cortex that stained for cytochrome oxidase (CO) differed markedly from other V1 neurons. Cells within these so-called “(CO) blobs” showed strong chromatic selectivity. But they were not tuned for orientation unlike neurons located in “inter-blob regions” (Livingstone & Hubel, 1984, 1988; Lu & Roe, 2008; Ts’o & Gilbert, 1988). In keeping with the suggested functional segregation, human fMRI showed that shifting attention between the contrast and the orientation of gratings leads to stronger response changes in V1 when the gratings are purely chromatic instead of achromatic (Song, Rowland, McPeek, & Wade, 2011). This is because attention to chromatically defined orientation or contrast is mediated by more distinct neural populations in blob and inter-blob regions, respectively, while the overlap between neural populations coding for contrast and orientation of achromatic stimuli is relatively larger. This leads to differences between the chromatic and achromatic conditions in the discriminability of fMRI response patterns in V1.

The distinction between the processing of chromatic versus spatial information in blob and inter-blob regions, however, is not as clear-cut as the initial experiments make it appear. For instance, Leventhal, Thompson, Liu, Zhou, and Ault (1995) did not find evidence for a difference between neural spatial and chromatic tuning in CO blobs and inter-blobs.

Yet even when disregarding the putative roles of neurons in blobs and interblob regions, the relationship between the processing of color and form in general is relevant to the concept of “double opponency”. The term describes a receptive field structure in which the center and surround receive cone opponent inputs but with different signs (e.g., L-M center and -L+M surround, see Figure 4). The preferred input for these cells would be edges of particular chromatic contrasts. This combination of cone opponency with spatial opponency (hence *double* opponency) may lie at the basis of important form-color interactions, possibly reflecting the dependence of color constancy on the 3D structure of the visual input (Maloney, 1999; Radonjic, Cottaris, & Brainard, 2015). Reports of chromatic contrast phenomena document the effect of spatial context on color appearance (e.g., Monnier & Shevell, 2003; Ware & Cowan, 1982).

Friedman, Zhou, and von der Heydt (2003), for instance, did not find any dependence between color tuning and orientation or edge selectivity. To the contrary, they observed that color-selective neurons were often also tuned to specific edge orientations. These properties would make them effective chromatic edge detectors and fit the description of double-opponent cells. These observations are consistent with other reports of double-opponent cells (Conway, Hubel, & Livingstone, 2002; Johnson, Hawken, & Shapley, 2001; Michael, 1978) including the seminal work by Livingstone and Hubel (1984) that suggested the functional distinction between blob and interblob regions in color and form processing.

Other forms of spatio-chromatic integration may involve modulation of neural responses by stimulation occurring outside the classical receptive field of that cell. Stimulation in these locations would not, by themselves, influence neural firing. But in the presence of a stimulus within the classical receptive field they would modulate the neural response to this stimulus. These effects are therefore referred to as “extra-classical receptive field effects”. Wachtler, Sejnowski, and Albright (2003) recorded neural responses to color patches that were presented within a surrounding color context outside the classical RF. A chromatic shift of the context caused a change in color tuning that was consistent with the direction of the context shift. Similar effects had previously been found in lightness perception (MacEvoy & Paradiso, 2001). Compared to double-opponent cells, these forms of spatio-chromatic interactions are expected to occur at a lower spatial scale because they do not, by definition, depend on the local receptive field structure.

In human V1, BOLD activity reflects cone opponency as well (Engel, Zhang, & Wandell, 1997). The BOLD signal is in fact more sensitive along the L-M cone opponent than the L+M luminance direction (see also Liu & Wandell, 2005; Wade, Augath, Logothetis, & Wandell, 2008). In this respect, the cone opponent signals match the perceptual contrast thresholds obtained in psychophysical experiments. The BOLD signal, however, does not parallel perception in other respects as the increased thresholds for L-M and S cone contrasts for higher temporal frequencies is not accompanied by a corresponding change in BOLD responsivity (Liu & Wandell, 2005). Although double opponency has not yet been addressed in humans, neuroimaging evidence suggests that neural populations conjunctively coding for spatial and chromatic stimulus features exist in human V1 as well (Seymour, Clifford, Logothetis, & Bartels, 2009, 2010).

2.3 Area V2

Similar to primary visual cortex, some portions of V2 also stain for cytochrome oxidase (Livingstone & Hubel, 1984). They are arranged in stripes that alternate between thick and thin stripes with pale stripes, which do not react to CO staining, sandwiched between them. Retrograde tracing studies showed that V1 input to thin stripes in V2 originates mainly in blobs whereas thick and pale stripes receive inputs primarily from inter-blob regions (Federer, Williams, Ichida, Merlin, & Angelucci, 2013; Sincich, Jocson, & Horton, 2007, 2010; Xiao & Felleman, 2004).

In agreement with this connectivity pattern, Hubel and Livingstone (1987) found cells in thin stripes to code for color in an opponent way while showing poor tuning for orientation (see also Lu & Roe, 2008). Orientation tuning was commonly associated with thick and pale stripes, which was initially interpreted as the cortical equivalent of the functional segregation seen in parvocellular and magnocellular LGN layers, respectively.

Again, later work called this functional segregation into question as, contrary to the idea of strict functional segregation, orientation-tuning was also found within thin stripes and color selectivity was also observed in thick and pale stripes (Gegenfurtner, Kiper, & Fenstemaker, 1996; Kiper, Fenstemaker, & Gegenfurtner, 1997). Gegenfurtner (2003) points out that although there is good agreement across studies in the tuning patterns observed for color, orientation, direction, and size, it is mainly the conclusions drawn from the data that differ (see Shipp & Zeki, 2002).

Similarly, just as in area V1, Friedman et al. (2003) found color selectivity and responsivity in V2 to be uncorrelated with one another as well. Correlations between chromatic and spatial indices, however, may depend on the exact layer location of a V2 neuron (Shipp, Adams, Moutoussis, & Zeki, 2009). An analysis of superficial and deeper layers (associated with feedback signals) on the one hand and middle layers on the other thus confirmed the lack of correlation for feedback layers but found a negative correlation for middle layers. This was interpreted as a neural mechanism for attention-based feature-binding.

To study the anatomical organization of color responses in V2, Xiao, Wang, and Felleman (2003) recorded optical imaging signals elicited by stimuli with progressively varying hues (e.g., red, orange, yellow, and so on). They found that activity in the thin stripes was topographically organized in hue maps so that perceptually similar hues activated neurons in anatomically adjacent locations.

This is consistent with the observation that, compared to V1 neurons, color tuning was more narrow and covered a broader range of hues (Kiper et al., 1997). Also extraclassical field effects differ between areas V1 and V2. A direct comparison demonstrated that surround suppression in V1 neurons did not affect chromatic tuning in the classical receptive field whereas for V2 neurons chromatic tuning was more consistent between classical RF and surround (Solomon, Peirce, & Lennie, 2004). The authors speculate that their findings in V1 may differ from those obtained in awake monkeys (Wachtler et al., 2003) because they used anesthetized monkeys and there may be more feedback to V1 in the awake state.

The findings regarding color processing in thin stripes may generalize to the human brain as well. Thin stripes can be mapped noninvasively with fMRI in the nonhuman primate (Conway, Moeller, & Tsao, 2007, Figure 2) and their color responsivity has recently been demonstrated in the human brain as well (Nasr, Polimeni, & Tootell, 2016). Chromatic tuning in V2 as measured with human fMRI resembled that in V1, i.e., it also showed a higher sensitivity along the L-M axis than the achromatic axis (Engel et al., 1997). Area V2 may play a particularly prominent role in the perceived binding between color and spatial features (X. Zhang, Qiu, Zhang, Han, & Fang, 2014), which would be consistent with electrophysiological evidence implicating V2 in attention-based feature binding in its feedback layers (Shipp et al., 2009).

The study by Xiao et al. (2003) emphasizes an important criterion for a neural correlate of color perception: similar hues should lead to similar activation patterns. The hue maps were interpreted as an important signature for the neural basis of color perception (Conway, 2003). It is interesting that patient R.M., who suffered from selective color constancy impairments suffered from a lesion that was anatomically confined to V2 (Rüttiger et al., 1999). Although the selectivity of the functional deficit was well established in this study it found that these deficits were associated with lesions in various different brain regions across the patient sample.

When measuring responses to color stimuli that were evenly sampled from a perceptually uniform chromaticity plane, the response patterns in V2 (and also V1) did not reflect this perceptual similarity in contrast to hV4 and VO1 (Brouwer & Heeger, 2009). However, the selectivity of neural populations for hues at intermediate angles with respect to the cardinal axes can also be measured in the human brain where it is shown to be at least as pronounced for earlier areas V1, V2, and V3 as in V4 (Kuriki, Sun, Ueno, Tanaka, & Cheng, 2015). Adaptation effects to hue angles bisecting the cardinal axes did not affect the responses to stimuli located on the adjacent cardinal axes but were restricted to the hue of the adapting stimulus only. This would not be expected if the voxel activity reflected the spatial sum of neural responses tuned to the cardinal axes and it demonstrates that neurons in the earliest human visual are tuned to the wide range of perceptually relevant hues.

2.4 Area V4

Area V4 has played a distinctly prominent role in the neuroscientific study of color vision to the extent that some researchers have referred to this area as the “color center” of the brain (Zeki, 1990) because lesions in this area are often accompanied by an inability to perceive color (see also Bouvier & Engel, 2005). In a series of early studies (Zeki, 1980, 1983) the response properties of cells in V4 and the primary visual cortex were characterized with regard to color constancy, which Zeki deemed a hallmark of color perception. V1 neurons were tuned to wavelengths and could be made to fire in response to a surface of any color as long as it reflected light that matched its spectral tuning. V4 neurons on the other hand were triggered by the surface color and featured invariance to the spectral power distribution of the reflected light. These neurons thus tracked rather the perceived color of the surface and not the spectral properties of the light. They exist in V4 and can also be found anterior to it in the (posterior part of the) temporo-occipital area (TEO), which Zeki

(1996) called V4α. Together with V4, it constitutes the "V4-complex". The neural mechanism underlying the reported color constancy effects may consist in extraclassical receptive field effects (Kusunoki, Moutoussis, & Zeki, 2006; Schein & Desimone, 1990). Chromatic stimulation in the surround of a V4 neuron shifts the peak in the wavelength tuning curve away from the chromaticity of the surround, thereby making the neuron respond to color contrast.

Together with the different anatomical connectivity patterns between V2 and areas V4 on the one hand and V5/MT on the other (DeYoe & Van Essen, 1985; Shipp & Zeki, 1985), the spectral properties of V4 neurons suggested the continuation of segregation between the processing of color and spatial features. The organization of color responses within V4 follows a degree of specialization that is akin to CO-active V1 blobs and V2 thin stripes: color-responsive cells are clustered together in "globs" in V4 and regions anterior to it (Conway et al., 2007; Conway & Tsao, 2009). Glob cells showed hue preferences to specific hues with neighboring neurons coding for similar hues and they are spatially less tuned than "inter-glob" neurons. V4 color responses seem to be arranged in a hue map that is separate from an angle map for orientation tuning (Li, Liu, Juusola, & Tang, 2014; Tanigawa, Lu, & Roe, 2010).

It is, however, clear that V4 is at least as involved in the analysis of object form (Bushnell & Pasupathy, 2012; Desimone & Schein, 1987; Pasupathy & Connor, 2002) and V4 lesions are accompanied by color as well as form vision deficits (Bouvier & Engel, 2005).

Although area V4 in nonhuman primates is a useful model for human area V4, the relationship between the two is complicated by anatomical differences. While V4 has a ventral and a dorsal component in nonhuman primates, only a ventral area exists in the human brain, which is commonly referred to as hV4 and which represents the entire visual hemifield (Goddard, Mannion, McDonald, Solomon, & Clifford, 2011; Wade et al., 2008; Wade, Brewer, Rieger, & Wandell, 2002; Winawer, Horiguchi, Sayres, Amano, & Wandell, 2010), although other models of retinotopic organization exist (K. A. Hansen, Kay, & Gallant, 2007). Color-responsivity is also found in regions anterior to area hV4 in the VO complex (Brewer, Liu, Wade, & Wandell, 2005), which may correspond to the two compartments V4 and V4α, as described by Zeki (1996). Accordingly, color scenes do elicit stronger BOLD signal in the ventral stream than luminance-matched achromatic stimuli in two clusters along the ventral pathway (Bartels & Zeki, 2000; Beauchamp, Haxby, Jennings, & DeYoe, 1999; McKeefry & Zeki, 1997). These have been thought to correspond to V4 and V4α, respectively, although their relationship to the retinotopically based distinction between hV4 and the VO complex is not yet clear.

Interestingly, V4 also responds to color vision when there actually is no chromatic stimulation. This is the case when subjects engage in color imagery in order to make color judgments about objects (Howard et al., 1998; Rich et al., 2006). As another example, in people with synesthesia specific spoken or visually presented numbers, letters, or words that are semantically unrelated to color (e.g., weekdays) can nevertheless induce the sensation of color. These synesthetic experiences are accompanied by increased activation in V4 (Hubbard & Ramachandran, 2005; Nunn et al., 2002) although this has not been observed consistently (Gould van Praag, Garfinkel, Ward, Bor, & Seth, 2015; Rich et al., 2006). Further evidence comes from research on a rare condition called "Charles Bonnet syndrome" (CBS). CBS patients suffer from (partial) blindness due to damage to their sensory input pathways but still experience strong (often colorful) visual hallucinations, which has been interpreted as evidence for the idea that the brain implements a generative model (Reichert, Seriès, & Storkey, 2013) V4 activation in CBS is found be correlated with the color content of the hallucinations (ffytche et al., 1998).

The perceptual relevance of V4 and also VO1 in humans is further underscored by the fact that, although also BOLD signal in earlier areas encode colors, only responses in these higher regions preserve the perceptual distances between them (Brouwer & Heeger, 2009) and reflect color categories derived behaviorally in a color naming task (Brouwer & Heeger, 2013). Although fMRI data is only correlational, there is evidence that V4 is causally involved in color perception. Electrical stimulation of the anterior ventral pathway (V4α) in a patient with an implanted electrode, for instance, causes him to experience a blue-purple color (Murphey, Yoshor, & Beauchamp, 2008).

Since color constancy is about the robust estimation of a surface property, i.e., reflectance, it is worth to consider the role of V4 in surface perception in a more general sense. The usage of abstract color stimuli ignores that color is not a unitary entity but instead must be associated with one of several perceptual primitives such as illumination, surfaces, light-transmitting substances, etc. (Mausfeld, 2003). In this light, and in keeping with its responsiveness to form, human V4 is activated by surface perception (Bouvier, Cardinal, & Engel, 2008; Mendola, Dale, Fischl, Liu, & Tootell, 1999). It is therefore interesting that V4, in contrast to earlier areas, tends to interpret color as a surface property and not just in terms of a distribution of chromatic contrast across the retina (Seymour, Williams, & Rich, 2015). The assignment of color to surfaces (i.e., a spatial object) may thus in fact rely on the conjunctive coding of color and spatial features (Seymour et al., 2009, 2010). The role of V4 in processing of form (and other spatial features), which had previously caused researchers to dismiss the conception of V4 as a color center (e.g., Shapley & Hawken, 2011, p. 713–714), may in this light actually be regarded as a prerequisite for surface color perception.

2.5 Color Processing in Other Extrastriate Brain Regions

As already mentioned, color-sensitive brain regions are found beyond area V4 (see also Komatsu, Ideura, Kaji, & Yamane, 1992; Tootell, Nelissen, Vanduffel, & Orban, 2004) and even anterior to the VO complex extending near the anterior temporal lobe (Lafer-Sousa & Conway, 2013; Lafer-Sousa, Conway, & Kanwisher, 2016). Single-cell recordings in monkeys have demonstrated that color responses in IT cortex were enhanced in a category-dependent way when monkeys performed a color categorization rather than a discrimination task (Koida & Komatsu, 2007) and may form the basis for the perceptually relevant "unique hues" (Stoughton & Conway, 2008). Unique hues are perceptually special in that their appearance is so pure that it cannot be described more accurately with reference to another color, i.e., "a greenish blue". Effects of categorical color perception in the human brain on the other hand have so far been found in the middle frontal gyrus while visual cortex encoded color in terms of their physical properties (Bird, Berens, Horner, & Franklin, 2014). The search for categorical color representations in the brain offers another approach to the study of the neural basis of color perception as humans experience color categorically as well (Bornstein & Korda, 1984).

3 RESEARCH AIMS

Most of the research reviewed in the previous section focused on the relationship between the spectral properties of the light impinging on the retina and the neural signals they elicit in response. This is surprising given that the wavelength composition of the incoming light bears no simple relation to color experience in real life. Only in artificial viewing conditions that abstract from such natural viewing conditions, specifically, when color stimuli are shown without any surrounding context (Zeki (1983) refers to this condition as "void mode"), color appearance is determined by the spectral power distribution of the light. Under natural conditions, however, color experience mostly is not directly linked to the wavelength spectrum of light and may in fact occur even without any stimulation whatsoever such as in dreams, mental simulation, and so on. Rather, the goal of color vision is to infer (possibly using prior object knowledge) a constant representation of a distal stimulus property, surface reflectance. This inference is based on proximal sensory input that is highly confounded by the spectral lighting conditions.

It is therefore important to investigate how the neural color processing architecture, which has mostly been characterized with simple chromatic stimuli, relates to color vision phenomena that most strikingly exemplify the complex connection between the proximal and distal stimuli in color vision. The three research projects described here address exactly this question.

Another important shortcoming of many previous human neuroimaging and electrophysiology experiments was that they often only tested if a brain region showed stronger overall activity in response to color (but see, e.g., Brouwer & Heeger, 2009; Seymour et al., 2009, for exceptions). This approach therefore remains mute regarding the question what information about the particular color being seen was actually encoded in this activity. Multi-voxel pattern analysis is a more sensitive approach to fMRI data analysis that can overcome this limitation (Haxby, Connolly, & Guntupalli, 2014; Haynes & Rees, 2006; Norman, Polyn, Detre, & Haxby, 2006). In each of the three experiments described here, multivariate pattern classification algorithms were used to examine neural activity for its information content regarding color.

Experiment I took up the long-standing question of the influence of "memory colors" on vision. Ewald Hering (1920) introduced the term to describe the influence of prior knowledge on color vision. He was aware that the incoming light signals do not completely specify our percepts. Instead observers' prior experience automatically biases the way they see an object towards its "real" color. There is psychophysical evidence for the influence of memory colors on the perception object colors (T. Hansen, Olkkonen, Walter, & Gegenfurtner, 2006; Olkkonen, Hansen, & Gegenfurtner, 2008; Witzel, Valkova, Hansen, & Gegenfurtner, 2011) but how they affect color processing in the brain remained unknown. The first experiment therefore investigated if and where in the brain there is a common representation for memory color and their corresponding abstract colors.

Experiment II addressed the relationship between illumination and surface colors. The stimuli used in this study consisted of simulated 3D scenes that were complex enough to approximate natural viewing conditions but at the same time were abstract enough so as to exclude the influence of object information. Surface reflectance and the spectral properties of the light sources were systematically varied to identify neural responses in viewers that represented surface color in a way that was robust against illumination changes and how they relate to color constancy as measured psychophysically. Another aim of the study was to test how brain regions differ in terms of their tendencies to interpret incoming light as a signal related to surface color or illumination.

Experiment III deals with the representation of object colors in visual imagery. Since there is no wavelength information in imagery pertaining to the object that is mentally visualized, this study focuses on how colors viewed in void mode relate to colors that are only experienced very subjectively, i.e., in one's imagination. It complements the first experiment because object imagery is a form of top-down influence based on voluntary and controlled effort distinguishing it from memory color which, according to Hering, is thought to operate automatically. The purpose of the third experiment hence was to localize in the brain the common neural substrate for perceiving wavelength-defined colors and color as a feature in the subjective experience of object imagery.

4 Experiment I: Decoding the Yellow of a Gray Banana

Author contributions: AB and MMB designed research, MMB collected and analyzed data, AB and MMB interpreted results and wrote the manuscript.

Current Biology 23, 2268–2272, November 18, 2013 ©2013 Elsevier Ltd All rights reserved http://dx.doi.org/10.1016/j.cub.2013.09.016

Report

Decoding the Yellow of a Gray Banana

Michael M. Bannert[1,2,*] **and Andreas Bartels**[1,2,*]
[1]Vision and Cognition Lab, Werner Reichardt Centre for Integrative Neuroscience, University of Tübingen, 72076 Tübingen, Germany
[2]Bernstein Center for Computational Neuroscience, 72076 Tübingen, Germany

Summary

Some everyday objects are associated with a particular color, such as bananas, which are typically yellow. Behavioral studies show that perception of these so-called color-diagnostic objects is influenced by our knowledge of their typical color, referred to as memory color [1, 2]. However, neural representations of memory colors are unknown. Here we investigated whether memory color can be decoded from visual cortex activity when color-diagnostic objects are viewed as grayscale images. We trained linear classifiers to distinguish patterns of fMRI responses to four different hues. We found that activity in V1 allowed predicting the memory color of color-diagnostic objects presented in grayscale in naive participants performing a motion task. The results imply that higher areas feed back memory-color signals to V1. When classifiers were trained on neural responses to some exemplars of color-diagnostic objects and tested on others, areas V4 and LOC also predicted memory colors. Representational similarity analysis showed that memory-color representations in V1 were correlated specifically with patterns in V4 but not LOC. Our findings suggest that prior knowledge is projected from midlevel visual regions onto primary visual cortex, consistent with predictive coding theory [3].

Results

Hering [4] postulated that memory color exerts a significant influence on the perception of object color. This prediction is supported by recent psychophysical studies showing that the color appearance of color-diagnostic objects is biased toward their corresponding typical colors even when they are presented achromatically [1, 2]. However, the implementation of these cognitive influences in the neural architecture of color processing has remained unknown.

We hypothesized that somewhere in the color-processing pathway of the visual system, bottom-up signals representing sensory chromatic input share a common neural representation with top-down color signals based on object knowledge. We used human functional magnetic resonance imaging (fMRI) in combination with pattern classification to test our hypothesis.

In the first four runs of our fMRI experiment, 18 naive participants (eight females) with normal color vision viewed achromatic images of eight color-diagnostic objects (Figure 1) representing four different memory colors (red, green, blue, and yellow, with two objects per category). Each object was presented in a separate miniblock. We used grayscale photos of real objects (instead of line drawings, for instance) because previous psychophysical research had indicated that the impact of object knowledge on color appearance depends critically on the stimuli appearing natural [2]. In the last six runs, participants were shown real chromatic ring-shaped stimuli from four different hue categories (red, green, blue, and yellow), each at two luminance levels (in separate miniblocks) to maximize subsequent classifier generalization (Figure 1). We asked participants to perform a motion discrimination task at all times in order to ensure naivety with regard to the purpose of the experiment, to maintain balanced attention across all trials, and to direct attention to an attribute different than color or objects.

Memory-Color Decoding: Searchlight Analysis

We first performed a whole-brain searchlight analysis (4-voxel radius) to find out where in the brain local fMRI patterns of blood oxygen level-dependent (BOLD) responses to real color were also predictive of memory color [5]. Four-way color classifiers were trained on all local activity patterns elicited by the chromatic ring stimuli to distinguish between the four color categories across both luminance levels. These classifiers were then tested on the local fMRI responses to each of the eight object images, with each object image being labeled by its memory color. In this way, we obtained a whole-brain map of decoding accuracies (chance level = 25%) for every participant. These maps indicated where object colors could be predicted based on real-color training. We found the largest significant cluster of informative voxels within visual cortex bilaterally near the calcarine sulcus (Figure 2A) ($t_{17} > 2.57$, $p_{\text{voxel}} < 0.01$, cluster size ≥ 68, $p_{\text{cluster}} < 0.001$). Additional significant clusters were found in the left hemisphere along the occipital and temporal lobes, near the left supramarginal gyrus, as well as the postcentral gyri and the posterior portion of the frontal lobe bilaterally.

Memory-Color Decoding: ROI Analysis

To verify the anatomical location of memory-color encoding, we repeated the above analysis for functionally defined regions of interest (ROIs) of visual areas V1–V3 and V4+ (union of areas hV4 and VO-1; see the Experimental Procedures) and object-responsive lateral occipital cortex (LOC), identified independently in nine of our participants. We used recursive feature elimination (RFE) for voxel selection and permutation tests for statistical inference (see the Experimental Procedures). As shown in Figure 2B, activity patterns in V1 allowed prediction of object colors above chance (34%, one-tailed permutation tests, p = 0.0005, Bonferroni corrected for five ROIs), but not V3 (27%, p = 0.22, uncorrected), V4+ (23%, p = 0.89, uncorrected), or LOC (28%, p = 0.048, uncorrected). A trend toward significant decoding was observed in V2 as well, although it marginally failed to reach significance (30%, p = 0.052, Bonferroni corrected). These results were replicated in the independent group of nine subjects for whom retinotopic mapping data were unavailable and anatomical masks were used instead (Figure 2C): the prediction of memory color based on real-color decoders worked only in V1 (30%, p = 0.012, Bonferroni corrected for three ROIs) and not in V2 (23%, p = 0.80, uncorrected) or fusiform gyrus (25%, p = 0.47, uncorrected).

*Correspondence: michael.bannert@tuebingen.mpg.de (M.M.B.), andreas.bartels@tuebingen.mpg.de (A.B.)

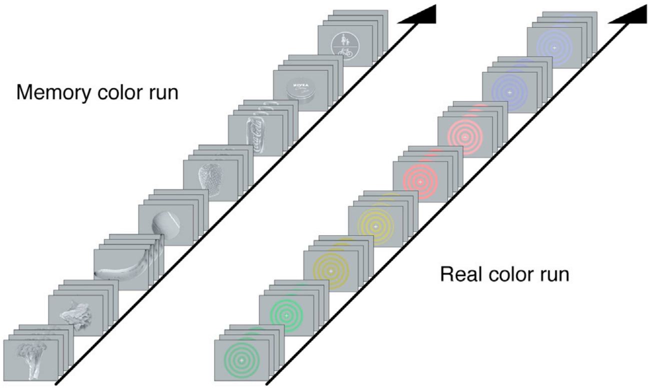

Figure 1. Experimental Design

Left: in the first four runs, participants viewed grayscale images of eight different color-diagnostic objects (broccoli, lettuce, banana, tennis ball, strawberry, coke can, Nivea tin, and blue traffic sign) in miniblocks of four stimuli (stimulus duration = 2 s, interstimulus interval [ISI] = 1 s). Objects rotated by 3°/s, and participants indicated the motion direction (clockwise or counterclockwise). Right: in the last six runs, participants viewed real-color stimuli of four different hues (red, green, blue, and yellow) at two luminance levels. See also Table S1 available online.

Real-Color Decoding: Searchlight and ROI-Based Analyses

As the prediction of memory color was restricted to early visual cortex with no successful decoding in higher regions such as V4+ and LOC, it appeared to be worthwhile to investigate whether this result was specific to memory-color decoding or whether it reflected classification accuracies for real-color stimuli. We therefore trained and tested color decoders on the data from the six "real-color runs" using an n − 1 cross-validation technique, leaving out a different run on every iteration that was then used for testing. A corresponding searchlight analysis showed that the whole occipital cortex, including the fusiform region, encoded real colors (Figure 2D). Accordingly, all ROIs encoded real colors. For functionally localized ROIs, decoding accuracies were as follows: V1 = 40%, V2 = 38%, V3 = 34%, V4+ = 37%, and LOC = 31%. Each ROI achieved p = 0.005 in one-tailed permutation tests, Bonferroni corrected for five ROIs (see Figure 2E). For anatomically defined ROIs, decoding accuracies were as follows: V1 = 39%, V2 = 40%, and fusiform gyrus = 35%. Each ROI achieved p = 0.003 in one-tailed permutation tests, Bonferroni corrected for three ROIs (see Figure 2F). Therefore, as real color could be decoded successfully from every ROI, including V4+, the absence of information predictive of memory color in extrastriate areas cannot be explained simply by the potential poor signal quality that has been shown to be a problem with measurements of the V4+ region in some individuals [6].

Feedback from Extrastriate Visual Areas to V1

The fact that V1 encoded memory colors of objects shown in grayscale strongly suggests that feedback from higher visual areas was involved (see the Discussion). V4+ and LOC are potential candidates for such feedback, as the former is involved in high-level color perception (e.g., [7]) and the latter in shape and object processing (e.g., [8]). We conducted two additional analyses to examine whether the data provide support for both regions being potential candidates as sources for feedback to V1.

First, we tested the possibility that memory color may be represented in V4+ and LOC, yet in a way that differs from the representation of real color. We trained classifiers to discriminate between colors on one half of the objects that represented four memory colors (e.g., strawberry, banana, lettuce, and Nivea tin) and tested them on the remaining half (e.g., coke can, tennis ball, broccoli, and traffic sign). We averaged the results over all 16 possible partitions into training and test set. The assumption was that generalization would only work if the classifier relied on memory color of objects. This analysis showed that memory color could be decoded significantly better than chance in V4+ (37%, p = 0.002, one-tailed permutation tests, Bonferroni corrected for two ROIs) and LOC (30%, p = 0.002) (see Figure 3A).

The alternative account for these results would be that classifiers relied on low-level or shape features that could have been by chance more similar among exemplars of the same memory color. In order to test for this alternative, we performed the same classification analysis using simulated data instead of fMRI data. We used a physiologically plausible computational feed-forward model of object recognition to calculate feature vectors (corresponding to C2 layer responses in HMAX [9]) for our stimuli in a way that mimics the filtering processes thought to be carried out by V4+ and IT circuitry. The classification based on the modeled data (see the Experimental Procedures) was not significant (28%, p = 0.216, one-tailed permutation test, same correction as used for V4+ and LOC). This analysis suggests that shape-related information is unlikely to account for the across-object decoding in V4+ and LOC, which in turn suggests that both regions encode memory color, yet differently than V1. However, as this validation relies on a computational model, it cannot fully rule out the alternative interpretation.

Second, we therefore sought to identify an additional way in which memory-color representations in V1 may be related to activity in V4+ and LOC. We used representational similarity analysis (RSA) [10] to probe whether the representational structure between real colors and memory colors was similar in V1 and in the extrastriate regions. To this end, we calculated the correlation coefficients between every activity pattern related to each of the real colors and to each memory color, yielding one representational dissimilarity matrix (RDM) for each ROI. We then examined which ROIs achieved highest similarity of the obtained matrix with that obtained for V1. We found that the average correlation between RDMs was significant only between V1 and V4+ ($r = 0.53$, $t_8 = 3.79$, $p = 0.005$, one-tailed t test, Bonferroni corrected for two ROIs), and not between V1 and LOC ($r = 0.17$, $t_8 = 1.188$, $p = 0.135$, one-tailed t test, uncorrected), with the former being significantly higher (one-tailed paired t test, $t_8 = 5.37$, $p < 0.001$). These results show that V1 and V4+ resemble each other significantly in terms of the similarity relationships between patterns encoding memory and real colors, respectively.

Discussion

In the present study, we addressed a fundamental question in color vision, namely the effect of prior knowledge on color processing. Our results show that color decoders could predict, from fMRI activity in V1, the true color of eight color-diagnostic objects, representing four different color categories, in the complete absence of chromatic stimulation. The results were found in naive observers carrying out a motion task and therefore appear to be the result of an automatically occurring process during object vision rather than of active imagery. A potential source of the memory-color signal in V1 may be

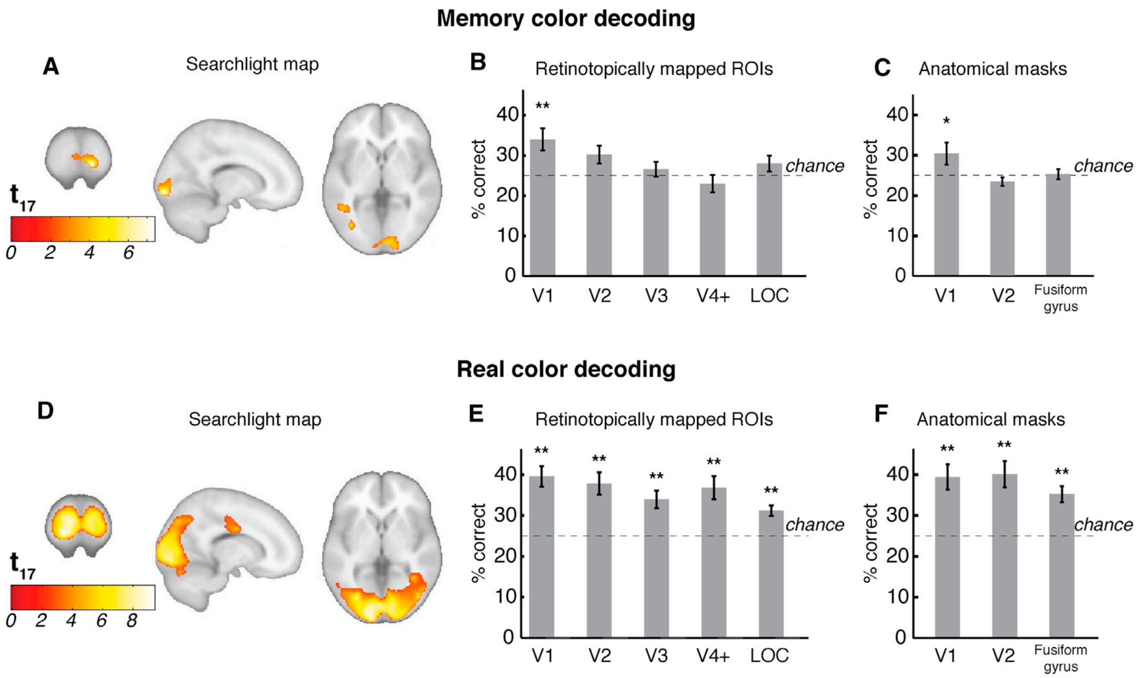

Figure 2. Multivoxel Pattern Analysis Results

(A) Whole-brain searchlight analysis across all 18 subjects. Prediction of the memory colors of the grayscale object images based on training on the real-color runs using local fMRI activity patterns was significantly above chance in early visual cortex. Brain sections are centered on position x = 14, y = −94, z = 0 in MNI space. Searchlight maps are cluster-size-corrected at $p_{voxel} < 0.01$, $p_{cluster} < 0.001$.

(B) Among functionally localized ROIs in nine subjects, prediction of memory color based on real-color training was successful only in area V1.

(C) In anatomically defined ROIs of the remaining nine subjects, prediction of memory color based on real-color training was successful only in area V1.

(D) Searchlight analysis. Prediction of real colors based on real-color training was significantly above chance in the entire occipital cortex. The same conventions as those in (A) are used.

(E) Among functionally localized ROIs, prediction of real colors was significantly above chance in all visual areas and in LOC.

(F) In anatomically defined ROIs, real-color prediction based on real-color training worked in each of the three ROIs.

Bar plots depict mean decoding accuracies. Error bars represent the SEM. *p < 0.05, **p < 0.01 (one-tailed permutation tests, Bonferroni corrected). See the Supplemental Experimental Procedures for further analyses. See also Figure S1.

V4+, as they shared a strong correlation in the structure of memory-color representations. Some authors interpret such similarity as "representational connectivity" between brain regions [10], which in this case fits well with our interpretation of the results that information is projected from higher-level visual regions onto primary visual cortex. The neural substrates revealed in the present findings may underlie several perceptual effects having to do with top-down influences of prior knowledge involving color [1, 2, 11–13]. To our knowledge, the present results are the first to demonstrate that memory color influences neural activity at the earliest levels of cortical processing, in the primary visual cortex.

The results are consistent with numerous experiments showing that, instead of encoding a veridical representation of the physical environment, V1 activity is in fact strongly modulated by top-down feedback, which can be readily detected with fMRI [14]. V1 activity has been shown to represent perceived lightness rather than physical stimulus intensity [15], to represent perceived rather than the physical size of stimuli [16, 17], to encode context-dependent feedback in the visual field [18], and to signal high-level grouping effects of global Gestalt cues [19].

Our own and the discussed results are consistent with predictive coding theory [3]. In the context of the present study, the assumption is that higher visual areas send predictions of expected object colors to V1, where they are compared to bottom-up information. Predictive coding is efficient in that it can enhance weak sensory input through prior knowledge and at the same time boost neural processing of unexpected (as opposed to predictable) aspects of the environment [20]. Thus, the omission of an expected visual stimulus can, for instance, even lead to stronger fMRI responses in V1 than its presence [21].

In this context, the BOLD signal in V1 represented either the mismatch between expected and incoming color signals or the predictive signal fed back to early visual cortex. Interestingly, this prediction-related activity resembled the expected signal driven by real-color input. Similarly, the agreement between representational structures in V1 and V4+ suggest that V4+ may be both receiver and source of color signals in V1 during sensory color stimulation and object viewing, respectively.

Based on previous imaging studies (e.g., [7, 22]), it may seem surprising that significant decoding was observed in V1 only but not in color-sensitive V4+. Several reasons may have contributed to this. Color signaling in V4+ could have been weakened due to vascular artifacts [6] and due to the attentional focus on motion rather than on color [23, 24]. Slotnick [25], for instance, does find upmodulation of V4+ in a memory task when subjects actively remembered that an abstract figure had previously been presented in color in the study phase. However, these reasons cannot fully account for the lack of decoding in V4+ as decoding of real colors was well above chance in V4+. Perhaps one reason lies in

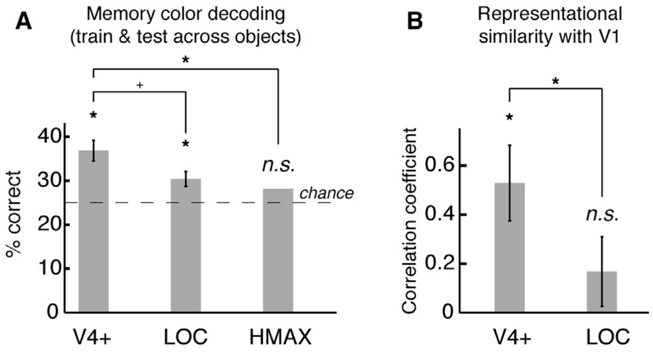

Figure 3. Color Generalization across Objects and Representational Similarity with V1

(A) Classifiers trained on responses to one set of color-diagnostic objects and tested on the other (with no overlap of object identities in the two sets) correctly predicted the memory color of objects in the test set in V4+ and LOC, with an advantage for V4+ [V4+ versus HMAX: $t(8) = 3.69$, $p = 0.009$, Bonferroni corrected for three comparisons; V4+ versus LOC: $t(8) = 2.33$, $p = 0.024$, uncorrected]. In contrast, classifiers failed to predict low-level and shape features between objects of the training and test sets. Features were extracted using the physiologically plausible HMAX algorithm.

(B) Representational similarity with V1 was found in V4+ but not LOC. This shows that the similarity relationships between patterns encoding memory and real colors were significantly correlated between V1 and V4+.

Error bars represent the SEM. *$p < 0.01$ (Bonferroni corrected) +$p < 0.05$ (uncorrected). One-tailed paired t tests (df = 8) were used for pairwise comparisons.

the categorical perception of grayscale rather than colored objects.

Also, classification accuracy may in principle be related to differences in the spatial inhomogeneity of feature-selective neuronal assemblies across voxels [26], which has recently also been suggested in context of color encoding along cardinal color axes [27]. Accordingly, our analyses cannot exclude the possibility that memory color may also be present in high-level regions (as is indeed suggested by our across-object classification).

Our data provide, to our knowledge, the first evidence for encoding of color in V1 in the absence of any chromatic input. The present results therefore add one more dimension, color, to a growing body of literature showing that V1 activity is heavily influenced by feedback from higher-level regions, encoding perceived rather than purely physical stimulus properties even if they are absent from bottom-up input. The present results offer a neural account for previously observed perceptual effects of memory color and provide additional evidence for a role of V1 as convergence zone between bottom-up input and top-down predictive signals. The present findings have implications beyond color vision, as they show how object knowledge can serve as a prior to constrain the inferences the visual system makes at earliest processing stages about the appearance of complex natural scenes.

Experimental Procedures

Participants

Eighteen volunteers (mean age 27.2 years, SD 4.1 years, eight female) with normal color vision, as assessed using Ishihara plates, participated in the main experiment. All provided written informed consent, and the ethics committee of the University Hospital Tübingen approved the experiment. Participants were naive with respect to the purpose of the study. Instead, they were told that its aim was to investigate motion using object and color stimuli. Nine participants (mean age 28.4, SD 5.1 years, four female) took part in a retinotopic mapping experiment.

Behavioral Tasks and Imaging Paradigm

In the first four fMRI runs, participants were required to view slowly rotating grayscale images of objects and to indicate for each stimulus by button press whether rotation occurred in a clockwise (right button) or counterclockwise (left button) direction. The images were isoluminant grayscale photos of eight different color-diagnostic objects, two for each color category: a strawberry and coke can for red, broccoli and lettuce for green, a traffic sign and Nivea tin for blue, and a tennis ball and banana for yellow (see the Supplemental Experimental Procedures for details). Every image was presented for 2 s, and the ISI was 1 s. Object images were presented in miniblocks of four trials containing the same object but with random rotation direction on each presentation (see Figure 1). Each run contained 32 miniblocks. The sequence of objects was pseudorandomized such that every object was preceded equally often by all objects.

In the last six runs, participants viewed chromatic stimuli consisting of abstract color rings similar to those used by Brouwer and Heeger [7] (see the Supplemental Experimental Procedures for details). Each ring was defined by its color (red, green, blue, or yellow) and brightness (high or low: ±10% around the object's luminance), yielding eight stimuli that were presented in separate miniblocks (see Table S1 for chromaticity coordinates). In each trial within a miniblock, rings randomly either expanded or contracted. The design was identical to that of the object runs. Participants performed a one-back matching task that amounted to a motion task in the majority of trials since hue and luminance were constant within miniblocks.

fMRI Scan Parameters and Preprocessing

Data were collected on a 3T fMRI system with a resolution of 3 mm isotropic voxel size across 33 slices and preprocessed with SPM5 (http://www.fil.ion.ucl.ac.uk/spm/). Neural responses to objects and color rings were estimated with a separate general linear model for each run and with separate boxcar regressors for each miniblock (see the Supplemental Experimental Procedures for details).

Retinotopic Mapping and Anatomical Masks

In nine subjects, polar angle maps were obtained using standard methods. Areas hV4 and VO-1 are reported as joint ROI "V4+" since segregated analyses yielded same results as for the joint ROI. To confirm memory-color decoding in V1 for the remaining nine participants, we used the automatic cortical parcellation provided by Freesurfer to obtain ROI masks of V1, V2, and fusiform gyrus (see the Supplemental Experimental Procedures and Figure S1).

Multivoxel Pattern Analysis

We analyzed our data with in-house Matlab code based on the Princeton multivoxel pattern analysis toolbox (http://www.pni.princeton.edu/mvpa/). For all analyses, we applied linear discriminant analysis for pattern classification using shrinkage estimation to make sure that covariance matrices were nonsingular [28, 29]. We obtained a whole-brain map of decoding accuracies for every participant. After smoothing with a 6 mm Gaussian kernel, group statistics were calculated using one-sample t tests. Results were corrected for multiple comparisons using a cluster size threshold determined on the basis of Monte-Carlo simulations [30]. For ROI-based decoding of memory color based on real-color training, we first used a feature selection algorithm (RFE) to identify those voxels that contributed most strongly to the discrimination of real colors (see the Supplemental Experimental Procedures). We used the more accurate approach of permutation tests for statistical inference in the ROI-based classification analyses [28]. This involved permuting the labels of the training data repeatedly to bootstrap a null distribution for every statistic (see the Supplemental Experimental Procedures).

Reported p values represent the fraction of permutations yielding classification accuracies that were at least as high as the observed one.

Supplemental Information

Supplemental Information includes Supplemental Experimental Procedures, one figure, and one table and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2013.09.016.

Acknowledgments

This work was funded by grants from the German Federal Ministry of Education and Research (grant number FKZ 01GQ1002), the Deutsche Forschungsgemeinschaft (DFG grant number BA4914/1-1 to A.B.), and the Centre for Integrative Neuroscience (DFG grant number EXC 307), University of Tübingen.

Received: June 28, 2013
Revised: August 15, 2013
Accepted: September 6, 2013
Published: October 31, 2013

References

1. Hansen, T., Olkkonen, M., Walter, S., and Gegenfurtner, K.R. (2006). Memory modulates color appearance. Nat. Neurosci. *9*, 1367–1368.
2. Olkkonen, M., Hansen, T., and Gegenfurtner, K.R. (2008). Color appearance of familiar objects: effects of object shape, texture, and illumination changes. J. Vis. *8*, 13.1–13.16.
3. Rao, R.P., and Ballard, D.H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. *2*, 79–87.
4. Hering, E. (1920). Grundzüge der Lehre vom Lichtsinn (Berlin: Springer).
5. Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. Proc. Natl. Acad. Sci. USA *103*, 3863–3868.
6. Winawer, J., Horiguchi, H., Sayres, R.A., Amano, K., and Wandell, B.A. (2010). Mapping hV4 and ventral occipital cortex: the venous eclipse. J. Vis. *10*, 1–22.
7. Brouwer, G.J., and Heeger, D.J. (2009). Decoding and reconstructing color from responses in human visual cortex. J. Neurosci. *29*, 13992–14003.
8. Malach, R., Reppas, J.B., Benson, R.R., Kwong, K.K., Jiang, H., Kennedy, W.A., Ledden, P.J., Brady, T.J., Rosen, B.R., and Tootell, R.B. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. Proc. Natl. Acad. Sci. USA *92*, 8135–8139.
9. Serre, T., Oliva, A., and Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. Proc. Natl. Acad. Sci. USA *104*, 6424–6429.
10. Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P.A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron *60*, 1126–1141.
11. Naor-Raz, G., Tarr, M.J., and Kersten, D. (2003). Is color an intrinsic property of object representation? Perception *32*, 667–680.
12. Tanaka, J.W., and Presnell, L.M. (1999). Color diagnosticity in object recognition. Percept. Psychophys. *61*, 1140–1153.
13. Mitterer, H., and de Ruiter, J.P. (2008). Recalibrating color categories using world knowledge. Psychol. Sci. *19*, 629–634.
14. Logothetis, N.K. (2008). What we can do and what we cannot do with fMRI. Nature *453*, 869–878.
15. Boyaci, H., Fang, F., Murray, S.O., and Kersten, D. (2007). Responses to lightness variations in early human visual cortex. Curr. Biol. *17*, 989–993.
16. Murray, S.O., Boyaci, H., and Kersten, D. (2006). The representation of perceived angular size in human primary visual cortex. Nat. Neurosci. *9*, 429–434.
17. Sperandio, I., Chouinard, P.A., and Goodale, M.A. (2012). Retinotopic activity in V1 reflects the perceived and not the retinal size of an afterimage. Nat. Neurosci. *15*, 540–542.
18. Smith, F.W., and Muckli, L. (2010). Nonstimulated early visual areas carry information about surrounding context. Proc. Natl. Acad. Sci. USA *107*, 20099–20103.
19. Zaretskaya, N., Anstis, S., and Bartels, A. (2013). Parietal cortex mediates conscious perception of illusory gestalt. J. Neurosci. *33*, 523–531.
20. Summerfield, C., and Egner, T. (2009). Expectation (and attention) in visual cognition. Trends Cogn. Sci. *13*, 403–409.
21. den Ouden, H.E.M., Friston, K.J., Daw, N.D., McIntosh, A.R., and Stephan, K.E. (2009). A dual role for prediction error in associative learning. Cereb. Cortex *19*, 1175–1185.
22. Brewer, A.A., Liu, J., Wade, A.R., and Wandell, B.A. (2005). Visual field maps and stimulus selectivity in human ventral occipital cortex. Nat. Neurosci. *8*, 1102–1109.
23. Bartels, A., and Zeki, S. (2000). The architecture of the colour centre in the human visual brain: new results and a review. Eur. J. Neurosci. *12*, 172–193.
24. Kastner, S., and Ungerleider, L.G. (2000). Mechanisms of visual attention in the human cortex. Annu. Rev. Neurosci. *23*, 315–341.
25. Slotnick, S.D. (2009). Memory for color reactivates color processing region. Neuroreport *20*, 1568–1571.
26. Bartels, A., Logothetis, N.K., and Moutoussis, K. (2008). fMRI and its interpretations: an illustration on directional selectivity in area V5/MT. Trends Neurosci. *31*, 444–453.
27. Parkes, L.M., Marsman, J.B.C., Oxley, D.C., Goulermas, J.Y., and Wuerger, S.M. (2009). Multivoxel fMRI analysis of color tuning in human primary visual cortex. J. Vis. *9*, 1–13.
28. Pereira, F., and Botvinick, M. (2011). Information mapping with pattern classifiers: a comparative study. Neuroimage *56*, 476–496.
29. Schäfer, J., and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat. Appl. Genet. Mol. Biol. *4*, Article32.
30. Slotnick, S.D., Moo, L.R., Segal, J.B., and Hart, J., Jr. (2003). Distinct prefrontal cortex activity associated with item memory and source memory for visual shapes. Brain Res. Cogn. Brain Res. *17*, 75–82.

Current Biology, Volume *23*

Supplemental Information

Decoding the Yellow of a Gray Banana

Michael M. Bannert and Andreas Bartels

Figure S1

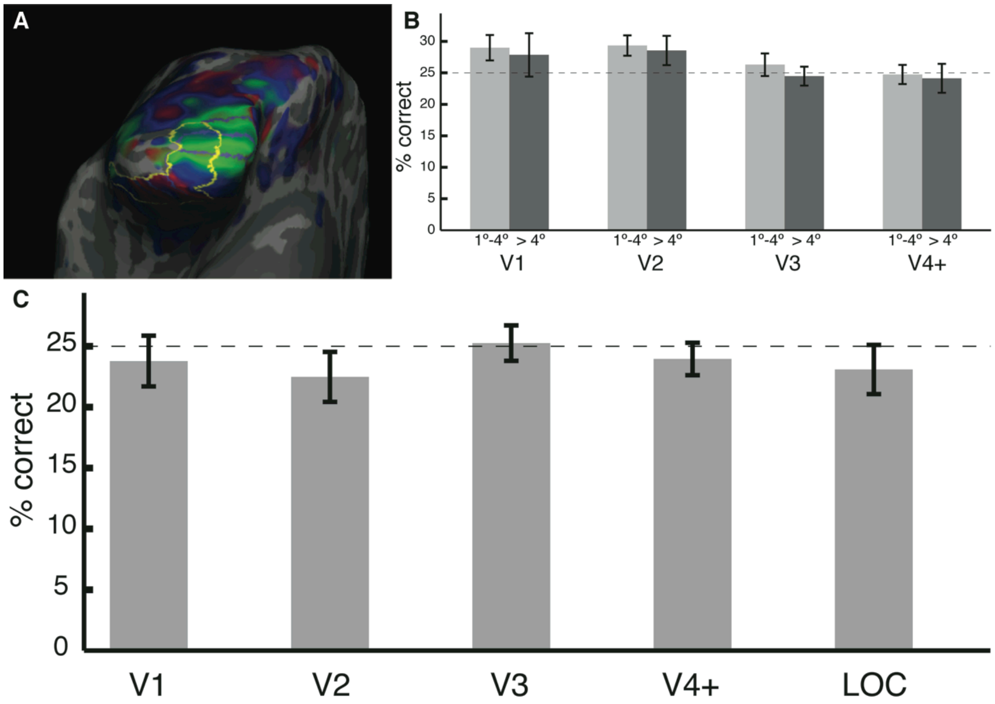

Figure S1 (related to main figure 2): Definition of retinotopic ROIs and classification accuracies at different eccentricities. (A) Polarity map of the left hemisphere of a representative participant illustrating the definition of retinotopic ROIs. Vertices significantly modulated at stimulus frequency are shown in color (ventral visual field: blue to green). Retinotopic subdivisions were only demarcated for voxels whose phase reversals were clearly visible (i.e. significantly modulated at the frequency of the mapping stimulus at p<0.01, uncorrected) and that could be assigned unambiguously to one of the visual areas. The foveal confluence was hence not included (see gray patch to the left of the yellow outline). Yellow outline: cortical surface representing eccentricities between 1 and 4 degrees. **(B)** Decoding accuracies of memory color based on real-color training per ROI at high (> 4 deg) and low (1-4 deg) eccentricities. There was no significant decoding advantage for either eccentricity. **(C)** Decoding accuracies based on mean activation levels**.** Shown are decoding accuracies of memory color based on real-color training when only mean signal per trial instead of voxel patterns were submitted to the classifiers for both training (on real color responses) and testing (memory color responses). Decoding did not work using mean signal only.

Table S1

	x	y
Red (high luminance)	0.3589	0.3291
Green (high luminance)	0.3087	0.4132
Blue (high luminance)	0.2825	0.2791
Yellow (high luminance)	0.3726	0.4281
Red (low luminance)	0.388	0.3292
Green (low luminance)	0.3099	0.388
Blue (low luminance)	0.2676	0.2544
Yellow (low luminance)	0.3744	0.4311

Table S1 (related to main figure 1). CIE (x,y) chromaticity coordinates of the eight color stimuli.

Supplemental Experimental Procedures

Object and Color stimulus details

Object stimuli:

Object images were gray-scale photos of eight different color-diagnostic objects, two for each color category: strawberry and coke can for red, broccoli and lettuce for green, traffic sign and Nivea tin for blue, tennis ball and banana for yellow. All stimuli had the same mean luminance (180 cd/m^2), same contrast (root-mean-square (RMS) of 12.95 % of image mean, standard deviation: 0.46 %), and same object size in terms of pixels subtended (number of pixels between 88746 and 89100). The average radius of the smallest circle enclosing each stimulus was 6.31 degrees (SD = 1.90 degrees).

At the beginning of every run, a dummy block was presented that corresponded to the last block of the previous run to ensure history-matching. Dummy blocks were excluded from analysis.

Color stimuli:

Color stimuli consisted of concentric expanding or contracting rings in red, green, blue, and yellow hues, each presented at high or low luminance, resulting in 8 stimuli that were presented in separate mini-blocks. Rings expanded or contracted at 1 deg/s, reaching a maximal radius of 7.19 deg. All stimuli were presented on an achromatic background of 180 cd/m^2 luminance. The transparency of the rings was modulated sinusoidally over space (0.5 cycles per visual degree). The minimum flicker technique [S1] was applied for each participant individually prior to scanning in order to ensure that all hues were equiluminant at high and low luminance, respectively. To this end, each hue was presented against two achromatic backgrounds with different luminance values (± 10% around the object's luminance: low luminance: 162 cd/m^2; high luminance: 198 cd/m^2). The CIE coordinates of the calibrated stimuli are shown in **Supplemental Table S1**.

fMRI object localizer and retinotopic mapping

LOC localizer:

Participants passively viewed images of objects and their phase-scrambled counterparts in a block design: 16 images of the same category were shown for 400 ms each (ISI = 600 ms) on an isoluminant gray background during 16 s. Each condition was repeated 9 times. The task was to attend to the stimuli presented and to internally name them if they were recognizable. Localizer data were spatially filtered with a 6 mm full width at half maximum (FWHM) Gaussian kernel. Lateral occipital cortex (LOC) was defined as the set of voxels in lateral occipital cortex responding more strongly to objects than to their scrambled counterpart.

Retinotopic mapping:

Polar angle maps were obtained by asking participants to fixate the center of the screen while performing a demanding attention task that was displayed on top of a slowly rotating wedge through which a flickering checkerboard stimulus was displayed. Attention is known to increase the SNR of the BOLD signal in this setting [S2]. Our attention task has previously been described in detail elsewhere [S3]. The angle of the wedge was 30º and it extended to the edge of the screen. For eccentricity mapping the checkerboard stimulus was masked by an annulus that was either contracting or expanding. The period of both stimuli was 49.92 s and there were 10 cycles per run. The contrast of the checkerboard was reversed at a frequency of 6 Hz. Check sizes were scaled logarithmically to take into account cortical magnification. Retinotopic mapping data were preprocessed and analyzed using the Freesurfer package [S4] available at http://surfer.nmr.mgh.harvard.edu/. Visual areas V1-V3, as well as hV4 and VO-1 were demarcated based on horizontal and vertical meridian boundaries [S5, S6]. Areas hV4 and VO-1 are reported as joint ROI 'V4+', since segregated analyses yielded same results as for the joint ROI.

Automatic anatomical parcellation:

To confirm memory color decoding in V1 for the remaining nine participants, we used the automatic cortical parcellation provided by Freesurfer to obtain ROI masks of V1, V2 [S7], and fusiform gyrus [S8] for every participant for whom no retinotopic maps were available. The analyses performed with these masks were identical to the ones carried out with the retinotopically mapped ROIs (see below).

fMRI details and preprocessing

fMRI details:

Functional data were acquired on a Siemens 3T Trio system using a 12-channel phased-array head coil (Siemens, Erlangen, Germany) with a T2*-weighted gradient-echo echoplanar (EPI) sequence at a spatial resolution of 3 mm isotropic voxel size (64 x 64 acquisition matrix, 33 slices) with TR = 2.3 s, TE = 35 ms, 79º flip angle. High-resolution structural scans (1 mm isotropic voxel size) were acquired using a T1-weighted MP-RAGE sequence. Functional runs had 179, 195, and 178 volumes for the object, localizer, and real color runs, respectively. The first 3 TRs from each run were discarded to allow for T1 equilibrium effects. Retinotopic mapping was performed at a spatial resolution of 2 mm isotropic voxel size with TR = 3.12 s, TE = 39 ms.

Preprocessing for classification:

Data were preprocessed with SPM5 (http://www.fil.ion.ucl.ac.uk/spm/), including slice-time-correction, spatial realignment, and unwarping. For classification, data were left unsmoothed in subjects' native spaces. The results were later normalized to MNI space for group analysis.

Neural responses to objects and color rings were estimated with a GLM for each run separately, with separate boxcar-regressors for each mini-block. Each mini-block was modeled with a boxcar regressor that was shifted by 5 s forward in time to take into account the hemodynamic lag. The resulting time series of beta estimates for each voxel were corrected for temporal drifts by subtracting the fit of a second-order polynomial from the original time course of beta estimates. Time-courses were then z-transformed. Outlier beta estimates with absolute values higher than 2 were re-set to 2 or -2, respectively. We repeated all ROI-based analyses with and without removal (reported in main figures) of the mean of every pattern. Results were virtually not affected, and all reported statistical inferences apply for both analyses equally. See also **Figure S1** for absence of mean-based decoding of memory color.

Recursive feature elimination (RFE) and permutation tests

Recursive feature elimination (RFE) applied in ROI analyses:

For decoding of memory color based on real color training, we first used a feature selection algorithm (RFE) to identify those voxels that contributed most strongly to the discrimination of real colors. We performed RFE in combination with linear discriminant analysis: In the first iteration, LDA classifiers were trained and tested including all voxels in the ROI in an n-1 cross-validation using only data from the six real color runs as detailed above. Classification accuracy was calculated and 15 % of the voxels with the lowest average absolute weights were removed from the feature set. Using only the surviving voxels in the next iteration, new classifiers were again trained and tested and the least discriminative voxels were selected out. There were 15 iterations in total, thus yielding 15 different voxel sets with 15 corresponding decoding accuracies. The feature set with the highest decoding accuracy on the real colors was then used for the prediction of memory color. It is important to note that the selection of the optimal feature set depended only on data from the real color runs and that the memory color runs were kept completely separate. Data from the object runs were used as test set only to evaluate generalization performance of the color decoder.

Statistical inference using permutation tests:

We used the more exact approach of permutation tests for statistical inference of the ROI-based classification analyses [S9] to ensure that statistical inference was based on a more accurate estimate of the true distribution of decoding accuracies under the null hypothesis of no hue information in the patterns. Since category labels are interchangeable under the null hypothesis, we repeated the classification using 10^4 (memory color decoding) or 10^3 (all other classifications) random permutations of the training labels (one of which was the true one) to bootstrap a null distribution for every statistic.

Reported p-values represent the fraction of permutations yielding classification accuracies that were at least as high as the observed one.

Classification of HMAX filter responses

In order to rule out the possibility that the classification of memory color across object identities was driven by low-level stimulus properties that happened to be shared by objects within a given color category, we performed a classification on simulated data obtained from a computational feed-forward model of object recognition (HMAX, available at http://maxlab.neuro.georgetown.edu/hmax/). We calculated the C2 layer filter responses to the object images, which are thought to approximate neural activity patterns in areas V4+ and IT [S10], and conducted the same classification on the filter responses as we did on the actual fMRI activity [S11, S12].

Retinotopic ROI definition and classification accuracies at different eccentricities

At low eccentricities, the demarcation of visual areas based on phase reversals becomes less precise. To minimize error in the definition of ROIs, we followed a rather conservative analysis strategy and left out all cortical surface vertices at low eccentricities where phase reversals were not clearly visible and that could not be assigned unambiguously to any of the visual areas. Specifically, we included only those vertices in our retinotopic mapping analysis that were significantly modulated at the frequency of the polarity mapping stimulus (at $p < 0.01$, uncorrected). These vertices had clear phase-assignments and could be unambiguously attributed to one of the visual areas (see **Figure S1A**).

To test whether decoding accuracies for memory color were different at high and low eccentricities within visually stimulated cortex, we separated those voxels that were included by our feature selection procedure into those that fell within 1 to 4 degrees of eccentricity of the visual field and those that represent the visual field beyond 4 degrees. We chose 4 degrees because at this particular eccentricity our ROIs were split into two approximately equally large sub-ROIs in most of the hemispheres. In analogy with our original analysis we again trained classifiers to distinguish between real colors for both eccentricities separately and tested them on the responses to the object stimuli. A three-way repeated measures ANOVA with ROI and eccentricity as fixed factors and participants as random factor failed to detect any significant effect of eccentricity $F(1,8) = .312$ ($p = .592$) (see **Supplemental Figure S1B**). We conclude that visual field representations were equally affected by the feedback signal at low and high eccentricities.

Classification based on mean activation levels

We tested to what extent our decoding results could be driven by differences in mean activation levels between conditions. To this end, we removed the pattern structure from every vector by replacing all components of each vector in the training and test sets by its respective mean value. It was not possible to predict the memory color of objects with color decoders trained on responses to real colors (10^4 permutations, $p \geq 0.4603$) (see **Figure S1C**). We can therefore conclude that there was no systematic difference in the overall activity evoked by our stimuli that could explain the results from our memory color classification.

Supplemental References

S1. Kaiser, P.K. (1991). Flicker as a function of wavelength and heterochromatic flicker photometry. In Limits of Vision, J.J. Kulikowski, V. Walsh and I.J. Murray, eds. (Basingstoke (United Kingdom): MacMillan), pp. 171-190.

S2. Bressler, D.W., and Silver, M.A. (2010). Spatial attention improves reliability of fMRI retinotopic mapping signals in occipital and parietal cortex. NeuroImage *53*, 526-533.

S3. Fischer, E., Bülthoff, H.H., Logothetis, N.K., and Bartels, A. (2012). Human areas V3A and V6 compensate for self-induced planar visual motion. Neuron *73*, 1228-1240.

S4. Fischl, B. (2012). FreeSurfer. NeuroImage *62*, 774-781.

S5. Sereno, M.I., Dale, A.M., Reppas, J.B., Kwong, K.K., Belliveau, J.W., Brady, T.J., Rosen, B.R., and Tootell, R.B. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. Science *268*, 889-893.

S6. Brewer, A.A., Liu, J., Wade, A.R., and Wandell, B.A. (2005). Visual field maps and stimulus selectivity in human ventral occipital cortex. Nat. Neurosci. *8*, 1102-1109.

S7. Fischl, B., Rajendran, N., Busa, E., Augustinack, J., Hinds, O., Yeo, B.T.T., Mohlberg, H., Amunts, K., and Zilles, K. (2008). Cortical folding patterns and predicting cytoarchitecture. Cereb. Cortex *18*, 1973-1980.

S8. Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage *31*, 968-980.

S9. Pereira, F., and Botvinick, M. (2011). Information mapping with pattern classifiers: a comparative study. NeuroImage *56*, 476-496.

S10. Serre, T., Oliva, A., and Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. Proc. Natl. Acad. Sci. USA *104*, 6424-6429.

S11. Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P.A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron *60*, 1126-1141.

S12. Connolly, A.C., Guntupalli, J.S., Gors, J., Hanke, M., Halchenko, Y.O., Wu, Y.-C., Abdi, H., and Haxby, J.V. (2012). The representation of biological classes in the human brain. J. Neurosci. *32*, 2608-2618.

5 Experiment II: Invariance of Surface Color Representations Across Illuminant Changes in the Human Cortex

Author contributions: AB and MMB designed research, MMB collected and analyzed data, AB and MMB interpreted results and wrote the manuscript.

Invariance of Surface Color Representations Across Illuminant Changes in the Human Cortex

Michael M. Bannert and Andreas Bartels
Werner Reichardt Centre for Integrative Neuroscience, Eberhard Karls University
Bernstein Center for Computational Neuroscience
Max Planck Institute for Biological Cybernetics
Department of Psychology, Eberhard Karls University
Tübingen

A central problem in color vision is that the light reaching the eye from a given surface can vary dramatically depending on the illumination. Despite this, our color percept, the brain's estimate of surface reflectance, remains remarkably stable. This phenomenon is called color constancy. Here we investigated which human brain regions represent surface color in a way that is invariant with respect to illuminant changes. We used physically realistic rendering methods to display natural yet abstract 3D scenes that were displayed under three distinct illuminants. The scenes embedded, in different conditions, surfaces that differed in their surface color (i.e. in their reflectance property). We used multivariate fMRI pattern analysis to probe neural coding of surface reflectance and illuminant, respectively. While all visual regions encoded surface color when viewed under the same illuminant, we found that only in V1 and V4α surface color representations were invariant to illumination changes. Along the visual hierarchy there was a gradient from V1 to V4α to increasingly encode surface color rather than illumination. Finally, effects of a stimulus manipulation on individual behavioral color constancy indices correlated with neural encoding of the illuminant in hV4. This provides neural evidence for the Equivalent Illuminant Model. Our results provide a principled characterization of color constancy mechanisms across the visual hierarchy, and demonstrate complementary contributions in early and late processing stages.

Keywords: Color constancy, fMRI, V1, V4, surface perception

INTRODUCTION

Color constitutes a fundamental quality of visual experience, and supports a large variety of behavioral tasks (Mollon, 1989). However, the fact that the light reflected from surfaces depends both on the surface color (i.e. its reflectance) and the color of the incident light (Land & McCann, 1971) poses a challenging problem to the visual system. It is therefore impossible to know the reflectance of a surface without any knowledge of the illumination. As numerous psychophysical studies have documented, however, the perception of surface color is fairly robust even in the face of changes in illumination. This property of the visual system is referred to as "color constancy". It is unclear how the human brain transforms the highly ambiguous incoming color signals to create surface color representations that are stable across illuminants. What factors at the neural level are involved in color constancy?

Early investigations demonstrated the involvement of area V4 in color perception in monkeys (Wild, Butler, Carden, & Kulikowski, 1985; Zeki, 1983) and inspired neuroimaging studies suggesting a similar role for human V4 (Bartels & Zeki, 2000; Beauchamp, Haxby, Jennings, & DeYoe, 1999; Lueck et al., 1989). Also the functional organization of color responses in this area was shown to reflect perceptually relevant stimulus dimensions in both non-human (Conway & Tsao, 2009; Kusunoki, Moutoussis, & Zeki, 2006; Li, Liu, Juusola, & Tang, 2014) and human primates (Brouwer & Heeger, 2009, 2013).

Human lesion studies have accordingly implicated a connection between area V4 and achromatopsia (but also form vision deficits) (Bouvier & Engel, 2005). As for selective color constancy impairments, evidence appears less conclusive with some work suggesting a link with V4 (Clarke, Walsh, Schoppig, Assal, & Cowey, 1998; Kennard, Lawden, Morland, & Ruddock, 1995) while different research highlights the involvement of other areas (Rüttiger et al., 1999), including V1 (Kentridge, Heywood, & Weiskrantz, 2007).

Human neuroimaging indeed shows increased responses to color already in early areas V1 and V2 (Bartels & Zeki, 2000; Beauchamp et al., 1999; Engel, Zhang, & Wandell, 1997). But also in non-human primates, the chromatic context modulation of neural color tuning (Wachtler, Sejnowski, & Albright, 2003) and the double opponency of neurons

(Conway & Livingstone, 2006; Johnson, Hawken, & Shapley, 2001, 2008; Michael, 1978) outline possible early color constancy mechanisms in V1.

While prior monkey studies shed some light on neural responses encoding perceptual (i.e. color constant) versus physical color properties in isolated regions, to our knowledge no prior study examined regional encoding of perceptually constant colors or of illuminant systematically across the whole ventral visual pathway, neither in monkey nor human brains.

We used multi-voxel pattern analysis as a test for color constancy: if a neural surface color representation is invariant across illumination changes, distinctions between representations of different surface colors should generalize across these changes. Using physically realistic rendering methods we displayed natural yet abstract 3D scenes that were displayed under three distinct illuminants. Surfaces that differed in their surface color were embedded in these scenes. We designed our experiment in this way to achieve higher ecological validity: color vision in real life occurs in 3D environments, and it is well established that some surface color cues depend on 3D scene structure (Maloney, 1999; Radonjic, Cottaris, & Brainard, 2015). Participants performed an attention task (that was independent of illuminant or color) on these surfaces during fMRI recording.

To test our hypotheses, we trained classifiers to discriminate BOLD responses to two surfaces ("blue" and "yellow") under two out of three illuminations (e.g. "neutral", and "blue") and tested them on new BOLD responses measured in the third illumination condition (e.g. "yellow"), which was not part of the training set. This analysis showed that activity patterns in V1 and V4α encoded surface colors in a way that generalized across illumination conditions, i.e. in a color-constant way.

Furthermore, we tested a prediction from equivalent illuminant models of color constancy. We hypothesized that the neural accuracy of encoding the illuminant of a scene predicts the behavioral accuracy of constant color perception. We collected behavioral color constancy indices, including a cue conflict manipulation that abolishes behavioral color constancy. We collected fMRI data for the same stimuli. In accord with the equivalent illuminant model, we found that the behavioral effect of the cue conflict manipulation on color constancy could be predicted from neural decoding of the illuminant in hV4.

Lastly, we examined how visual areas interpret two different surfaces that reflect the same light because they are presented under different illuminations. These surfaces were perceived as having distinct colors, but emitted the same light. These stimuli can be discriminated on the basis of their surface reflectance or illumination. Our analysis revealed that higher visual regions hV4, VO1, V4α weighted the difference in surface reflectance more strongly than earlier visual areas.

In sum, the results provide a detailed account of the contributions of different visual areas to color constancy.

MATERIALS AND METHODS

Participants

Our sample consisted of 20 healthy observers (15 female, 5 male) from the Tübingen University community between the ages of 19 to 35 (mean age: 24.5). All participants had normal color vision as ascertained with Ishihara plates (Ishihara, 2011). They provided written informed consent to participation in the experiment prior to the first session. The local ethics committee of the University Hospital Tübingen approved the study. Data from the fMRI main experiment of subject 12 could not be analyzed due to a data handling error. We therefore discarded this subject's dataset altogether, allowing us to use data of 19 subjects for the reported analyses.

Stimuli

Rendering Method. We used the RenderToolbox3 (Heasly, Cottaris, Lichtman, Xiao, & Brainard, 2014) to devise stimuli very similar to those used previously in a typical psychophysics experiment about color constancy (Xiao, Hurst, MacIntyre, & Brainard, 2012). The RenderToolbox3 provides a MATLAB-based framework for the development of stimuli with the 3D modeling software Blender 2.73 (www.blender.org) and the rendering software Mitsuba (www.mitsuba-renderer.org). The toolbox enables the user to control material properties (like reflectance) and the spectral power distributions (SPDs) of light sources within a 3D scene and creates 2D images of that scene based on a physically accurate rendering algorithm. The domains of reflectance functions and SPDs ranged from 380 *nm* to 730 *nm* and were discretized in steps of 10 *nm*.

Scenes, illuminants, and surfaces of interest. We rendered complex yet abstract three dimensional scenes, similar to those used in prior behavioral studies on color constancy (Xiao et al., 2012). The key motivation for using complex scenes was evidence that color constancy benefits from complex surroundings, presumably as the latter provides better cues to estimate the illuminant (Maloney, 1999; Radonjic et al., 2015). We embedded several surface patches in these scenes (see Figure 1*a*). The patches had, in different experimental conditions, two different surface reflectance functions. These functions were chosen such that the surfaces appeared blue and yellow, respectively, under "neutral", daylight illumination, which was approximated by the standard illuminant D65 (Figure 1*a*). From here onward we refer to these surfaces of interest simply as "surfaces" or "surface patches". The full scene (including the surfaces), was rendered under a total of three illuminants: the standard day-

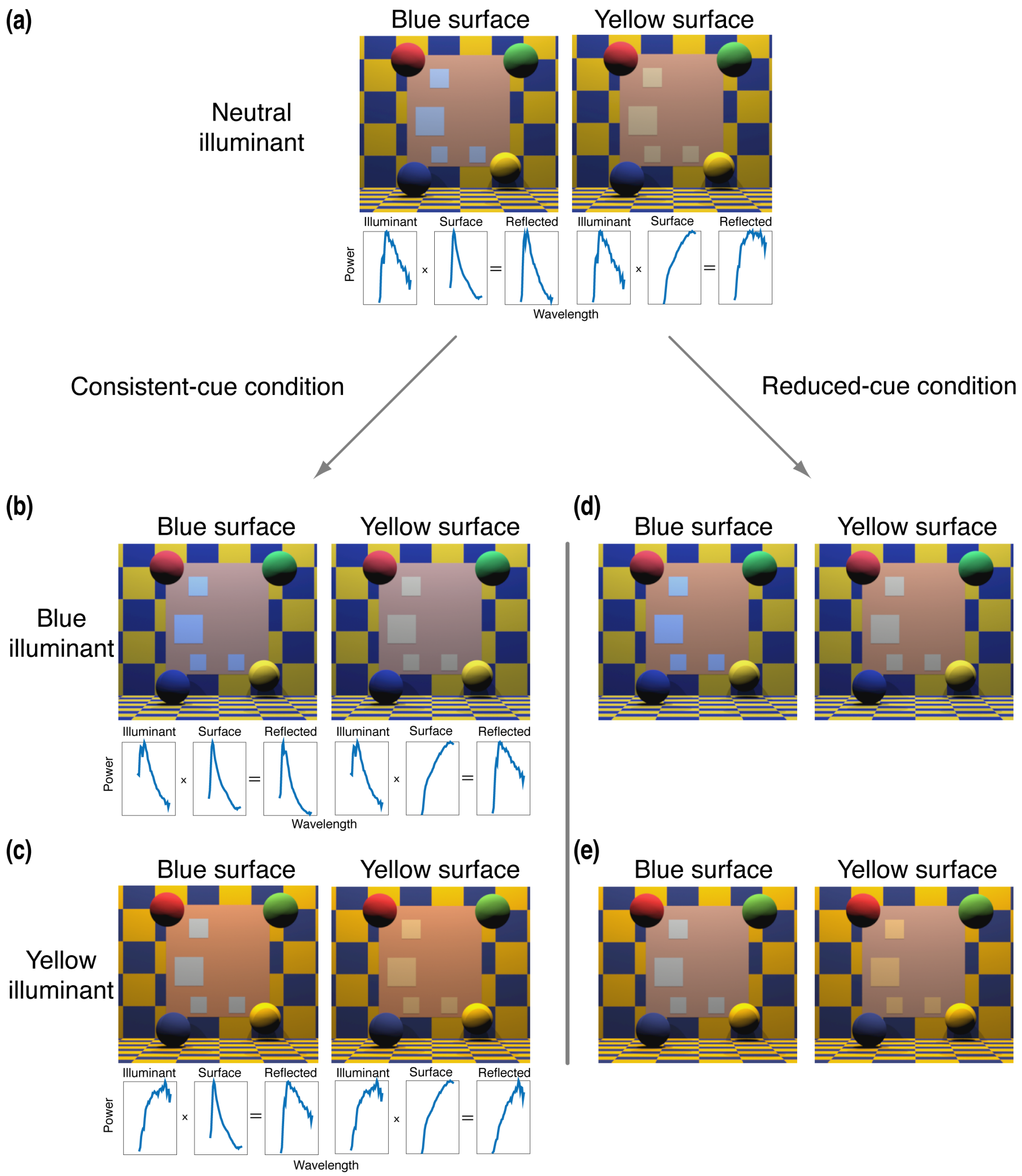

Figure 1. **Experimental Procedure and Stimuli.** **(a)** Rendered 3D scenes contained four surfaces that were either blue or yellow. In even-numbered subjects the location of the four surfaces was mirrored horizontally at the center (not shown). Note that 3 surfaces fell in one hemifield, 1 in the other. Plots below each scene image show SPDs for the illuminant, the reflectance functions for the surface, and light reflected from the surfaces. **(b, c)** Same scenes as in (a) but rendered under blue and yellow illumination, respectively. Note that the SPDs of the lights reflected from the yellow surface under blue illumination and from the blue surface under yellow illumination are the same. **(d, e)** Same as in (b) and (c) except this time the background rectangle was replaced with a different surface that reflected the same light as the original one under neutral illumination shown in (a). Note that this figure serves as illustration only, and does not allow judgment of actual color constancy effects produced by the stimuli shown in experimental conditions. Robustness of color constancy in experimental conditions is quantified in Figure 5.

light illuminant D65 (Figure 1*a*), plus two more illuminants located on the black body curve at correlated color temperatures of 10,925 K (Figure 1*b*) and 4,500 K (Figure 1*c*). For simplicity, the first illuminant is referred to as the "neutral" illuminant, the second and third as "blue" and "yellow" illuminants due to their appearances compared to daylight. The reflectance functions of the two surfaces were chosen such that the blue surface under the yellow illuminant reflected the same light as the yellow surface under the blue illuminant (compare SPDs of reflected lights in Figure 1*b* on the right and Figure 1*c* on the left). Together, the two surfaces and three illumination conditions constitute six conditions.

The colored surfaces appeared predominantly in the left visual hemi-field for odd-numbered participants (Figure 1) and in the right hemi-field in even-numbered participants (not shown). Three surface patches were shown in one hemifield and one in the other. We chose this stimulus design to exploit the retinotopic organization in visual cortex, which allowed us to test for hemispheric lateralization of BOLD responses.

Note that each rendered scene (including surface patches) provides cues on its illuminant: the illuminant illuminates the whole scene. However, and importantly, the scene excluding the surface patches does not provide any information on the surface color. Surface color cues are hence exclusively provided by the surfaces (within a given illuminant condition), and by the *combination* of the scene and the surfaces (across illuminants): under a given illuminant, the light emitted from the surfaces alone does differentiate between the two surface colors. But this is not the case across illuminants, as under the blue illuminant the yellow surface emits identical light as the blue surface under yellow illumination. To achieve color constancy, the brain needs to integrate information from the scene *and* the surface to infer the reflectance property of the surface, as it requires information of the illuminant and of the light reflected from the surface to infer its color.

Figure 2 shows the chromaticity coordinates of the two surfaces under each of the three illuminants. Average patch luminance in the neutral condition was 284.4 cd/m^2. Blue and yellow surfaces were matched in luminance. We also matched stimulus luminance across illumination conditions. The Michelson contrast between patches and the immediately surrounding (i.e., within 5 pixels) background surface was 0.179.

Cue conflict conditions. In addition to the above, we introduced cue conflict conditions shown to strongly impair color constancy judgments (Delahunt & Brainard, 2004; Xiao et al., 2012). The manipulation consists in replacing the central large background rectangle on which the blue or yellow colored surfaces appeared with a background that reflected the same light under each illuminant as the original background under neutral illumination (Figure 1*d* and *e*). We refer to the condition involving the original background as the *consistent-cue* condition and to the condition with the replaced background as the *reduced-cue* condition (following Xiao et al., 2012). The manipulation was applied for the blue and yellow illuminant and both surfaces, yielding four cue conflict conditions – the neutral cue-conflict condition was identical to the neutral consistent-cue condition. In total, we had ten different conditions.

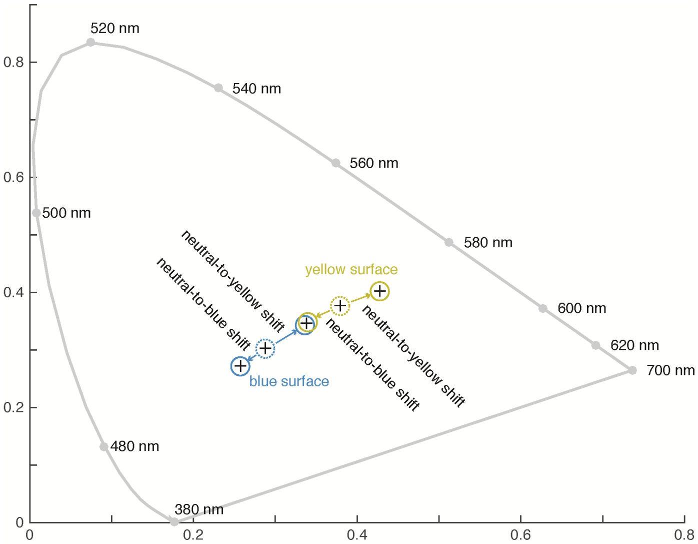

Figure 2. **Stimulus chromaticities.** Stimulus coordinates (x, y) in the CIE 1931 chromaticity plane. Blue (yellow) circles mark chromaticities of the blue (yellow) surface. Dotted circles indicate chromaticities under D65 illumination. Arrows represent chromaticity shifts induced by changes in illumination. Note that a chromaticity shift from neutral to blue for the yellow surface results in the same chromaticity as a shift from neutral to yellow for the blue surface as indicated by overlapping blue and yellow circles.

To make sure that classification of surface colors was driven only by differences between the chromaticities of the two surfaces and not by luminance differences, took two measures: first, we equated luminances across both surface colors and all illumination conditions. This was achieved by setting the images to the average luminance values across the ten conditions on a pixel-by-pixel basis. Second, one half of the stimuli was presented with luminance increased and the other half with luminance reduced by 10 %. This ensured that classifiers generalized across possible differences in luminance and made discriminations based on chromatic differences. To accomplish this, we converted the RGB values of all images to XYZ tri-stimulus space using the transformation matrix obtained from the display calibration. We then calculated the mean luminance (Y) across all images. We converted each XYZ image to CIE xyY space to separate chromaticity from luminance components and applied the mean luminance vector to every image (including the two 10 % luminance modulations). The images were subsequently transformed back to RGB space for presentation on the gamma-corrected displays.

fMRI experiment

Participants lay supine in the scanner and viewed the scene images via a mirror mounted on top of the head coil. Stimuli were presented against a screen located at the end of the scanner bore using a gamma-calibrated projector (NEC PE401H, CalibrateMonSpd.m function from Psychtoolbox, spectroradiometer by Photo Research PR-670). The size of each image on the projection screen was 16.8° and 15° of visual angle in horizontal and vertical directions, respectively, at a resolution of 1024 x 768 pixels.

Stimulus presentation

The sequence of trials is shown in Figure 3. Each scene appeared four times for 1.5 *s* within a stimulus block in alternation with a luminance matched color mask that lasted for 1 *s*, leading to a block-duration of 10 *s*. Color masks consisted of three RGB layers created independently from 1/f noise. All stimuli were presented using MATLAB and Psychtoolbox (Kleiner, Brainard, & Pelli, 2007). Stimulus blocks appeared on the screen in a pseudo-randomized sequence that made sure that every pair-wise transition between conditions was equally likely across the whole experiment (Brooks, 2012). All ten conditions were presented 40 times each in a total of eight runs. Each run lasted 8 min 20 *s* plus 8.7 *s* for 11 dummy scans and 8 *s* extra scan time.

Task

Participants maintained fixation on the fixation cross in the center of the screen while paying attention to the colored surfaces surrounding it. Their task was to respond via button press whenever they detected a target event. Targets were a temporary decrease in the luminance of the white fixation cross or a temporary change in the number of colored surfaces. One of the surfaces at the 11, 9, 7, or 5 o'clock positions sometimes disappeared for a short period or an additional surface appeared at the 1 or 3 o'clock position. The target event lasted 0.5 *s* in each case. Each of the six surface events occurred in 3.75 % of the trials while fixation cross events occurred in 22.5 % of the trials yielding a total target probability of 45 % (= 22.5 % + 6 locations * 3.75 %) for each condition. To increase motivation, participants received feedback about their performance at the end of each run.

Retinotopic mapping & ROI definition

Visual areas V1-V3, hV4, and VO1 were localized using standard retinotopic mapping techniques (Sereno et al., 1995). In brief, observers fixated in the middle of the screen while attending a flickering black and white checkerboard visible through a wedge-shaped aperture on a gray background. Check sizes increased logarithmically from the center to the visual periphery to account for cortical magnification. The wedge subtended the entire screen within an angle of 90° and rotated with a period of 55.64 *s* in clockwise or counter-clockwise direction. Participants viewed ten cycles of this stimulus in four polar mapping runs with stimulus direction alternating between runs. The functional data were analyzed using Freesurfer software (http://surfer.nmr.mgh.harvard.edu/) to obtain polar maps on a flattened 2D reconstruction of individual brains that allowed for the detailed demarcation of single visual areas.

In addition to retinotopic ROIs we also created a group ROI located anterior to the retinotopically defined ROIs based on the results of our within-illuminant searchlight analysis (see section "fMRI pattern classification: searchlight analysis" in Results). This ROI was defined as the smallest sphere that encompassed an entire cluster of information located in a brain region that has previously been labeled V4α (Barbur & Spang, 2008; Bartels & Zeki, 2000). We hence refer to this region as putative V4α (or pV4α for short). The MNI coordinates of the center of the sphere were x = 39, y = -58, z = -11 and its radius was 6 *mm*. Importantly, the searchlight analysis used to define pV4α was independent from the subsequent classification tests carried out on this ROI.

fMRI scan details

We used a 64-channel coil to record functional activity at 3 T magnetic field strength (Siemens Prisma) with a GRAPPA accelerated T2*-weighted 4-fold parallel imaging sequence (GRAPPA factor 2, 0.87 *s* repetition time (TR), 57° flip angle, 30 *ms* echo time (TE), 56 slices without gaps, 96 x 96 acquisition matrix, 2 *mm* isotropic voxel size). A T1-weighted MP-RAGE sequence was used to acquire an anatomical image of each subject's brain at a resolution of 1 x 1 x 1 mm^3. To correct for magnetic field inhomogeneity, we measured a Gradient Echo field map in each fMRI session.

fMRI data preprocessing

Preprocessing was carried out using SPM8 (http://www.fil.ion.ucl.ac.uk/spm). We applied motion correction and unwarping (using the voxel displacement maps obtained from the Gradient Echo field map) to our functional data, corrected them for slice acquisition time, and co-registered them to the anatomical scan. The anatomical scans were segmented and normalized to MNI space along with the functional data. The co-registered functional data from the retinotopic mapping experiment were further smoothed with a 4 *mm* (FWHM) Gaussian kernel.

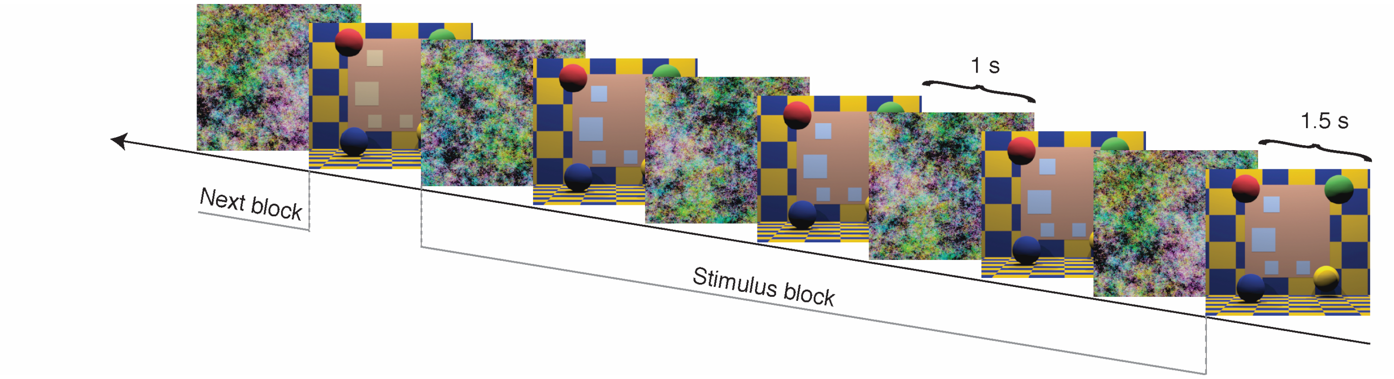

Figure 3. **Stimulus Presentation Sequence.** Each scene was presented four times for 1.5 s, each time followed by a random color mask shown for 1 s. Observers were instructed to press a button whenever the fixation cross was dimmed or when the number of colored surfaces changed.

fMRI pattern estimation

To estimate the patterns of fMRI responses to our stimuli under various conditions, we modeled the normalized, unsmoothed voxel time series with GLMs. Every stimulus block (Figure 3) was modeled separately as a boxcar function shifted forward in time by 5 *s* to account for the hemodynamic lag. Additional regressors modeled the estimated motion parameters. For each time step voxel values were scaled by the global mean value. Instead of applying SPM's frequency cutoff, the resulting series of beta estimates were temporally filtered by removing linear and quadratic trends (and intercept) from each voxel per run. Finally, we z-scored the beta estimates for each voxel and replaced values above 2 with 2s and values below -2 with -2s to handle outliers as recommended by the LIBLINEAR authors (see below).

fMRI pattern analysis

All classification analyses were carried out using leave-one-run-out cross-validation, and all inference was based on bootstrapping and family-wise error correction (FWE) (for inference see section on statistical inference below). In detail, classification accuracy was calculated by taking the sum of correct percentages across validation folds weighted by the number of predictions made in that run. If training sets were unbalanced (i.e., contained more samples from one category than the other), we imputed the missing values by setting them to the average pattern of that category. We used the LIBLINEAR implementation (https://www.csie.ntu.edu.tw/~cjlin/liblinear/) of SVM (Fan, Chang, Hsieh, Wang, & Lin, 2008). The training algorithm determined the optimal value for C by performing 3-fold cross-validation on training data and selecting the value that minimized prediction error. C is a penalty parameter that controls the trade-off between training set error and generalization to prevent overfitting. Candidate values for C were powers of 2: $2^{-18}, 2^{-17}, \ldots, 2^{1}$.

Within illuminant classification. To provide a benchmark of decoding accuracies for subsequent analyses, we first tested if the two surface colors could be predicted from fMRI activity patterns when training and test set data came from the same illuminant conditions. In this case above-chance classification does not specifically depend on perceived surface color but can be achieved simply on the basis of different wavelength compositions of the lights the surfaces reflect. Discriminability of surface color within the same illumination conditions is rather a prerequisite if classification is expected to work when training and testing are performed on different illumination conditions. This analysis can be regarded as a replication of previous research (e.g., Brouwer & Heeger, 2009) but extends it to the more complex context of our 3D rendered scene that embeds the surfaces of interest. We proceeded by training classifiers to distinguish between blue and yellow surfaces under each two pairs of illuminants (neutral/blue-consistent-cue, neutral/yellow-consistent-cue, blue-consistent-cue/yellow-consistent-cue) and tested them on responses to stimuli from the same two conditions (recorded in the withheld run) Figure 4*a*). The average across all three possible combinations was taken as the dependent variable in this analysis.

We provided the algorithm with vectors of voxel response patterns (one for each trial), and for each one label corresponding to the surface color (regardless of illuminant) that was presented to the observer when that measurement was made. The algorithm was "blind" with regards to the illumination that the measured pattern came from. The classifier thus had to treat any variation due to our luminance modulation or illumination as within-class variation. This approach helps the classifier select especially those voxels that are more likely to represent color in a constant way while avoiding voxels that are driven by luminance differences. At the same time, this approach allowed us to exploit as much of our data as possible in our analyses. This training proce-

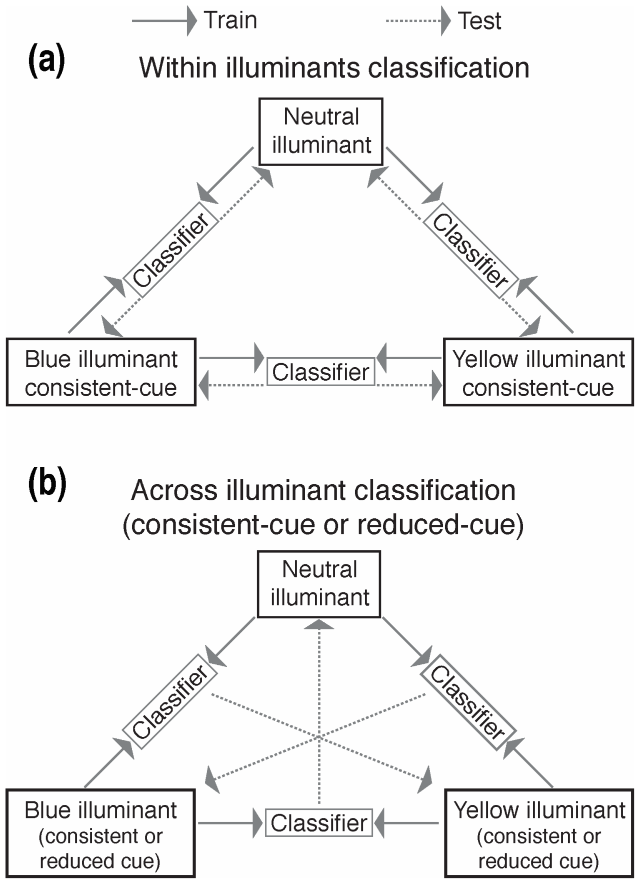
Train
Test
(a)
Within illuminants classification
Neutral illuminant
Classifier
Classifier
Blue illuminant consistent-cue
Classifier
Yellow illuminant consistent-cue
(b)
Across illuminant classification (consistent-cue or reduced-cue)
Neutral illuminant
Classifier
Classifier
Blue illuminant (consistent or reduced cue)
Classifier
Yellow illuminant (consistent or reduced cue)

Figure 4. **Classification Analyses** Illustration of the first three classification analyses performed on fMRI response patterns. These classifications aimed to distinguish between blue and yellow surface conditions, within or across varying illumination conditions, respectively. **(a)** Three classifiers were trained to distinguish responses to blue and yellow surfaces under two different illuminants and tested on responses to stimuli from the same illumination conditions that were not used for training. No generalization across illuminations was required. Results were averaged across the three analyses. **(b)** Training as in (a), but this time classifiers were tested on responses to stimuli from the third illumination condition that was not included for training. The analysis for (b) was carried out separately for consistent-cue and for reduced-cue stimuli, respectively.

dure was chosen for all three analyses (Figure 4) to allow for comparison.

The critical difference between the classifications within-illuminants and across-illuminants is this: In within-illuminant classification the test data from the withheld run came from the same two illuminant conditions while in across-illuminant classification the test data came from the third illuminant condition.

The reason why we used data of two illuminants also for training in our control ("within illuminant") analysis is described in the following. For "within illuminant" we trained a classifier to distinguish "blue" and "yellow" surfaces that were viewed under two illuminants (e.g. "neutral" and "blue"). Testing was done for the same two illuminants on trials of a left-out run. Hence, for ROIs that did not have illuminant-invariant representations, the classifier would learn two patterns for "blue" surface (one for each illuminant) and another two for "yellow" surface, and perform well on testing using trials taken from either illuminant. Testing on a new (untrained) illuminant would fail. The advantage of this training procedure on two illuminants, and the reason why we used it, was that for ROIs that did have illuminant-invariant representations, training would encourage the classifier to rely on voxel patterns that were invariant with respect to the illuminant. Thus, for such ROIs, classification on a new (untrained) illuminant would work.

This was hence the best training scheme to maximize success of our main classification (across illuminants). In order to have a control analysis that was as similar as possible to the main analysis, we used the same training scheme – only the test trials differed between control and main analyses.

Across illuminant classification (consistent-cue condition). The critical test for the invariance of surface color representations is to examine how well they generalize across changes in illumination. In other words, how well can a classifier predict surface color when the data used to train the classifier come from different illumination conditions than the data used for testing? Three classifiers were trained to distinguish between blue and yellow surfaces under each two pairs of illuminants, as described above for within-illuminant decoding. However, testing was carried out on the left-out illumination for each of the classifiers and their accuracies were averaged. In detail: one classifier was trained to distinguish blue and yellow surfaces on responses from "neutral" and "blue-consistent-cue" illumination conditions and tested on responses from the "yellow-consistent-cue" illumination condition; one was trained on data from "neutral" and "yellow-consistent-cue" conditions and tested on data from the "blue-consistent-cue" condition; and one was trained on "blue-consistent-cue" and "yellow-consistent-cue" condition responses and tested on "neutral" condition data. Each classifier was thus trained to discriminate between surface colors under two illuminants and then applied to responses to the same surfaces under the third illuminant that was not included in the training set (Figure 4*b*). Note that a change in illumination alters the wavelength compositions of the surfaces and that generalization accuracy above chance indicates that the color representation of the surface is invariant with respect to the illuminant under which it is viewed. It is also important to appreciate that this classification scheme was comparable to the one applied for the classification within illuminants because the same data were used for training.

Across illuminant classification (reduced-cue condition). The purpose of the reduced-cue condition was to investigate the sensitivity of any neural signature of constant colors to manipulations that are known to reduce color constancy as measured psychophysically. The classification scheme in this analysis is identical to the one described above with one exception: instead of using the responses elicited under the consistent-cue blue condition and the consistent-cue yellow condition we used the responses elicited by stimuli from the reduced-cue blue and reduced-cue yellow conditions, respectively (Figure 4*b*).

Ipsilateral and contralateral ROIs. As depicted in Figure 1*a*, the four blue or yellow surfaces appeared mostly in one visual hemifield. The surface locations shown in Figure 1 correspond to the locations for odd-numbered subjects. Their locations were mirrored horizontally for even-numbered subjects. We analyzed the ROI data separately for ipsilateral and contralateral regions with respect to the hemifield where more surfaces appeared.

Searchlight analysis. To obtain a more global view of the brain responses coding for surface color, the three types of classifications detailed above were also carried out in searchlight analyses (Kriegeskorte, Goebel, & Bandettini, 2006). SVM classifiers were trained and tested in the same way as in the ROI analyses (within illuminants, across illuminants / consistent-cue, and across illuminants / reduced-cue) on local patterns of fMRI responses within a radius of 3 voxels.

Statistical inference: ROI results. Non-parametric permutation tests were performed to test whether classification accuracy was above chance (chance level being 50 % for binary classifications). Class labels in the training set were shuffled randomly 10^3 times and new classification models were fit to the data and tested on the intact withheld test set. When bootstrapping null distributions for classifications that required averaging across classification accuracies, we took care that the same label permutations were applied in each of the classification analyses involved. This procedure yielded a distribution of mean classification accuracies at the group level that would be expected if no relationship existed between the multivariate data and the class labels in the training set. To correct for multiplicity due to the number of ROIs tested, we controlled the family-wise error (FWE) in the following way (e.g., Blair & Karniski, 1993): within each permutation step, the randomized label assignments were kept identical for all ROIs within individual subjects. Group mean ROI classification values were then obtained for each ROI. Only the maximum group mean value across all ROIs was used for the null distribution. Therefore, a common null distribution was used for all ROIs that controlled the error probability of at least one null hypothesis being falsely rejected.

Statistical inference: searchlight analyses. Individual searchlight maps were spatially smoothed with a 6 *mm* (FWHM) Gaussian kernel. We tested if decoding was significantly better than chance (i.e., 50 %) using a one-sample *t* test across participants (df = 18) and null hypotheses were rejected for *p* values below .001 (uncorrected). The maps were masked with each participant's whole brain mask only.

Surface/illuminant bias analysis. The purpose of the aforementioned analyses was to test to what extent surface color representations are invariant with respect to changes in illumination. Since illumination changes alter the SPDs of the light reflected off those surfaces, we considered a complementary question as well: what happens if instead the wavelength distribution of the reflected light is identical for two stimuli that differ with respect to surface and illumination colors? Some brain regions may be biased to encode reflectance, others illumination.

We devised stimuli and performed an analysis that can estimate this bias (Seymour, Clifford, Logothetis, & Bartels, 2009, 2010; Seymour, Williams, & Rich, 2015): Figure 8 shows that the light reflected from the yellow surface under blue illumination is identical to the light from the blue surface under the yellow illumination (top-right and bottom-left stimuli in Figure 8). Classifiers were trained to distinguish between responses to these two stimuli, but tested on responses to the other two stimuli. The classifier results would hence reveal whether a ROI encodes primarily the perceived surface color (i.e. reflectance), or illumination. Either classifier response would be "correct", but reveal the encoding bias of a given ROI. If the classifier assigns the same labels to stimuli in the same row, the neural representation weights the influence of illumination on the incoming signal more strongly. If it assigns the same labels to stimuli in the same column, the neural representation emphasizes the difference in surface reflectance.

In contrast to our previous ROI analyses, we used a two-tailed permutation technique to test if classifier predictions indicated a representational bias that was significantly different from 50 %. This means that the group null distribution this time did not consist of the largest classification value per permutation step but the value with the largest absolute difference from chance level (i.e., 50 %). This leads to a bimodal null distribution. Reported *p* values are the proportion of samples in the null distribution that are above or below the observed value (whichever is smaller) multiplied by two.

In order to interpret a bias in favor of illuminant or surface encoding, it is informative to check if illuminant or surface information can be decoded when classifiers are explicitly trained to discriminate along one or the other stimulus dimension. A region that showed no bias for illuminant or surface encoding for instance, would have the same bias as a region that contained only noise. To distinguish between these scenarios, we conducted two control analyses using the same data as in the bias analysis. In explicit surface decoding, classifiers were trained to discriminate between responses to blue and yellow surfaces (under blue or yellow illumination) and tested on how well they could predict surface color in an independent test set (using leave-one-run-out cross-validation). In explicit illuminant decoding, illuminant and surface switched roles such that classifiers were instead trained and tested on how well they could predict illuminant irrespective of surface color. If activity within a ROI was only noise, none of these classifications would reveal any information.

Finally, we complemented this analysis with a comparison between representational dissimilarity matrices (RDMs) of the measured voxel responses and with two different model RDMs (Nili et al., 2014). For each ROI, we calculated one 4-by-4 RDM. Each entry represented the pairwise dissimilarities (1 - Pearson correlation coefficient) between the average response vectors (voxel values for a given condition of a given ROI) of two conditions. The RDM was calculated for the same four conditions used in the bias analysis. These are given by the two binary factors surface (blue or yellow) and illumination (blue or yellow). The observed RDMs were tested for agreement with two different model RDMs (shown in Figure 9*c*) using rank correlations. The surface model assumed that neural activity in a ROI represented differences in surface reflectance but did not distinguish between illuminations. It hence predicted a maximal dissimilarity of 1 between patterns of responses to different surfaces and a dissimilarity of 0 between illumination responses (i.e., a complete surface bias). The illuminant model made the opposite prediction that neural responses only reflect differences in illumination without discriminating between surfaces. It therefore predicted a dissimilarity of 1 between responses to different illuminations and a dissimilarity of 0 between different surface representations (i.e., a complete illuminant bias). Multidimensional scaling (MDS, metric stress criterion) was used to discover relative commonalities between similarity structures in neural encoding within multiple ROIs and the hypothetical model RDMs predicted by theory. We performed hierarchical clustering (average linkage, Euclidean distance) and examined the resulting dendrograms to check whether the representational similarity structures within ROIs preferred a clustering by surface conditions over illumination conditions or vice versa.

Psychophysics

Human observers differ in their ability to perceive a given surface as the same color when the illumination varies, i.e. in their color constancy. This can be quantified by the color constancy index (CI). We measured color constancy in all our participants behaviorally using an alternating staircase procedure described by (Xiao et al., 2012). This method finds the chromaticity of a color that appears achromatic under the illumination of the scenes presented to observers. A CI can be computed with respect to a pair of illuminants from the achromatic settings made under each of them (Brainard, 1998). The color constancy index is based on the notion that a perfectly color constant perceptual system should shift its achromatic point in the direction of the illuminant change while a perceptual system without any color constancy should exhibit identical achromatic points under both illuminants (see Supplementary Information for formulas). A CI of 0 thus indicates absence of color constancy while a CI of 1 means perfect color constancy (although CIs are not necessarily bound between 0 and 1).

We identified the achromatic point for observers under each viewing condition by instructing them to judge the appearance of a briefly flashed color circle (750 *ms*) in alternating blocks as either more reddish than greenish or as more bluish than yellowish. The judgments in every trial were used to update the chromaticity of the circle such that it appeared increasingly achromatic. Our methods were identical to those in (Xiao et al., 2012) except that we adjusted the chromaticity in the staircase procedure not in RGB space but in a subspace of the perceptually uniform CIE L*a*b* space. To accomplish this, the maximal red, green, blue, and yellow RGB values of the calibrated display were converted to a*, b* coordinates. We defined the red/green and the blue/yellow directions as the unit difference vectors between the two coordinate pairs for each direction. Another difference was that the color circle for which the color judgment had to be made did not appear in the center of the screen but in the part of the rectangle where no colored surfaces were present, i.e., in the top-right corner for participants with odd subject numbers, in the top-left for all others.

In addition to determining the CI for each participant, we also determined $\mathrm{CI}_{\text{reduced-cue}}$ by applying the same cue conflict manipulation also applied for part of the imaging stimuli (see “Stimuli: cue conflict conditions”). This manipulation is known to heavily impair color constancy judgments in human observers (e.g., Delahunt & Brainard, 2004; Xiao et al., 2012), and allowed us to relate its behavioral effect to individual brain decoding results (see section below).

Hence, two CIs were calculated for each participant by taking the average of all CIs in the consistent-cue and reduced-cue conditions, respectively (ignoring conditions in which staircases did not converge, see Results).

Psychophysical data analysis

The mean CIs in the consistent-cue and reduced-cue conditions were compared using a paired one-tailed *t*-test to ascertain that the cue validity manipulation did in fact show a decrease of color constancy in the cue conflict condition relative to the consistent-cue condition. Finally, we were interested in the psychophysiological relationship between fMRI measurements of neural activity and behavioral measurements of color constancy. Specifically, we tested a prediction from the Equivalent Illuminant Model of color constancy (Brainard & Maloney, 2011). It assumes that color constancy depends on the chromaticity of the illuminant estimated by the perceptual system. Whether color constancy fails or not depends on the accuracy of this estimate (the "equivalent illuminant"). To test this prediction, we calculated the difference in prediction accuracies for the three illuminants between consistent-cue and reduced-cue conditions in each ROI and subject. We then examined if these differences between conditions predict the individual decrease in color constancy induced in the reduced-cue condition using linear regression models.

RESULTS

We rendered a complex scene containing four surfaces that appeared either blue or yellow. Three different illuminations were simulated: neutral D65, a blue illumination, and a yellow illumination (Figure 1 *a*, *b*, *c*). Additionally, we introduced a reduced-cue condition for the blue (Figure 1*d*) and yellow (Figure 1*e*) illumination conditions: in these conditions the background square on which the colored surfaces appeared was replaced with a surface that was chosen such that the light reflected from it was the same as in the neutral condition. This manipulation is known to strongly impair color constancy in human observers (Delahunt & Brainard, 2004; Xiao et al., 2012). We reasoned that a neural correlate of color constancy should be just as susceptible to this manipulation as behavioral color constancy measures. Furthermore, the reflectance of the blue and the yellow surfaces were chosen such that the light reflected from the yellow surface under blue illumination (see power spectra in insets in Figure 1*b* on the right) was the same as the light reflected from the blue surface under yellow illumination (see insets in Figure 1*c* on the left). This allowed us to study the degree to which different brain regions are biased to encode surface color versus illuminant properties in a situation when the reflected light is physically identical for both stimuli.

Participants completed a neuroimaging session in which they viewed the rendered scenes on a projection screen while lying supine in the scanner and performing a fixation and spatial attention task (Figure 3). They participated in a behavioral experiment to measure color constancy and a retinotopic mapping experiment.

We trained linear classifiers to discriminate between the two surface colors on the basis of fMRI activation patterns and examined how well they generalize to new samples that were measured either under the same illuminations (Figure 4*a*) or a new illumination condition that has not been included in the training set (Figure 4*b*). Generalization to a new illumination was used as an indicator for color constancy.

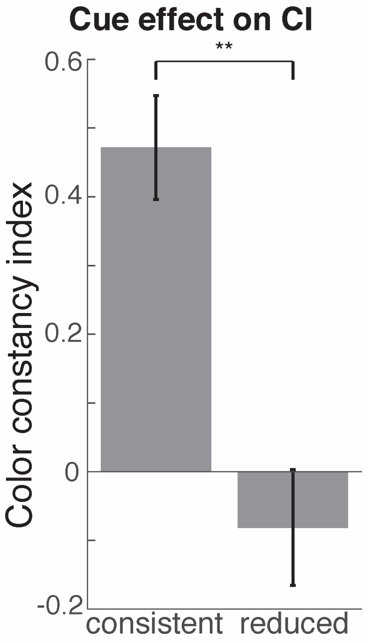

Figure 5. **Behavioral Results** Color constancy was significantly better in the consistent-cue than in the reduced-cue condition (paired *t* test, t_{17} = 10.566, $p < .001$).

Psychophysics: Cue condition effect

The reduced-cue conditions differed from the consistent-cue ones in that the rectangular background surface of the rendered scene reflected the same light in both blue and yellow illumination: that reflected under neutral illumination. As expected, this manipulation led to a strong decrease in color constancy relative to the consistent-cue condition – in fact, in the reduced-cue condition color constancy was completely abolished (Figure 5). Accordingly, mean CI was significantly larger in the consistent-cue than in the reduced-cue condition ($CI_{\text{consistent-cue}}$ = .4716, $CI_{\text{reduced-cue}}$ = -.0812, t_{17} = 10.566, $p = 3.4 \times 10^{-9}$, one-tailed). Note that the behavioral data from subject 14 could not be used because the staircases did not converge for any of the illuminant conditions.

ROI-based fMRI pattern classification: within illuminants

The first fMRI data analysis examined the simplest scenario in which classifiers were trained on samples of the same illuminant conditions as those in the test set. We trained linear classifiers to distinguish between blue and yellow surface colors under pairs of illuminations (e.g. neutral and blue) and tested them on left-out samples from the same illuminations; results were averaged for all combinations of illumination pairs (see Figure 4*a*). Training was done for pairs of illuminants to provide a comparison point to analogous subsequent analyses. This analysis does not test for any generalizability across illuminants. It tests for the discriminability of BOLD responses caused by lights of different wavelength compositions, and thus provides a baseline for decoding accuracy for the current data. Our ROIs comprised the visual areas we identified in the retinotopic mapping session and a functionally defined ROI anterior to them found in the searchlight analysis (for ROI definition, see "fMRI pattern classification: searchlight analysis" below). Consistent with previous research (Brouwer & Heeger, 2009), the two colors led to different patterns of fMRI activity in almost all ROIs examined (Figure 6*a*), especially in contralateral ROIs. Note that surface stimuli had been presented in both hemifields, but with an asymmetry of three in the contralateral and one in the ipsilateral hemifield (see methods). Classification accuracy ranged from 51.38 % (Cohen's $d = .46$) in ipsilateral hV4, marginally failing to reach significance ($p = .057$, FWE corrected), to 54.42 % (Cohen's $d = 1.2$, $p = .001$, FWE corrected) in contralateral V1. We conclude that surface color can be predicted from fMRI activity patterns in all ROIs (except ipsilateral hV4) when classifiers have been trained on samples from the same illuminant conditions as those in the test set. Although mean classification accuracies were not much higher than chance, the effect sizes are considerable in magnitude. The result for putative V4α in Figure 6*a* is reported for completeness only because it is based on voxels enclosing a cluster of information identified in the same (and therefore non-independent) within-illuminant analysis conducted with the searchlight technique. Note that pV4α results in Figure 6*b* and *c* are independent of the ROI-defining contrast.

ROI-based fMRI pattern classification: across illuminants (consistent-cue)

This analysis addressed the question whether voxel patterns evoked by the blue and yellow surfaces under one set of illuminants would generalize to new illuminants as well. Generalization of response patterns to new illuminant conditions would imply the invariance of surface color representations with respect to illumination. To this end, we trained classifiers in the same way as in the previous analysis within illuminants but this time tested them on data from an illuminant condition that had not been part of the training set (see Figure 4*b*). Our analysis showed (Figure 6*b*) that, among the retinotopically defined ROIs, only the V1 area contralateral to the three surfaces allowed predicting the surface color using a classifier trained on responses to stimuli simulating illuminations that were not included in the training set (54.41 %, $p = .001$, FWE corrected, Cohen's $d = .88$). Putative V4α also exhibited surface color decoding across illuminants significantly above chance (52.04 %, $p = .033$, FWE corrected, Cohen's $d = .43$). We did not, however, observe decoding accuracies significantly above chance in any other ROI, with the highest non-significant decoding accuracy found in contralateral V2 (51.21 %, $p = .427$, FWE corrected, Cohen's $d = .33$). Our analysis thus demonstrates that neural representations of surface color in V1 and pV4α are invariant with respect to illumination changes. This invariance was only observed in these two regions as the patterns of responses to surface color in all other ROIs were not found to generalize across illuminants.

Control analysis: classification across cue conditions. In order to rule out that the null results for across-illuminant surface decoding in hV4 and VO1 was driven by bad signal quality, we performed additional control analyses. In particular, previous research has suggested that fMRI signal quality in V4 can suffer from the presence of nearby blood vessels, which may put this region to a particular disadvantage (Winawer, Horiguchi, Sayres, Amano, & Wandell, 2010). We trained classifiers to discriminate between surface colors (blue vs. yellow) in the consistent-cue condition (using blocks from the blue and yellow illuminant conditions) and tested them using the corresponding blocks of the reduced-cue condition (and vice versa). The reasoning is the following: the surface-squares in consistent-cue and reduced-cue conditions have identical light emissions (i.e. wavelength-based information). However, the surface appearances were more clearly distinguishable as blue and yellow in the consistent-cue compared to reduced-cue blocks. Hence, a region encoding appearance (i.e. perception of constant color) should suffer in decoding performance in this cross-testing scenario, whereas a region encoding wavelength-based information should not. Paired *t* tests in the contralateral hemisphere showed that decoding accuracies were indeed larger in hV4 (56.58 %) than in V1 (53.29 %, $t_{18} = 2.5617$, $p = 0.0392$, all one-tailed and Bonferroni corrected for four comparisons) and pV4α (50.82 %, $t_{18} = 3.9504$, $p = .0019$, Bonferroni corrected, see Figure S1 for results from all ROIs). For VO1 (53.27 %), however, the differences relative to pV4α (50.82 %, $t_{18} = 1.4881$, $p = 0.077$, uncorrected) or V1 (53.29 %) were not significant. This finding rules out the possibility that signal quality in hV4 was generally worse than in V1 and pV4α.

ROI-based fMRI pattern classification: across illuminants (reduced-cue)

The observation that classifiers could predict surface color from response patterns in V1 and pV4α even when training and test data came from different illumination conditions raises an important question about the functional role of such activity for color constancy. If it is related to color constancy, there must be an experimental manipulation that simultaneously affects activity in these regions as well as psychophysical measures of color constancy. Our experimental design included a reduced-cue condition, in which a background surface was made to emit the same light in all three illumination conditions. How does this manipulation affect brain activity on the one hand and behavior on the other? As confirmed psychophysically, the reduced-cue condition completely abolished color constancy in our observers (see Figure 5). The classification analysis was carried out in the same way as in the analysis across illuminations in the consistent-cue condition (i.e., data in training and test set came from different illumination conditions) except that now only responses from the reduced-cue illumination conditions were used (Figure 1 *d*,*e*). In this analysis, classification accuracy was not different from chance in any of the examined ROIs (Figure 6*c*). Classification accuracy was highest in ipsilateral area V2 (51.47 %, *p* = .293, FWE corrected). Importantly, prediction accuracy was not significantly above chance in contralateral V1 either (50.68 %, *p* = .882, FWE corrected). Classification of responses in pV4α similarly did not exceed chance (49.76 %). Taken together with the psychophysical finding, this analysis shows that a manipulation that strongly impairs color constancy also causes surface color decoding across different illuminant conditions to fail. This suggests that V1 and pV4α activity may have contributed to color constancy in our experiment by encoding surface color in terms of chromatic contrast.

fMRI pattern classification: searchlight analysis

Since there may be color-responsive activity beyond retinotopically mapped regions relevant to our task, we repeated the same classifications performed for the ROIs also at the whole brain level by means of the searchlight method (3 voxel radius) (Kriegeskorte et al., 2006).

Figure 7*a* shows searchlight results for the within-illuminants analysis. Significant decoding was apparent throughout the occipital cortex including the calcarine gyrus, as well as in the fusiform gyrus primarily contralateral to the surface stimuli. Note that we left-right flipped searchlight maps for participants with even subject numbers to ensure that for all subjects contralateral stimulation was on the right side (positive values of x).

Figure 7*b* shows the results from the searchlight analysis across illuminants (consistent-cue). It revealed a cluster of voxels in the fusiform gyrus where differences between local patterns of fMRI activity distinguishing the two surface colors generalized across illuminants. This cluster was located anterior to the ROIs we had examined, with the MNI coordinates of the peak voxel being x = 34, y = -54, z = -10 (Figure 7*b*). This fusiform region overlaps with voxels that exhibit classification accuracies above chance in the within-illuminant searchlight analysis.

As expected from the null-findings in the ROI analyses, also the searchlight analysis on reduced-cue across-illuminant decoding did not reveal any significant results.

The functional properties of the anterior fusiform region thus resemble those of area V1. Both regions allowed predictions of surface colors from local brain activity within the same illuminant conditions as well as across different illuminants. The reduced-cue abolished generalizability across illuminants in both regions. Figure 7*c* shows the location of this cluster in relation to the retinotopically defined ROIs in a cortical surface rendering.

Previous studies have already identified two separate color-responsive regions in the fusiform gyrus (Barbur & Spang, 2008; Bartels & Zeki, 2000; Beauchamp et al., 1999; Wade, Augath, Logothetis, & Wandell, 2008). This region has often been referred to as V4α. The peak voxel of the cluster in our searchlight analysis was located in close vicinity to the peak voxels listed for V4α in the review by Bartels and Zeki (2000).

For a more direct comparison with the results from our ROI analyses, we created a ROI from all voxels within a sphere just large enough to encompass the whole information cluster circled in Figure 7*a* (see also "Retinotopic mapping & ROI definition" in Methods and Materials). We included this region as putative V4α (or pV4α) in our ROI analyses. Results for this region can be seen to the right of the dotted lines in the plots of Figure 6.

Due to the similar response properties of V1 and pV4α, we explored the connectivity between these regions. Partial correlation analyses of the mean residual time series per ROIs showed that there was a small, yet significant amount of unique variance shared between V1 and putative pV4α (*r* = -.038, *p* = .0479, two-tailed, Holm-Bonferroni corrected for 15 tests, Figure S2).

Surface/illuminant bias analysis

It is well known that color constancy depends on our estimation of the illuminant (Delahunt & Brainard, 2004; Xiao et al., 2012). For example, uncertainty about the illuminant within a scene can change the perceived color of a given foreground object dramatically (e.g., "the dress", Gegenfurtner, Bloj, & Toscani, 2015). It is hence likely that some neural representations primarily encode information about the current illuminant, while others primarily encode the color, i.e. estimated surface reflectance. While our previous analyses focused on the robustness of surface color representations in the face of changes in illumination and hence wavelength distribution, we here investigated whether ROIs had a bias in encoding surface color or illuminant information when the visual system is presented with lights composed of identical SPDs.

We had designed stimuli such that the yellow surface reflected the same light under blue illumination (SyIb) as the blue surface under yellow illumination (SbIy) (Figure 8). Hence, while the surfaces emitted identical wavelength information, their colors (i.e. their perceptual appearance) differed, as did the context of their illumination. When a classifier was trained to distinguish between responses to these two stimuli, it could rely on two types of information: perceived

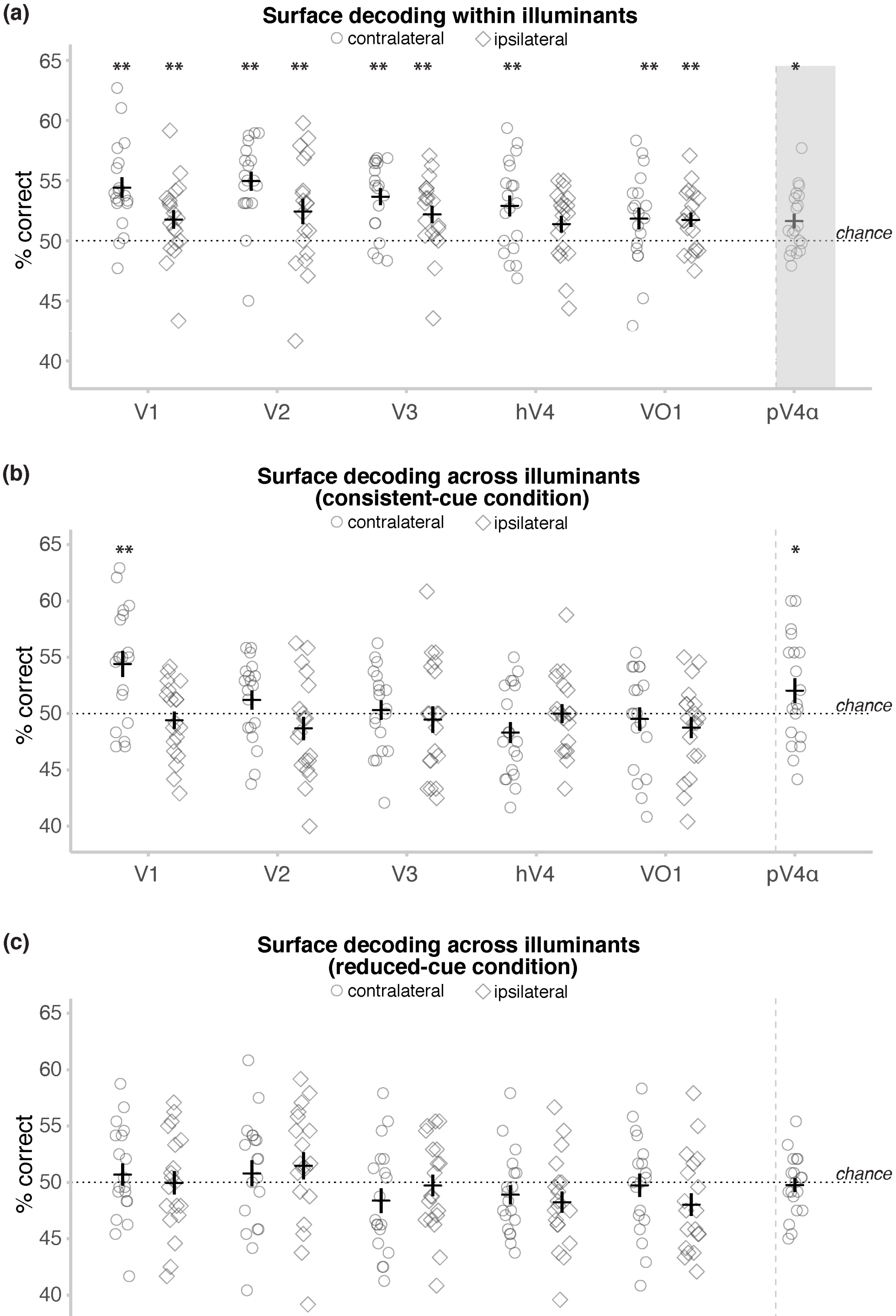

Figure 6. **Pattern Classification Results: ROI Analyses.** Mean decoding accuracies in all ROIs for surface color discrimination (the three analyses shown in Figure 3). **(a)** Surface color could be decoded significantly above chance in all ROIs (except ipsilateral hV4) when classification did not involve generalization across illuminant conditions. This panel presents pV4α in light gray as this analysis (within-illuminant decoding) was the same used to define this ROI. Note that the other panels show independent analyses. **(b)** When classifiers were trained on one set of illuminants and tested on an illuminant not present in the training set, classification was significant only in contralateral V1 and pV4α. **(c)** In the reduced-cue condition generalization across illuminant condition was not better than chance in any ROI. $*p < .05$, $**p < .005$, 10^3 permutations, FWE corrected, error bars: SEM.

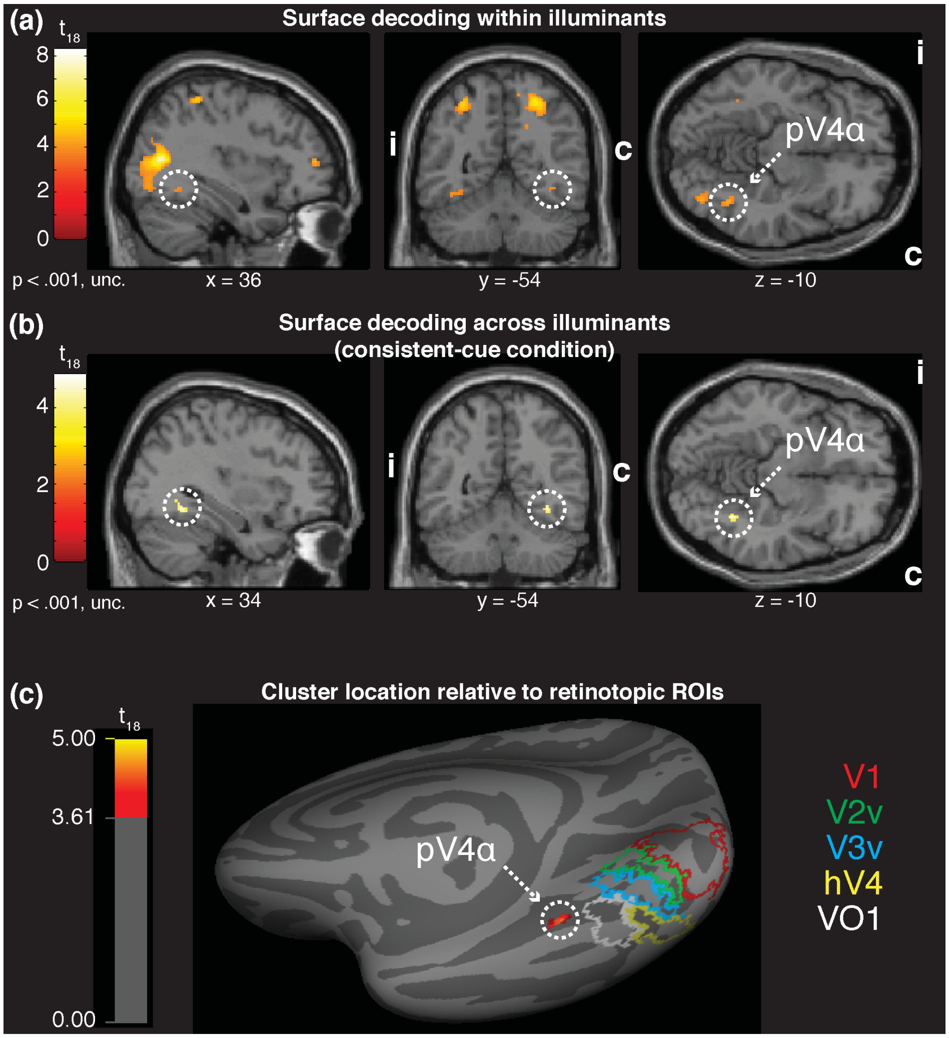

Figure 7. **Pattern Classification Results: Searchlight Analyses.** Searchlight maps (3 voxel radius) for the classification analyses shown in Figure 3 and Figure 5. Letters "i" and "c" denote ipsilateral and contralateral hemispheres with respect to location of most stimuli. **(a)** Surface classification within illuminants. Circle marks the cluster defined as putative V4α using within-illuminant decoding. **(b)** Surface classification across illuminants. The searchlight map reveals a cluster coinciding with pV4α. (c) Illustration of relative location of the ventral regions. Medial posterior view of the pV4α cluster identified in the searchlight analysis (dotted circle) and retinotopic ROIs overlaid on a cortical surface rendering in MNI space. The pV4α cluster was located anteriorly to the ROIs. Colored labels denote surface area falling into individually defined ROIs in at least 25 % of the participants.

color (yellow vs. blue surface), or illumination (yellow vs. blue). Which of the two dominated, could then easily be tested by cross-testing this classifier on the two unambiguous stimuli, i.e., blue surface under blue illumination (SbIb) and the yellow surface under yellow illumination (SyIy). If the classifier learned to rely primarily on illumination, it would classify e.g. SbIb as SyIb. However, if it relied primarily on surface color, it would classify SbIb as SbIy.

As can be seen in Figure 9*a*, all regions with the exception of pV4α showed an illuminant bias significantly different from 50 % (two-tailed). Areas V1-V3 showed the strongest illuminant bias (weakest in V3: 40.9 %, p = .002, FWE corrected), followed by areas hV4 and VO1 (weakest in hV4: 46.0 %, p = .01, FWE corrected). Only pV4α did not show an illuminant bias (50.9 %, p = .942, FWE corrected). Since there is no absolute baseline for this bias analysis, we tested for bias differences between ROIs using a repeated measures ANOVA with ROI as fixed and subjects as random factors: differences across all ROIs were significant ($F_{5,90}$ = 11.126, p = 2.51 x 10^{-7}, Greenhouse-Geisser corrected for non-sphericity ϵ_{GG} = 0.8381). Post-hoc contrasts with separate error terms showed that putative V4α indeed had a stronger

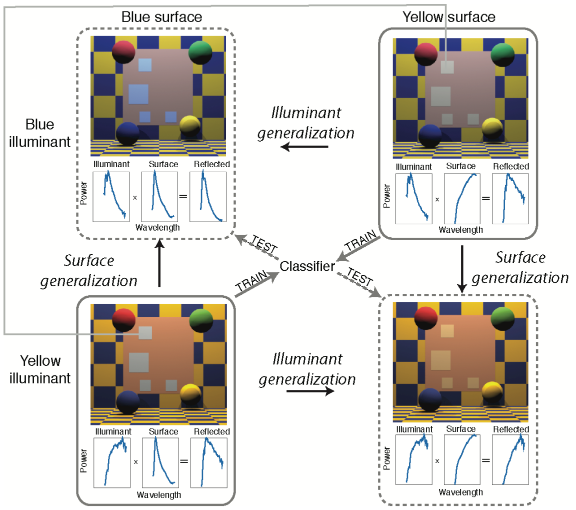

Figure 8. **Surface/Illuminant Bias Analysis: Procedure.** Analysis determining bias towards encoding of surface color versus illuminant, respectively. Illustration of stimuli and analysis used for the surface vs. illuminant bias analysis. Stimuli are identical to those shown in Figure 1*b,c*. Importantly, the yellow surface under blue illumination reflected the same light as the blue surface under yellow illumination, even though they were perceived differently (indicated by gray connector between surfaces). A first analysis trained classifiers to distinguish between these two stimuli. These classifiers could hence rely on perceptual surface color or on illumination. Classifiers were tested on responses to the other two stimuli, which revealed which of the two features the classifiers relied on. If stimuli in the same row were assigned the same label, this demonstrated a bias for illuminant representation. If stimuli in the same column were assigned the same label, this indicated a bias for surface representation. In a second and third analysis classifiers were trained and tested to distinguish between illumination (i.e. top versus bottom row) or between surface color (i.e. left vs. right column).

surface bias than the remaining ROIs (exceeding Roy-Bose critical value, $F_{5,14}$ = 5.064, p = .0148, Holm-Bonferroni corrected) and that hV4 and VO1 had stronger surface biases than V1-V3 (exceeding Roy-Bose critical value, $F_{5,14}$ = 3.149, p = .0412, Holm-Bonferroni corrected).

Since in this analysis 50% could mean "no bias", but equally well "chance level", we performed two control analyses using the data from the same four conditions (SbIb, SyIb, SbIy, SyIy). In the first analysis we trained and tested classifiers explicitly using leave-one-run-out cross-validation to distinguish between surfaces (SbIb and SbIy vs. SyIb and SyIy) while in the second analysis discriminations had to be made between illuminants (SbIy and SyIy vs. SbIb and SyIb). As can be seen in Figure 9*b*, such classifiers could successfully predict illumination and surface color from almost all ROIs (lowest decoding accuracy for surface decoding in pV4α: 53.5 %, p = .021, FWE corrected). Only illuminant decoding was not significantly above chance in this region (51.1 %, p = .653, FWE corrected). The fact that all ROIs exhibited significant decoding in surface and/or illuminant color decoding shows that none of them represented only noise.

Lastly, we calculated representational dissimilarity matrices (RDMs) to compare the similarity structures reflected in the neural responses within our ROIs with the similarity structures predicted for hypothetical brain regions that exclusively represent illuminant or surface color. While response properties in areas V1-V3 were closest to the illuminant model, pV4α resembled more the surface model while areas hV4 and VO1 were in between the two (Figure 9*c*). Dendrograms on response patterns provide further, descriptive evidence that, while all other ROIs emphasized the difference between illuminants, activity patterns in putative pV4α were clustered primarily according to surface color.

These results demonstrate that there is an increasing gradient from areas V1-V3 to hV4 and VO1 and finally to putative pV4α in preferentially encoding surface as opposed to illuminant color.

Predicting behavioral color constancy indices from fMRI activity

The Equivalent Illuminant Model proposes a simple two-stage procedure that explains surface color appearance in typical psychophysical experiments: the perceptual system first estimates the chromaticity of the illuminant (the "equivalent illuminant") and uses this estimate in a second step to calculate surface reflectance (Brainard & Maloney, 2011). Accurate estimates entail better behavioral color constancy. A direct and testable prediction would hence be that the degree to which patterns of fMRI responses to the three different illuminants can be decoded predicts the individual behavioral color constancy score.

We trained classifiers to discriminate between activity patterns elicited by the three different illuminants and cross-validated them by leaving out every run once for testing only. We performed this analysis for the consistent-cue and the reduced-cue conditions, respectively, yielding two decoding accuracies per participant. The decreases in decoding accuracies in each ROI between the two conditions entered simple linear regression models to predict the behavioral decrease in color constancy scores between the two conditions. As Figure 10 shows, there was a positive correlation in hV4 (r = .414, p = .0436, uncorrected). The correlation coefficients for the remaining ROIs are shown in Table 1. Although the effect did not survive correction for multiple comparisons, we report the result for hV4 due to its notable effect size.

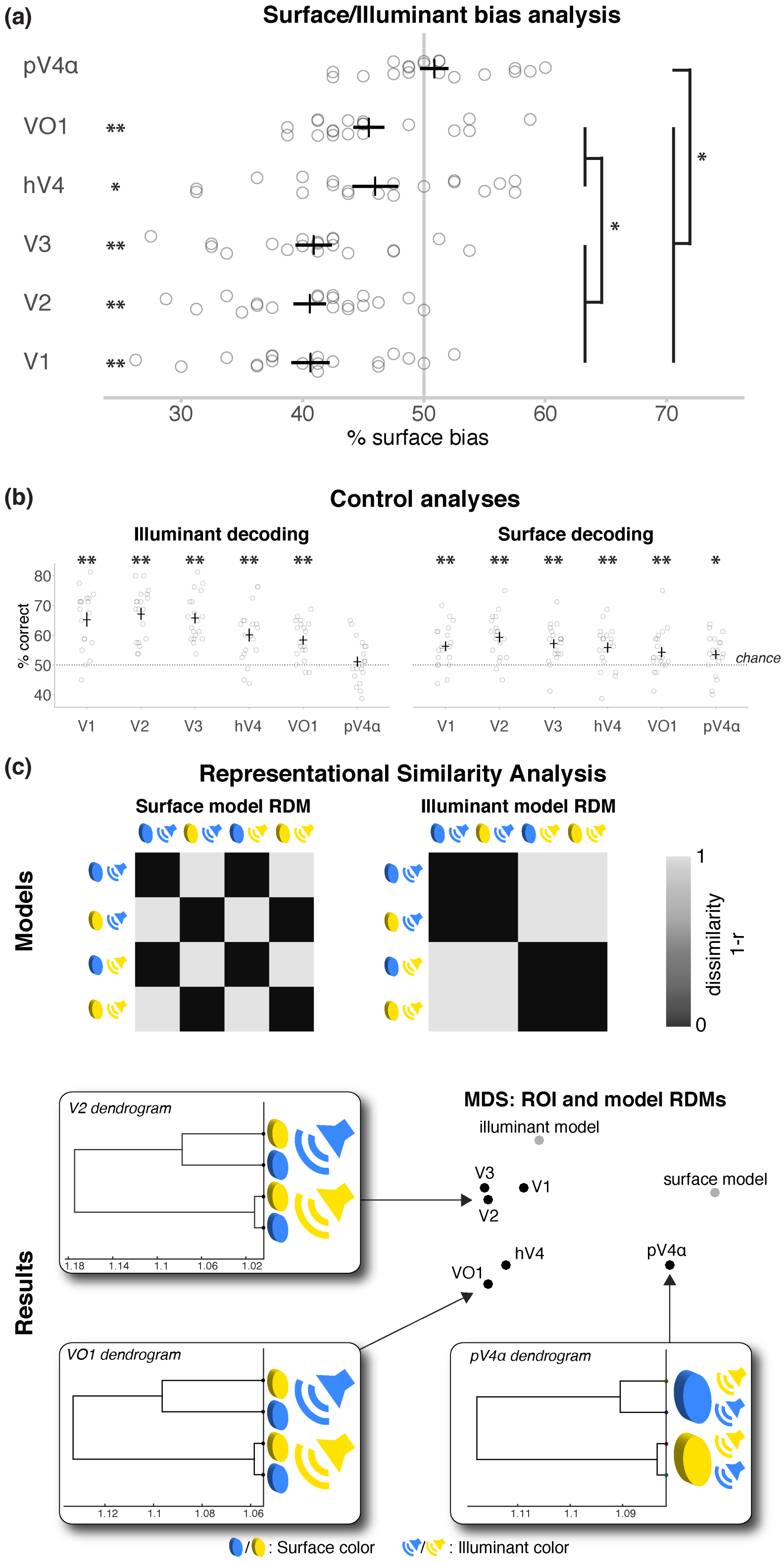

Figure 9. **Surface/Illuminant Bias Analysis: Results.** **(a)** High percentages indicate surface encoding bias while lower percentages indicate illuminant encoding bias. ROI results are tested for a significant deviation from 50 % (10^3 permutations, two-tailed). Post-hoc tests showed that pV4α had a relatively stronger surface bias than retinotopic ROIs and that hV4 and VO1 had stronger surface biases than areas V1-V3. **(b)** Mean decoding accuracies (error bars represent SEMs) for illuminant and surface color classifications. Accuracies were above chance for both analyses in all ROIs. **(c)** MDS solution in representational similarity analysis (minimizing metric stress): of all regions examined, areas V1-V3 were most similar to the illuminant model, followed by hV4 and VO1, and lastly pV4α, which was more similar to the surface model. Dendrograms show that, in contrast to the other ROIs, the grouping of the four stimuli in pV4α emphasized the difference between surfaces over illuminants.

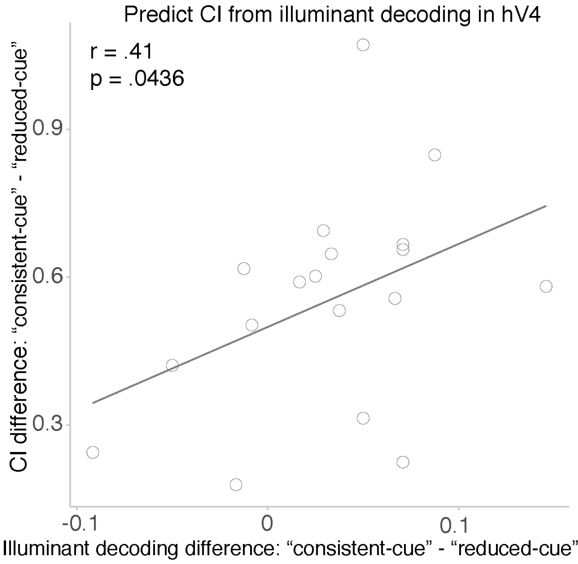

Figure 10. **Predicting Behavioral Color Constancy change due to cue conflict manipulation from neural decoding of illuminant in hV4.** Each dot represents a change in color constancy index and the corresponding change in illuminant decoding accuracy between consistent-cue or reduced-cue conditions. Larger differences in illuminant decoding were accompanied by larger differences in color constancy indices.

ROI	Pearson's r	p (uncorrected)
V1	-.01	.5081
V2	.14	.2946
V3	.06	.4134
hV4	**.41**	**.0436***
VO1	.13	.3040
V4α	-.12	.6853

Table 1

Predicting color constancy from illuminant decoding. *Correlations between changes in illuminant decoding and color constancy indices in consistent-cue and reduced-cue conditions.*

DISCUSSION

The present study is the first, to our knowledge, to directly investigate brain activity that explicitly represents surface color that is invariant with respect to illuminant changes, and to relate behavioral color constancy to neural codes of illumination. We found that the earliest cortical stage, V1, as well as one of the most anterior color-responsive regions in the fusiform gyrus, pV4α, encode color invariantly with respect to the illuminant. We also found that there is a gradient from early cortex to anterior fusiform regions to increasingly encode surface color rather than illuminant. Finally, we demonstrate that illuminant encoding in hV4 predicted the strength of the effect of a cue manipulation on behavioral color constancy, as predicted by the Equivalent Illuminant Model.

Color constancy computations in V1

We found that surface color could be decoded from fMRI activity in all visual areas when illumination did not change between training and test set. V1 and pV4α were the only regions where activity encoded surface color such that the information content generalized to new illuminants. The involvement of V1 in color constancy computations fits well with the observation that neurons in this area flexibly adjust their firing behavior to account for chromatic and achromatic changes in the illumination context well outside their receptive fields (RF) (MacEvoy & Paradiso, 2001; Wachtler et al., 2003). Other authors have emphasized the importance of double-opponent cells for color constancy (Conway & Livingstone, 2006; Friedman, Zhou, & von der Heydt, 2003; Johnson et al., 2001; Shapley & Hawken, 2011): these neurons detect chromatic contrast or color gradients of surfaces, which remain relatively constant across illuminant changes. In keeping with the perceptual relevance of chromatic contrast for surface perception, a recent study found strong edge enhancement effects for chromatically defined surfaces (Zweig, Zurawel, Shapley, & Slovin, 2015). Similarly, fMRI activity in V1 has been shown to reflect color appearance in perceptual filling-in (Hsieh & Tse, 2010). The two proposed mechanisms are not mutually exclusive and may contribute in a complementary way to color constancy computations in V1 (Hurlbert, 2003). Whatever the underlying mechanism for color constancy in primary visual cortex is, even if it is based on feedback, dysfunction of V1 is known to abolish color constancy judgments in patient D.B. although he could still discriminate between stimuli based on their spectral composition (Kentridge et al., 2007).

Invariance of surface color representations in V4 across illuminations

We did not find invariant surface color representations in retinotopically mapped areas hV4 or VO1, but more anterior, in pV4α. These null findings of course do not prove that information is not represented in those areas because differences in decoding accuracies may simply reflect differences in how such information is represented (e.g., chromatic representations in blobs in V1 versus thin stripes in V2), which may in turn influence the "sampling bias" in MVPA (Bartels, Logothetis, & Moutoussis, 2008). Although differences exist between the physiology of chromatic processing in human and nonhuman primate brains (Lafer-Sousa, Conway, & Kanwisher, 2016; Wade et al., 2008), the present findings are in conflict with previous observations that neurons in monkey V4 encode surface color and are robust against changes in wavelength composition (Kusunoki et al., 2006; Zeki, 1983), in particular given the sensitivity of MVPA to surface color in most ROIs in the within-illuminant classification analysis.

Our results may be explained by the fact that neural activity in V4 is strongly influenced by attention to color as demonstrated in both monkey electrophysiology (Maunsell & Treue, 2006; McAdams & Maunsell, 2000; Motter, 1994) and human fMRI studies (Bartels & Zeki, 2000; Brouwer & Heeger, 2013; Saenz, Buracas, & Boynton, 2002). We did not instruct our participants to specifically pay attention to the color of the surfaces in the scene. Alternatively, it is conceivable that V1 and V4 represent chromatic gradients and contrasts at different spatial scales due to differences in RF sizes (for a similar interpretation of Zeki's seminal findings (Zeki, 1983), see Maunsell and Newsome (1987)). Our complete stimulus image (16.8° x 15°), for instance, was considerably smaller than the stimuli used by Kusunoki et al. (2006) (30°). In contrast, Wachtler et al. (2003) examined contextual modulation of chromatic processing at distances of up to only 6° of visual angle from the RF the size of which ranged from 2.5° to 4.5°. In fact, attention may be implemented in form of RF tuning, as has been reported for V4 neurons (David, Hayden, Mazer, & Gallant, 2008; Klein, Harvey, & Dumoulin, 2014; Moran & Desimone, 1985). We should point out, however, given that RFs presumably are even larger in V4α than V4, a pure RF size account cannot fully explain the disparate findings in these two regions.

Our findings are in accord with evidence for the involvement of anterior fusiform gyrus, which includes V4 and V4α, in achromatopsia (Bouvier & Engel, 2005), with clinical observations being a key reason for the traditional view of this region to play a crucial role in color vision and color constancy (Zeki, 1990). The fact that electrical stimulation in V4α elicits color percepts in a human patient underscores the relevance of this area for color vision in general (Murphey, Yoshor, & Beauchamp, 2008).

A potential role of feedback in V1 signal

Finally, given the abundance of cortical feedback to V1 (Felleman & Van Essen, 1991; Muckli & Petro, 2013), it is also conceivable that the information we decode from V1 actually reflects feedback from higher visual areas, possibly V4 or V4α. Prior studies have shown that BOLD signal is particularly susceptible to feedback (Haynes, Deichmann, & Rees, 2005; O'Connor, Fukui, Pinsk, & Kastner, 2002; Wunderlich, Schneider, & Kastner, 2005), which is most likely due to its strong correlation with postsynaptic neural input (Bartels et al., 2008; Logothetis, 2008). Similarly, numerous fMRI studies that have found decoding or signal modulation specifically in V1, but not in V2, found this pattern of result in situations where the signal must be due to feedback: for memory color (Bannert & Bartels, 2013), size illusions induced by distance (Sperandio, Chouinard, & Goodale, 2012), shape perception (Murray, Kersten, Olshausen, Schrater, & Woods, 2002), and context effects in scene perception (Smith & Muckli, 2010). The illuminant invariant surface signals decoded here in V1 may hence also be the result of a complex interplay between V1, hV4 and especially V4α that encoded illuminant invariant surface color. The significant partial correlation between activity in V1 and V4α is consistent with this idea (Figure S2).

Functional gradient for surface vs. illuminant color representation from V1-VO1

When two differently colored surfaces reflect the same light, they can be discriminated perceptually either on the basis of their reflective properties or based on differences in their illumination. Our surface/illuminant bias analysis tapped into this crucial mechanism of color constancy. The results showed that the tendency to discriminate between illuminant color as opposed to surface color decreased along a gradient from V1 to V4α (Figure 9*a*). The propensity of higher visual areas (in particular hV4, VO1, V4α relative to earlier areas) to interpret the difference between stimuli as being between surface color (despite matched wavelength composition) is consistent with V4's role in figure-ground segmentation and surface perception (Bouvier, Cardinal, & Engel, 2008; Cox et al., 2013; Poort et al., 2012; Roe et al., 2012) and resembles gradients found for the perception of illusory contours (Mendola, Dale, Fischl, Liu, & Tootell, 1999) and chromatically defined figure/ground segmentation (Seymour et al., 2015).

A model for color constancy computations in visual cortex

A remarkable finding in the present study was the correlation of neural discriminability between distinct illuminants with the change in behavioral color constancy indices caused by our cue conflict manipulation. To our knowledge this is the first empirical evidence at the level of neural encoding for the Equivalent Illuminant Model, which links correct illuminant estimation with the ability to estimate surface reflectance (i.e. color) (Brainard & Maloney, 2011). In accord with this, we found that when the difference in neural illuminant decoding between cue conditions was large, the decrease of behavioral color constancy indices was also strong.

Conclusion

The present study adds an important new piece to the puzzle of human color vision. Experimental approaches seeking to discover isomorphic mappings between perceptual and neural color spaces have found area V4 to be involved in color perception (Brouwer & Heeger, 2009; Li et al., 2014). In the present study we examined two central components of color constancy in the human brain, namely the robustness of neural encoding of surface reflectance during changes in illumination, and the neural encoding of the illuminant itself. We found that the only regions robustly encoding surface color during varying illumination conditions were the primary visual cortex and a region in anterior ventral cortex previously implied in color vision, pV4α. Careful stimulus design allowed us to examine for each region whether it was biased in encoding of surface color or the illuminant. This was achieved by choosing distinct pairs of surface reflectance and illumination that resulted in matched reflected light. In such ambiguous situations, there was a gradient from early to higher ventral regions to preferentially encode surface reflectance relative to illumination. Finally, we found evidence suggesting a correlation between perceptual color constancy and neural encoding of the illuminant, as proposed by the equivalent illuminant model, in area hV4.

ACKNOWLEDGMENTS

This work was supported by a grant from the German Federal Ministry for Education and Research (BMBF), grant number: FKZ 01GQ1002, by the German Excellence Initiative of the German Research Foundation (DFG) grant number EXC307, and by the Max Planck Society, Germany. The authors declare no competing financial interests.

References

Bannert, M. M. & Bartels, A. (2013). Decoding the yellow of a gray banana. *Current Biology*, *23*, 2268–72.

Barbur, J. L. & Spang, K. (2008). Colour constancy and conscious perception of changes of illuminant. *Neuropsychologia*, *46*, 853–63.

Bartels, A., Logothetis, N. K., & Moutoussis, K. (2008). fMRI and its interpretations: an illustration on directional selectivity in area V5/MT. *Trends in Neurosciences*, *31*, 444–53.

Bartels, A. & Zeki, S. (2000). The architecture of the colour centre in the human visual brain: new results and a review. *European Journal of Neuroscience*, *12*, 172–93.

Beauchamp, M. S., Haxby, J. V., Jennings, J. E., & DeYoe, E. A. (1999). An fMRI version of the Farnsworth-Munsell 100-Hue test reveals multiple color-selective areas in human ventral occipitotemporal cortex. *Cerebral Cortex*, *9*, 257–63.

Blair, R. C. & Karniski, W. (1993). An alternative method for significance testing of waveform difference potentials.

Bouvier, S. E., Cardinal, K. S., & Engel, S. A. (2008). Activity in visual area V4 correlates with surface perception. *Journal of Vision*, *8*, 28.1–9.

Bouvier, S. E. & Engel, S. A. (2005). Behavioral Deficits and Cortical Damage Loci in Cerebral Achromatopsia. *Cerebral Cortex*, *16*, 183–191.

Brainard, D. H. (1998). Color constancy in the nearly natural image. 2. Achromatic loci. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, *15*, 307–25.

Brainard, D. H. & Maloney, L. T. (2011). Surface color perception and equivalent illumination models. *Journal of Vision*, *11*, 1–18.

Brooks, J. (2012). Counterbalancing for serial order carryover effects in experimental condition orders. *Psychological Methods*, 1–54.

Brouwer, G. J. & Heeger, D. J. (2009). Decoding and Reconstructing Color from Responses in Human Visual Cortex. *Journal of Neuroscience*, *29*, 13992–14003.

Brouwer, G. J. & Heeger, D. J. (2013). Categorical Clustering of the Neural Representation of Color. *Journal of Neuroscience*, *33*, 15454–15465.

Clarke, S., Walsh, V., Schoppig, A., Assal, G., & Cowey, A. (1998). Colour constancy impairments in patients with lesions of the prestriate cortex. *Experimental Brain Research*, *123*, 154–8.

Conway, B. R. & Livingstone, M. S. (2006). Spatial and temporal properties of cone signals in alert macaque primary visual cortex. *Journal of Neuroscience*, *26*, 10826–46.

Conway, B. R. & Tsao, D. Y. (2009). Color-tuned neurons are spatially clustered according to color preference within alert macaque posterior inferior temporal cortex. *Proceedings of the National Academy of Sciences*, *106*, 18034–9.

Cox, M. A., Schmid, M. C., Peters, A. J., Saunders, R. C., Leopold, D. A., & Maier, A. (2013). Receptive field focus of visual area V4 neurons determines responses to illusory surfaces. *Proceedings of the National Academy of Sciences*, *110*, 17095–100.

David, S. V., Hayden, B. Y., Mazer, J. a., & Gallant, J. L. (2008). Attention to stimulus features shifts spectral tuning of V4 neurons during natural vision. *Neuron*, *59*, 509–21.

Delahunt, P. B. & Brainard, D. H. (2004). Does human color constancy incorporate the statistical regularity of natural daylight? *Journal of Vision*, *4*, 57–81.

Engel, S., Zhang, X., & Wandell, B. A. (1997). Colour tuning in human visual cortex measured with functional magnetic resonance imaging. *Nature*, *388*, 68–71.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, *9*, 1871–1874.

Felleman, D. J. & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*, 1–47.

Friedman, H. S., Zhou, H., & von der Heydt, R. (2003). The coding of uniform colour figures in monkey visual cortex. *Journal of Physiology*, *548*, 593–613.

Gegenfurtner, K. R., Bloj, M., & Toscani, M. (2015). The many colours of ‘the dress’. *Current Biology*, *25*, R543–R544.

Haynes, J.-D., Deichmann, R., & Rees, G. (2005). Eye-specific effects of binocular rivalry in the human lateral geniculate nucleus. *Nature*, *438*, 496–499.

Heasly, B. S., Cottaris, N. P., Lichtman, D. P., Xiao, B., & Brainard, D. H. (2014). RenderToolbox3: MATLAB tools that facilitate physically based stimulus rendering for vision research. *Journal of Vision*, *14*, 1–22.

Hsieh, P.-J. & Tse, P. U. (2010). "Brain-reading" of perceived colors reveals a feature mixing mechanism underlying perceptual filling-in in cortical area V1. *Human Brain Mapping*, *1407*, 1395–1407.

Hurlbert, A. (2003). Colour Vision: Primary Visual Cortex Shows Its Influence. *Current Biology*, *13*, R270–R272.

Ishihara, S. (2011). *Ishihara’s tests for colour deficiency* (38 Plates). Tokyo, Japan: Kanehara Trading Inc.

Johnson, E. N., Hawken, M. J., & Shapley, R. (2001). The spatial transformation of color in the primary visual cortex of the macaque monkey. *Nature Neuroscience*, *4*, 409–16.

Johnson, E. N., Hawken, M. J., & Shapley, R. (2008). The Orientation Selectivity of Color-Responsive Neurons in Macaque V1. *Journal of Neuroscience*, *28*, 8096–8106.

Kennard, C., Lawden, M., Morland, a. B., & Ruddock, K. H. (1995). Colour identification and colour constancy are impaired in a patient with incomplete achromatopsia associated with prestriate cortical lesions. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *260*, 169–75.

Kentridge, R. W., Heywood, C. A., & Weiskrantz, L. (2007). Color contrast processing in human striate cortex. *Proceedings of the National Academy of Sciences*, *104*, 15129–31.

Klein, B. P., Harvey, B. M., & Dumoulin, S. O. (2014). Attraction of position preference by spatial attention throughout human visual cortex. *Neuron*, *84*, 227–37.

Kleiner, M., Brainard, D. H., & Pelli, D. G. (2007). "What’s new in Psychtoolbox-3?"

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, *103*, 3863–8.

Kusunoki, M., Moutoussis, K., & Zeki, S. (2006). Effect of background colors on the tuning of color-selective cells in monkey area V4. *Journal of Neurophysiology*, *95*, 3047–59.

Lafer-Sousa, R., Conway, B. R., & Kanwisher, N. G. (2016). Color-Biased Regions of the Ventral Visual Pathway Lie between Face- and Place-Selective Regions in Humans, as in Macaques. *Journal of Neuroscience*, *36*, 1682–97.

Land, E. H. & McCann, J. J. (1971). Lightness and retinex theory. *Journal of the Optical Society of America*, *61*, 1–11.

Li, M., Liu, F., Juusola, M., & Tang, S. (2014). Perceptual Color Map in Macaque Visual Area V4. *Journal of Neuroscience*, *34*, 202–17.

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature*, *453*, 869–78.

Lueck, C. J., Zeki, S., Friston, K. J., Deiber, M. P., Cope, P., Cunningham, V. J., … Frackowiak, R. S. (1989). The colour centre in the cerebral cortex of man. *Nature*, *340*, 386–9.

MacEvoy, S. P. & Paradiso, M. A. (2001). Lightness constancy in primary visual cortex. *Proceedings of the National Academy of Sciences*, *98*, 8827–8831.

Maloney, L. T. (1999). Physics-based approaches to modeling surface color perception. In K. R. Gegenfurtner & L. T. Sharpe (Eds.), *Color vision: from genes to perception* (pp. 387–416). Cambridge, UK: Cambridge University Press.

Maunsell, J. H. R. & Newsome, W. T. (1987). Visual processing in monkey extrastriate cortex. *Annual Review of Neuroscience*, *10*, 363–401.

Maunsell, J. H. R. & Treue, S. (2006). Feature-based attention in visual cortex. *Trends in Neurosciences*, *29*, 317–22.

McAdams, C. J. & Maunsell, J. H. R. (2000). Attention to both space and feature modulates neuronal responses in macaque area V4. *Journal of Neurophysiology*, *83*, 1751–5.

Mendola, J. D., Dale, A. M., Fischl, B., Liu, A. K., & Tootell, R. B. (1999). The representation of illusory and real contours in human cortical visual areas revealed by functional magnetic resonance imaging. *Journal of Neuroscience*, *19*, 8560–72.

Michael, C. R. (1978). Color vision mechanisms in monkey striate cortex: dual-opponent cells with concentric receptive fields. *Journal of Neurophysiology*, *41*, 572–88.

Mollon, J. D. (1989). "Tho’ she kneel’d in that place where they grew..." The uses and origins of primate colour vision. *The Journal of Experimental Biology*, *146*, 21–38.

Moran, J. & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science*, *229*, 782–784.

Motter, B. C. (1994). Neural correlates of attentive selection for color or luminance in extrastriate area V4. *Journal of Neuroscience*, *14*, 2178–2189.

Muckli, L. & Petro, L. S. (2013). Network interactions: non-geniculate input to V1. *Current Opinion in Neurobiology*, *23*, 195–201.

Murphey, D. K., Yoshor, D., & Beauchamp, M. S. (2008). Perception matches selectivity in the human anterior color center. *Current Biology*, *18*, 216–20.

Murray, S. O., Kersten, D., Olshausen, B. A., Schrater, P., & Woods, D. L. (2002). Shape perception reduces activity in human primary visual cortex. *Proceedings of the National Academy of Sciences*, *99*, 15164–9.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Computational Biology*, *10*, e1003553.

O'Connor, D. H., Fukui, M. M., Pinsk, M. A., & Kastner, S. (2002). Attention modulates responses in the human lateral geniculate nucleus. *Nature Neuroscience*, *5*, 1203–9.

Poort, J., Raudies, F., Wannig, A., Lamme, V. A. F., Neumann, H., & Roelfsema, P. R. (2012). The role of attention in figure-ground segregation in areas V1 and V4 of the visual cortex. *Neuron*, *75*, 143–156.

Radonjic, A., Cottaris, N. P., & Brainard, D. H. (2015). Color constancy supports cross-illumination color selection. *Journal of Vision*, *15*, 1–19.

Roe, A. W., Chelazzi, L., Connor, C. E., Conway, B. R., Fujita, I., Gallant, J. L., … Vanduffel, W. (2012). Toward a unified theory of visual area V4. *Neuron*, *74*, 12–29.

Rüttiger, L., Braun, D. I., Gegenfurtner, K. R., Petersen, D., Schönle, P., & Sharpe, L. T. (1999). Selective color constancy deficits after circumscribed unilateral brain lesions. *Journal of Neuroscience*, *19*, 3094–106.

Saenz, M., Buracas, G. T., & Boynton, G. M. (2002). Global effects of feature-based attention in human visual cortex. *Nature Neuroscience*, *5*, 631–632.

Sereno, M. I., Dale, A. M., Reppas, J. B., Kwong, K. K., Belliveau, J. W., Brady, T. J., … Tootell, R. B. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science*, *268*, 889–93.

Seymour, K. J., Clifford, C. W. G., Logothetis, N. K., & Bartels, A. (2009). The coding of color, motion, and their conjunction in the human visual cortex. *Current Biology*, *19*, 177–83.

Seymour, K. J., Clifford, C. W. G., Logothetis, N. K., & Bartels, A. (2010). Coding and binding of color and form in visual cortex. *Cerebral Cortex*, *20*, 1946–54.

Seymour, K. J., Williams, M. A., & Rich, A. N. (2015). The Representation of Color across the Human Visual Cortex: Distinguishing Chromatic Signals Contributing to Object Form Versus Surface Color. *Cerebral Cortex*, 1–9.

Shapley, R. & Hawken, M. J. (2011). Color in the cortex: single- and double-opponent cells. *Vision Research*, *51*, 701–17.

Smith, F. W. & Muckli, L. (2010). Nonstimulated early visual areas carry information about surrounding context. *Proceedings of the National Academy of Sciences*, *107*, 20099–103.

Sperandio, I., Chouinard, P. A., & Goodale, M. A. (2012). Retinotopic activity in V1 reflects the perceived and not the retinal size of an afterimage. *Nature Neuroscience*, *15*, 540–2.

Wachtler, T., Sejnowski, T. J., & Albright, T. D. (2003). Representation of color stimuli in awake macaque primary visual cortex. *Neuron*, *37*, 681–691.

Wade, A. R., Augath, M., Logothetis, N. K., & Wandell, B. A. (2008). fMRI measurements of color in macaque and human. *Journal of Vision*, *8*, 6.1–19.

Wild, H. M., Butler, S. R., Carden, D., & Kulikowski, J. J. (1985). Primate cortical area V4 important for colour constancy but not wavelength discrimination. *Nature*, *313*, 133–135.

Winawer, J., Horiguchi, H., Sayres, R. A., Amano, K., & Wandell, B. A. (2010). Mapping hV4 and ventral occipital cortex: The venous eclipse. *Journal of Vision*, *10*, 1–1.

Wunderlich, K., Schneider, K. A., & Kastner, S. (2005). Neural correlates of binocular rivalry in the human lateral geniculate nucleus. *Nature Neuroscience*, *8*, 1595–1602. arXiv: NIHMS150003

Xiao, B., Hurst, B., MacIntyre, L., & Brainard, D. H. (2012). The color constancy of three-dimensional objects. *Journal of Vision*, *12*, 1–15.

Zeki, S. (1983). Colour coding in the cerebral cortex: the reaction of cells in monkey visual cortex to wavelengths and colours. *Neuroscience*, *9*, 741–65.

Zeki, S. (1990). A century of cerebral achromatopsia. *Brain*, *113*, 1721–77.

Zweig, S., Zurawel, G., Shapley, R., & Slovin, H. (2015). Representation of Color Surfaces in V1: Edge Enhancement and Unfilled Holes. *Journal of Neuroscience*, *35*, 12103–12115.

Supplementary Information
"Invariance of Surface Color Representations Across Illuminant Changes in the Human Cortex"

Michael M. Bannert and Andreas Bartels
Werner Reichardt Centre for Integrative Neuroscience, Eberhard Karls University
Bernstein Center for Computational Neuroscience
Max Planck Institute for Biological Cybernetics
Department of Psychology, Eberhard Karls University
Tübingen

Keywords: Color constancy, fMRI, V1, V4, surface perception

Calculating Color Constancy Indices

We used the Stockman Sharpe 10° cone fundamentals (Stockman & Sharpe, 2000) for all calculations below and followed the approach that is detailed on pages 312–313 and Appendix A of Brainard (1998). The index quantifying color constancy for a shift from illumination 1 to 2 is given by:

$$CI = 1 - ||\mathbf{c_2} - \mathbf{c_d}||/||\mathbf{c_2} - \mathbf{c_1}|| \qquad (1)$$

Since CI is asymmetric, we always calculated CI for a shift from illumination 1 to 2 and vice versa and averaged both values. ||x|| is the L^2 norm of vector x. $\mathbf{c_1}$ and $\mathbf{c_2}$ are the u'v' chromaticities of illumination 1 and illumination 2. To calculate these, we plugged in the SPDs that we used in our rendering simulation and convolved them with the cone fundamentals. We then converted the resulting cone activations $\mathbf{e_1}$ and $\mathbf{e_2}$ to XYZ space (CIE 1931) and converted them to u'v' coordinates (Psychtoolbox function XYZTouv.m).

$$diag(\mathbf{D_0}) = \mathbf{a_2}./\mathbf{a_1} \qquad (2)$$

("./" means element-wise division) from the achromatic settings $\mathbf{a_1}$ and $\mathbf{a_2}$ (cone coordinates) participants made under the two illuminations. This matrix describes how, for a given observer, cone excitations for an achromatic setting under illuminant 1 are predicted to change to correspond to an achromatic match under illuminant 2.

This matrix $\mathbf{D_0}$ can be used to predict the "equivalent illuminant", i.e., the cone coordinates for an illuminant 2 that would match the chromaticity of the actual chromaticity of illuminant 1 in appearance (for the given observer):

$$\mathbf{e_d} = \mathbf{D_0} * \mathbf{e_1} \qquad (3)$$

The cone coordinates ed are then also converted to XYZ space to obtain chromaticities cd (as described above). If the observer is perfectly color constant, then $\mathbf{c_d} = \mathbf{c_2}$. The color constancy index would then be 1. An observer without color constancy, however, would have a diagonal matrix with only ones on the diagonal. Color constancy would then be 0. Importantly, the matrix $\mathbf{D_0}$ does not depend on the physical chromaticity of the achromatic setting that an observer made. The change in the achromatic settings determines $\mathbf{D_0}$.

Wavelength-based decoding across cue conditions

Our ROI analysis of surface color decoding across illuminants in the consistent-cue condition showed that activity patterns generalized across illuminants in V1 and V4α only. This implies that V1 and V4α encoded signal related to the color-constant appearance of the surfaces, even though their physical properties changed under changing illuminants. To control for the possibility that the absence of evidence for above-chance decoding in hV4 and VO1 could be explained by poor signal quality in those areas, we devised a pattern classification analysis for which one would expect the opposite results: successful decoding in regions that encode physical properties of the surfaces, but worse decoding in regions solely encoding appearance of the surfaces.

This test would be achieved by training classifiers using neural responses on consistent-cue conditions, and testing on those to reduced-cue ones, and vice versa. The rationale is the following: the physical properties of surfaces were identical in consistent-cue and reduced-cue conditions, yet their appearance differed, as color constancy broke down in the reduced-cue condition. This is the opposite scenario to decoding across illuminants, where appearance remained constant while physical properties changed.

This means that in our supplemental analysis on decoding across consistent-cue and reduced-cue conditions, color constant neural representations are now expected to show lower decoding accuracies than wavelength-based ones. This is analogous to the findings from patient D.B. (Kentridge, Heywood, & Weiskrantz, 2007) who only distinguished colors on their spectral basis in the impaired visual field whereas in the unaffected visual field behavioral matches reflected contextual modulation by the chromatic surround (as in healthy controls).

We performed this test to check if the null finding in hV4 can be explained by poor signal in that area. This is not the case because this wavelength-based decoding analysis works better in hV4 than in V1 and pV4α. This finding means that the wavelength-based representation of color in hV4 is more pronounced than in V1 and pV4α. Fig. S1 shows the results from this analysis and the comparisons between ROIs.

Connectivity analysis between ROIs

Since V1 and putative V4α had similar response properties in our ROI classification analyses, we checked if these regions also showed a correlation between their time series across the whole experiment. Such correlations could indicate functional connectivity between regions. We obtained the voxel time series by fitting a new GLM with SPM12 to the entire experimental data and regressed out only the motion parameters, the global mean signal (Desjardins, Kiehl, & Liddle, 2001), and the activity induced by the presence of the visual scene (as opposed to the color mask) convolved with a canonical HRF. We used SPM12 because it offers the possibility to save the residual time series after model estimation. Residual time series were averaged per ROI and entered into pairwise partial correlation analyses between all ROIs. We used partial correlation because we were interested in detecting variance that is uniquely shared between every two regions, which cannot be explained by the common influence of another ROI. Significance was assessed with two-tailed t tests (18 degrees of freedom) and we corrected for multiplicity using the Holm-Bonferroni method.

Fig. S2 shows that neighboring ROIs had positive partial correlations between their mean time series. However, the only significant negative partial correlation coefficient was found between V1 and putative V4α, which clearly are not located in close proximity to each other. This may indicate that the two regions are part of a functional network in this task.

Decoding based on mean activation levels

In order to check to what extent the decoding accuracies that we observed were driven by potential differences in overall activation levels between the responses to the blue and yellow surfaces, we repeated the analyses shown in main Fig. 4b (across-illuminant decoding using the mean signal strength across voxels. This means that each pattern of responses was now replaced with its arithmetic mean. The classification algorithm thus had to learn a discrimination boundary separating the (scalar) mean activations associated with blue and yellow surfaces, respectively. In all other respects the analysis was performed in the same way as in the analysis of across-illuminant decoding. Fig. S3 shows the results for the classification of the mean values in direct comparison with the original classification results. Classification accuracies did not significantly exceed chance in any of the regions tested ($p \geq .468$, one-tailed permutation test, 10^3 permutations, FWE corrected). This means that differences in mean activation levels cannot account for the constancy effects that we observed in our main analysis.

References

Brainard, D. H. (1998). Color constancy in the nearly natural image. 2. Achromatic loci. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, *15*, 307–25.

Desjardins, A. E., Kiehl, K. A., & Liddle, P. F. (2001). Removal of confounding effects of global signal in functional MRI analyses. *NeuroImage*, *13*, 751–758.

Kentridge, R. W., Heywood, C. A., & Weiskrantz, L. (2007). Color contrast processing in human striate cortex. *Proceedings of the National Academy of Sciences*, *104*, 15129–31.

Stockman, A. & Sharpe, L. T. (2000). The spectral sensitivities of the middle- and long-wavelength-sensitive cones derived from measurements in observers of known genotype. *Vision Research*, *40*, 1711–1737.

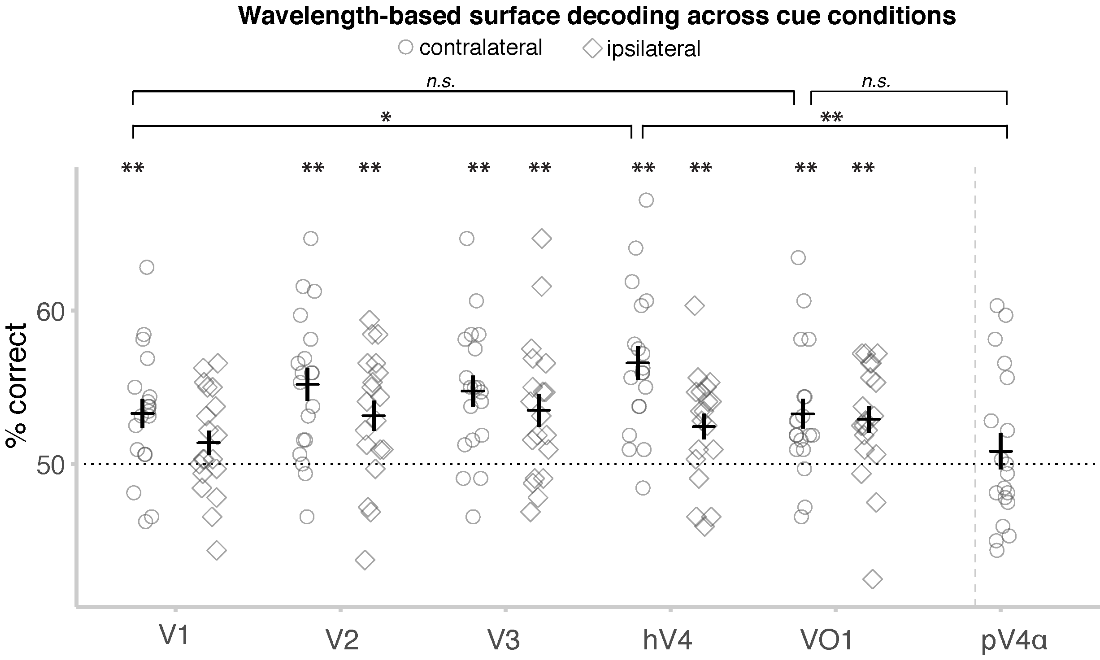

Figure 1. **Decoding across cue conditions.** Classifiers were trained to distinguish between surface colors in the consistent-cue condition and tested on samples from the reduced-cue condition (and vice versa). Classifications were conducted separately for the blue and yellow illuminant conditions and averaged. Classification accuracy was significantly above chance in all ROIs except ipsilateral V1 and putative V4α (see Methods in main article for statistical inference procedure applied in ROI analyses). Comparisons with one-tailed t tests showed that the classification accuracy was larger in contralateral hV4 than V1 and putative V4α (Bonferroni corrected). The comparisons with contralateral VO1 were not significant (see Results in main text for more detail). *p < .05, **p < .01.

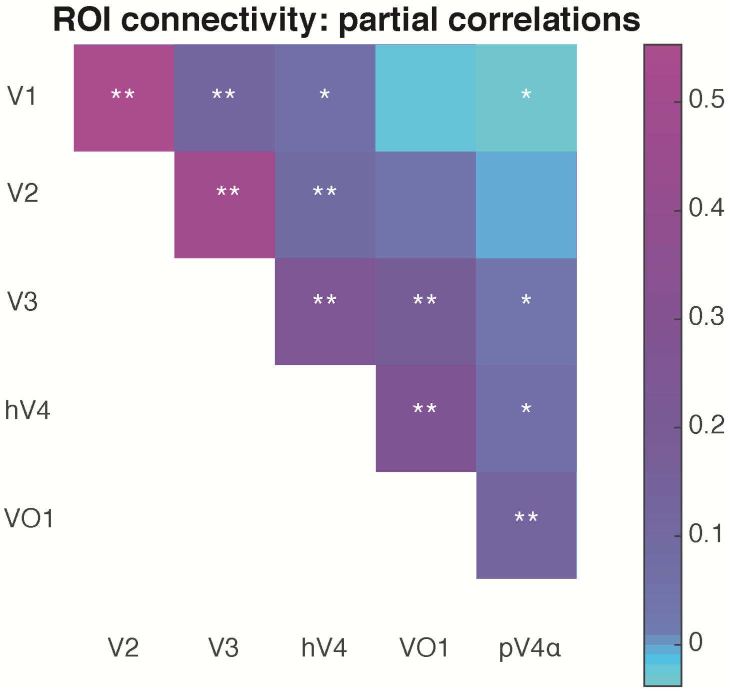

Figure 2. **Connectivity analysis between ROIs.** Pairwise correlations between all ROIs after partialling out all variance attributable to all remaining ROIs. Significant partial correlations are found between regions that are relatively close to each other and a significant long-range correlation is found between V1 and putative V4α. *p < .05, **p < .01, df = 18, Holm-Bonferroni corrected.

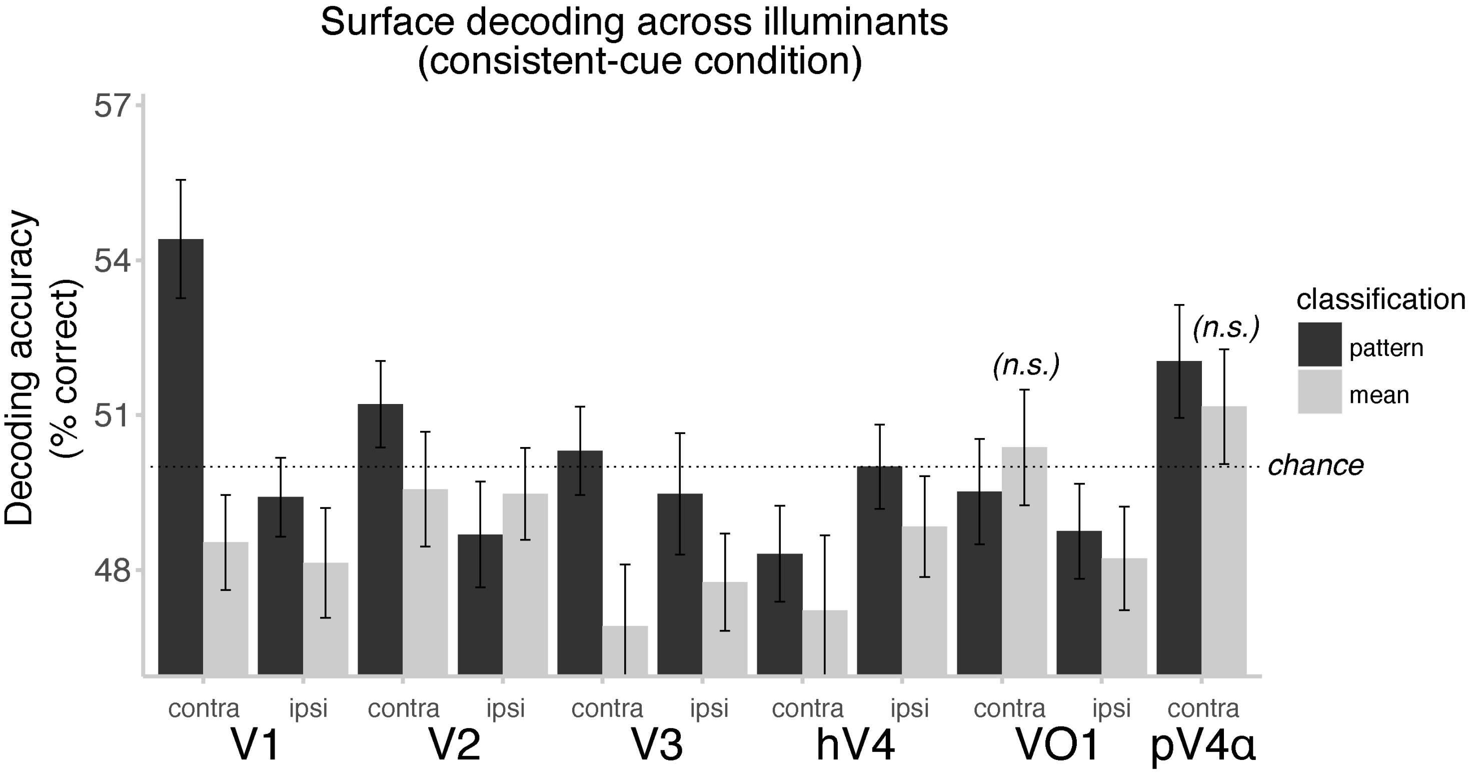

Figure 3. **Decoding based on mean activation levels.** The results from across-illuminant decoding in the consistent-cue condition (main Figure 4b) in comparison with the same analysis in which each pattern was replaced with the mean of the pattern. Results show that decoding based on mean activation was not significantly above chance in any region of interest (means across participants, error bars are sem).

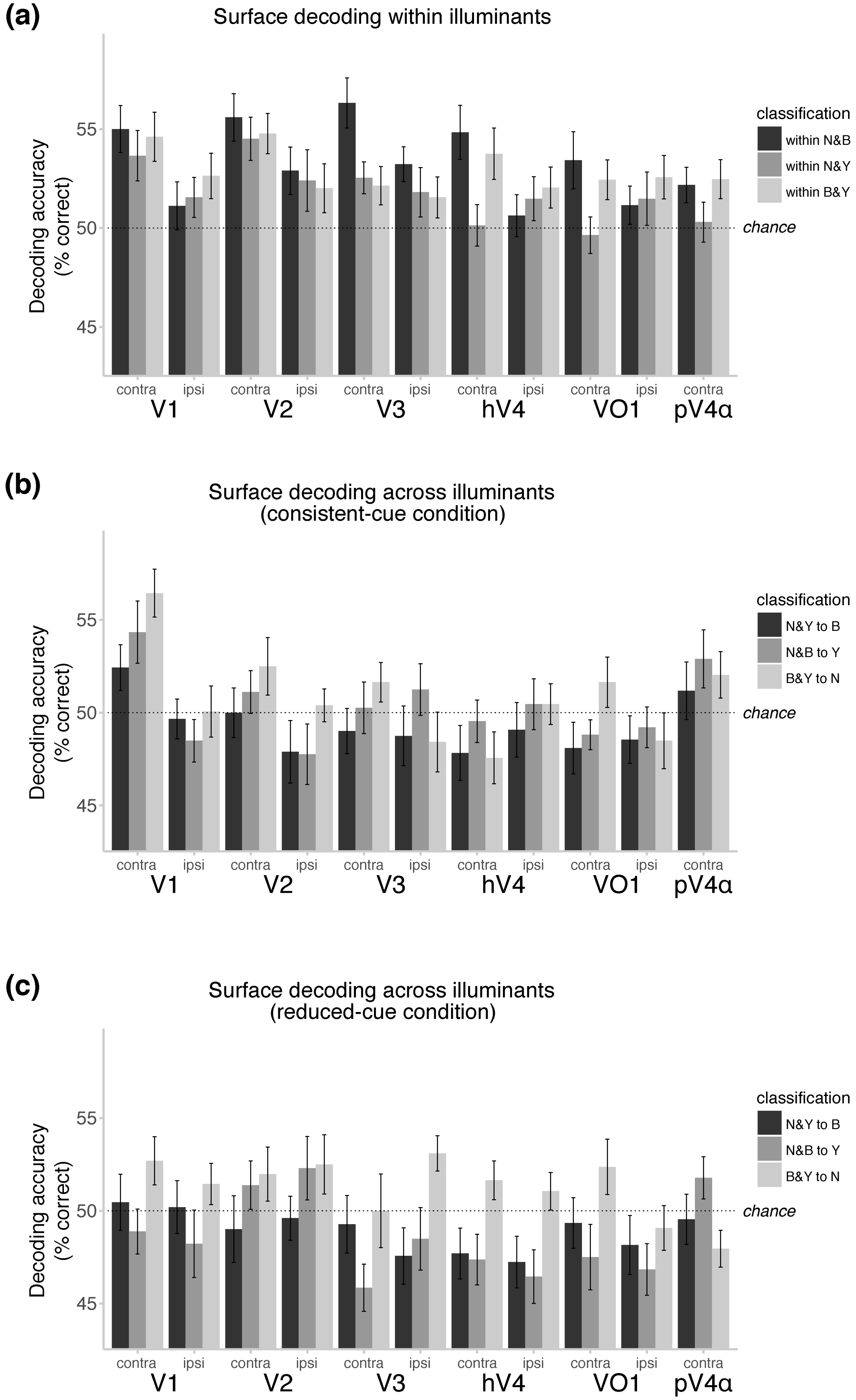

Figure 4. **Prediction accuracies for individual classification analyses.** Percentage of correctly predicted surface color labels in the three sub-classifications averaged across participants (error bars represent sem). Same data as in main Figure 6.

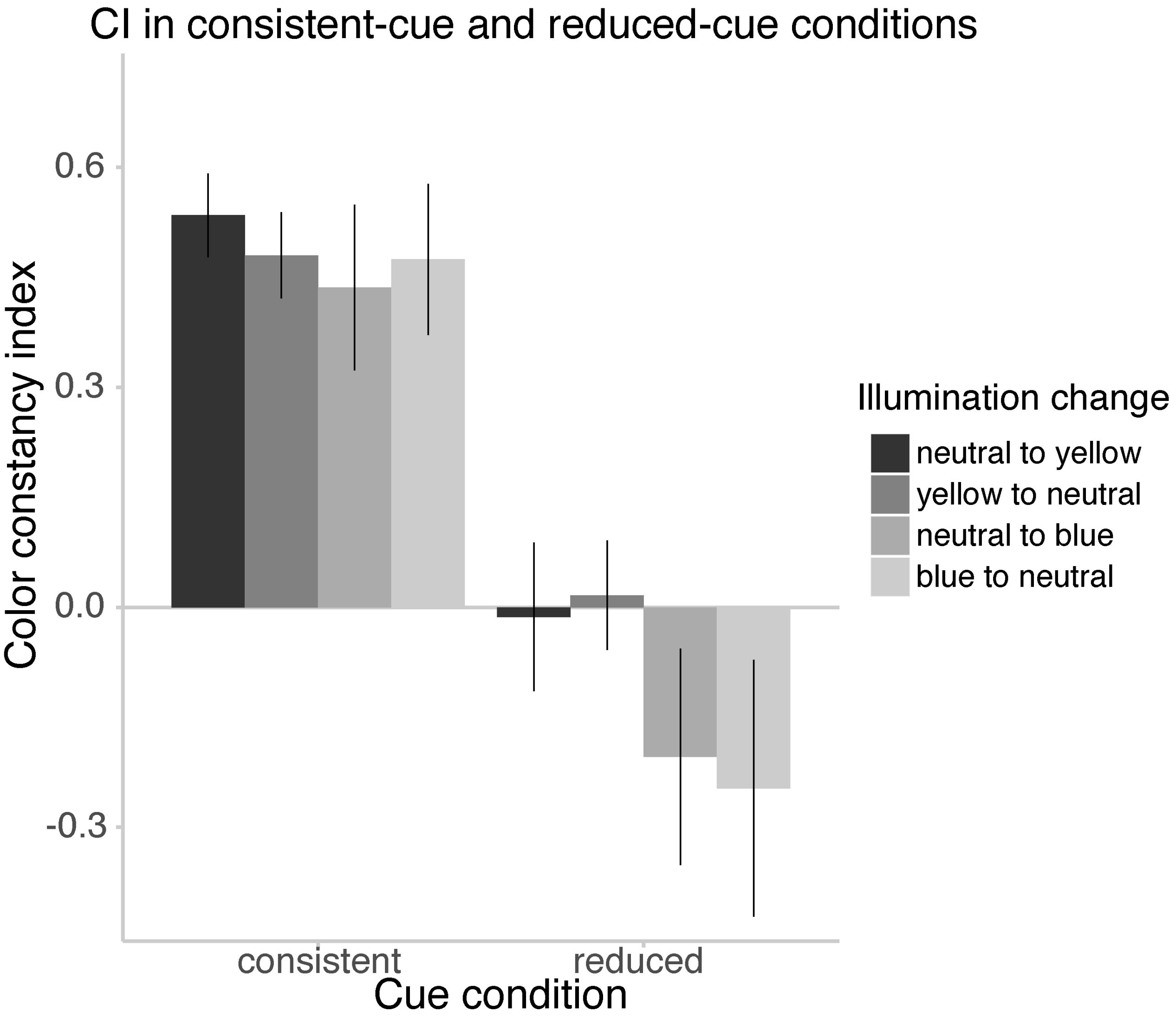

Figure 5. **Color constancy indices by illumination transitions.** Mean color constancy indices in the two cue conditions grouped by illumination conditions (error bars are sem). Same data as in main Figure 5.

6 Experiment III: Human V4 Encodes Object Color in Visual Imagery

Author contributions: AB and MMB designed research, MMB collected and analyzed data, AB and MMB interpreted results and wrote the manuscript.

Human V4 Encodes Object Color in Visual Imagery

Michael M. Bannert and Andreas Bartels
Werner Reichardt Centre for Integrative Neuroscience, Eberhard Karls University
Bernstein Center for Computational Neuroscience
Max Planck Institute for Biological Cybernetics
Department of Psychology, Eberhard Karls University
Tübingen

Among the multitude of elements making up visual experience, color stands out in that it can specify subjective experience and objective properties of the outside world. Whereas most neuroimaging research on human color vision has focused on external stimulation, the present study addressed this duality by investigating how externally elicited color vision is linked to subjective color experience induced by object imagery. We recorded fMRI activity while showing our participants abstract color stimuli (in "void" mode) that were either red, green, or yellow in half of the runs and asked them to produce mental images of colored objects corresponding to the same three categories in the remaining half. Although seen color could be predicted from all retinotopically mapped visual areas, only color decoders trained on responses in hV4 could also predict the color category of an object that was being imagined. Using a brain-based behavioral modeling approach, we demonstrated that the neural signal in hV4 was predictive of performance in the color judgment task on a trial-by-trial basis. The commonality between neural representations of perceived and imagined object color, in combination with the behavioral modeling evidence, hence identifies area hV4 as a "perceptual bridge" linking externally triggered color vision with color in self-generated object imagery.

Keywords: Color vision, object imagery, fMRI, pattern classification, drift diffusion modeling

INTRODUCTION

Color is a ubiquitous feature of our visual environment. This is because there is a large variability in the reflective properties of objects surrounding us, helping us perceive, search for, and recognize objects (Gegenfurtner & Rieger, 2000; Mollon, 1989). Not only can we perceive color effortlessly and automatically under most normal viewing conditions (Brainard & Maloney, 2011; Foster, 2011; Shevell & Kingdom, 2008), but also our memories, dreams, and thoughts can feature color just as naturally.

While a lot has been learned from human neuroimaging about the neural mechanisms underlying the perception of color (Bartels & Zeki, 2000; Beauchamp, Haxby, Jennings, & DeYoe, 1999; Brouwer & Heeger, 2009, 2013; Kuriki, Sun, Ueno, Tanaka, & Cheng, 2015; Lueck et al., 1989; Parkes, Marsman, Oxley, Goulermas, & Wuerger, 2009), it is not clear how such neural representations of perceived color relate to the color of imagined objects although it is known that color imagery can strongly influence color perception (Chang, Lewis, & Pearson, 2013; Wantz, Mast, & Lobmaier, 2015).

Previous fMRI research on object imagery showed that imagining and perceiving various classes of objects lead to similar patterns of blood-oxygen-level dependent (BOLD) activity in visual cortex (Cichy, Heinzle, & Haynes, 2012; M. Lee & Wagenmakers, 2012; Reddy, Tsuchiya, & Serre, 2010). If these effects are mediated by a neural correspondence at the level of object semantics or low-level features, and if so, which features, cannot be decided based on these results. One experiment demonstrated that imagery for complex artwork is encoded in the same low-level feature representations (e.g., orientation, spatial frequency) as perceiving the pictures (Naselaris, Olman, Stansbury, Ugurbil, & Gallant, 2015). However, this study examined neural responses in V1 and V2 only and did not analyze color responses. Given the specialization for color processing in the brain (Conway, Moeller, & Tsao, 2007; Livingstone & Hubel, 1984; Schiller, Logothetis, & Charles, 1990; Xiao, Wang, & Felleman, 2003; Zeki & Stutters, 2013), it now becomes obvious to examine the role of low-level features also in these specialized systems during imagery of natural objects.

Early imaging experiments probed color imagery by instructing observers to make color judgments about known objects, which led to increased BOLD activity either in V4 (Rich et al., 2006) or anterior to it (Howard et al., 1998). This approach leaves open the question if these BOLD responses show selectivity for the type of object color. Moreover it does not address the question if object colors and perceived color are encoded in visual cortex in a similar way.

Investigation of related phenomena involving internally generated color experience such as synesthesia yielded mixed results as well (for review, see Rouw, Scholte, & Colizoli, 2011): while some experiments found area V4 to be activated by both synesthetic experience and color perception (Hubbard, Arman, Ramachandran, & Boynton, 2005; Nunn et al., 2002), others did not (Gould van Praag, Garfinkel, Ward, Bor, & Seth, 2015; Rich et al., 2006). The one study that tested for color selectivity failed to find repetition suppression in visual cortex (van Leeuwen, Petersson, & Hagoort, 2010). The neural correlate of the exact phenomenological content in synesthetic color experience therefore remains elusive.

Other examples of nonretinal vision (Pearson & Kosslyn, 2015) includes working memory, which, as some have argued, can be mediated by the same neurocognitive mechanism as imagery (Albers, Kok, Toni, Dijkerman, & de Lange, 2013; Keogh & Pearson, 2011; Pearson & Kosslyn, 2015). Although working memory paradigms justifiably are not designed to address the question if object imagery rests on low-level color representations, their results suggest that, even in the absence of bottom-up sensory stimulation, information about low-level stimulus properties (e.g., orientation) can be selectively encoded in the patterns of BOLD activity in V1 as well as extrastriate visual areas (Harrison & Tong, 2009; S.-H. Lee & Baker, 2016). It is, however, questionable if this can be generalized to color since their short-term memory representations could only be decoded in V1 (Serences, Ester, Vogel, & Awh, 2009).

Memory has also been shown to induce color experience for achromatic images of color-diagnostic objects (Hansen, Olkkonen, Walter, & Gegenfurtner, 2006; Olkkonen, Hansen, & Gegenfurtner, 2008; Witzel, Valkova, Hansen, & Gegenfurtner, 2011). Using pattern classifiers, neuroimaging studies have found that the neural responses to grayscale images of color-diagnostic objects are similar to the neural responses to veridical color (Bannert & Bartels, 2013; Vandenbroucke, Fahrenfort, Meuwese, Scholte, & Lamme, 2016). Such effects of prior object knowledge on color perception cannot be generalized to object imagery without restrictions, as, unlike object imagery, they likely depend on "implicit" top-down processing (Albright, 2012).

Findings from related fields of nonretinal color vision are thus quite heterogeneous and none of them address the role of low-level color representations in object imagery. We therefore recorded fMRI activity while participants viewed either abstract color stimuli that did not convey object information or visually imagined objects. Classifiers trained to discriminate between BOLD activity patterns in hV4 elicited by veridical color perception were able to predict the color of the imagined objects. Decoding accuracy was correlated with behavioral performance in a 1-back color judgment task on a trial-by-trial basis, thereby highlighting the functional relevance of the decoded color representations in hV4.

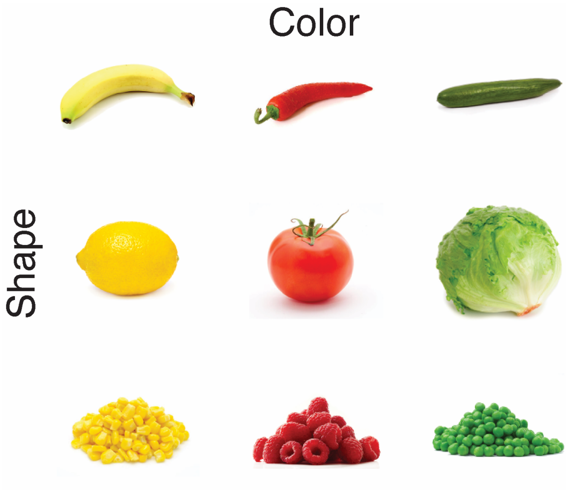

Figure 1. **Stimulus material.** Stimuli used in the imagery task. Each object belonged to one of three color categories (yellow, red, green). To reduce confounds unrelated to object color, objects were approximately matched in shape (longish, round, pile-shaped) and semantic associations (all of them were fruits/vegetables). Prior to scanning and before each imagery fMRI run, participants practiced to remember the images of 9 natural objects. The task required them to identify the correct stimulus, which they had to imagine ("is if they could see it on screen") in the imagery run, from a selection that included 3 additional distractor items (not shown) from the same basic level category. The experiment did not proceed before the participant completed the task without error.

MATERIALS AND METHODS

Participants

19 volunteers from the Tübingen University community participated in the experiment. They provided written informed consent prior to the first experimental session. One participant failed to complete the experiment. We thus analyzed data from the remaining volunteers (N = 18, 3 males, age between 22 – 35 years, mean = 25.8 years). Participants had normal color vision as measured with Ishihara plates (Ishihara, 2011).

Stimuli

In the beginning of the experiment participants were familiarized with the images of the 9 objects used in the imagery condition (Figure 1). Each belonged to one of three color categories (red, green, and yellow). We chose three objects per color that were matched in overall shape across categories: per category, one object was longish, another round, and third a third one “pile-shaped”. Furthermore all objects belonged to the same superordinate level category (“food”) to minimize semantic confounds. The colored rings in perception runs had mean chromaticities of x = .39, y = .35 in the red, x = .34, y = .41 in the green, and x = .41, y = .43 in the yellow conditions, respectively. Colored rings were shown at high and low luminance. Luminance values were determined using the minimal flicker procedure (Kaiser, 1991) which required participants to adjust the luminance of a color stimuli presented against achromatic backgrounds at high (184.9 cd/m^2) and low (151.3 cd/m^2) intensities until the amount of perceived flicker was minimal.

The colors determined in this procedure were used to construct the abstract color rings used in the perception runs of the fMRI experiment. They consisted of several concentric rings created by displaying a colored disc that had its alpha channel sinusoidally modulated as a function of eccentricity (1° visual angle cycle size). The rings drifted outwards at a velocity of $2.47°/s$. We varied the stimulus intensity within color categories to ensure that classifiers relied on chromaticity differences to distinguish color conditions while treating the luminance difference as noise.

The cue words in the imagery phase were presented in black letters in the center of the screen. Cue words were presented in German or English, depending on the participant’s preference. All stimuli were presented with Psychtoolbox 3 (Kleiner, Brainard, & Pelli, 2007).

fMRI experiment

Participants viewed the stimuli while lying supine in a scanner via a mirror fixed to the head coil. A projector (NEC PE401H) displayed the stimuli on a screen placed at the end of the scanner tube. The display was gamma-calibrated using a Photo Research PR-670 spectroradiometer (CalibrateMonSpd.m function from Psychtoolbox). Display size was 21.8° and 16.2° of visual angle along the horizontal and vertical direction, respectively, at a resolution of 800 x 600 pixels.

The fMRI experiment consisted of 6 perception and 6 imagery runs (Figure 2), which participants underwent in alternation (counterbalanced across participants). In perception runs participants viewed ring-shaped abstract color stimuli in blocks of 8.5 *s* and responded with a button press whenever the color changed luminance for a brief period of time (0.3 *s*). Participants received feedback about their task performance at the end of each block for motivation. Before the start of imagery runs, participants performed a task to memorize the objects they had to visualize during the measurement. This study task required them to distinguish the correct object from among three distractor objects of the same basic level category. For the imagery task, we told participants to imagine the objects as if they were seeing them on the screen. The participants had to correctly identify each of the nine objects at least once before the fMRI experiment was resumed. In imagery runs, participants fixated on a small circle in the middle of the screen. Cue words appeared at the beginning of each imagery block for 1.5 *s* indicating which of the nine objects to mentally visualize in the subsequent imagery block. The imagery block lasted 11.714 *s* and a new cue word of the subsequent trial appeared. Upon the appearance of a new cue word, participants performed a 1-back color judgment task, which required them to indicate if the object referred to by the new cue and the object they had been imaging previously were the same color. We instructed our participants to reply as quickly and accurately as possible. They had to make a decision while the cue word was still on the screen. At the end of each run they received feedback about reaction times and errors for motivation.

Conditions in perception and imagery runs were presented in a pseudo-randomized sequence that ensured that each condition was preceded by every other condition an equal number of times (Brooks, 2012). The first trial from each run was discarded from the analysis and only served to keep the back matched sequence of conditions intact across alternations between perception and imagery runs. Perception runs thus consisted of 36 trials and imagery runs consisted of 27 trials for a total of 216 perception trials and 162 imagery trials that entered the analysis.

Before the start of the scan, we familiarized the participants with the object images to ensure that they could correctly identify each object based on the word cue. They practiced the study task, which they performed later in the scanner before each imagery run.

Retinotopic mapping & ROI definition

Each volunteer participated in a retinotopic mapping session to identify visual areas V1, V2, V3, hV4, VO1, LO1, and LO2. We chose these areas because they had previously been shown to be involved in the processing of shape and color (e.g., Brouwer & Heeger, 2009; Larsson & Heeger, 2006; Seymour, Clifford, Logothetis, & Bartels, 2010). We used standard retinotopic mapping procedures to identify reversals in the angle map of the visual field representations that delineate the boundaries between these areas on the cortical surface (Sereno, McDonald, & Allman, 1994; Wandell & Winawer, 2011). Participants viewed a contrast-reversing checkerboard through a wedge-shaped aperture. Since the visual field representations are compressed at large eccentricities on the cortical surface (cortical magnification), the check sizes increased logarithmically with eccentricity. The aperture subtended the entire screen within a 90° angle at all eccentricities from the fixation dot. The wedge rotated at a period of 55.64 *s* for a total of 10 cycles per run. The rotation direction alternated between the four mapping runs.

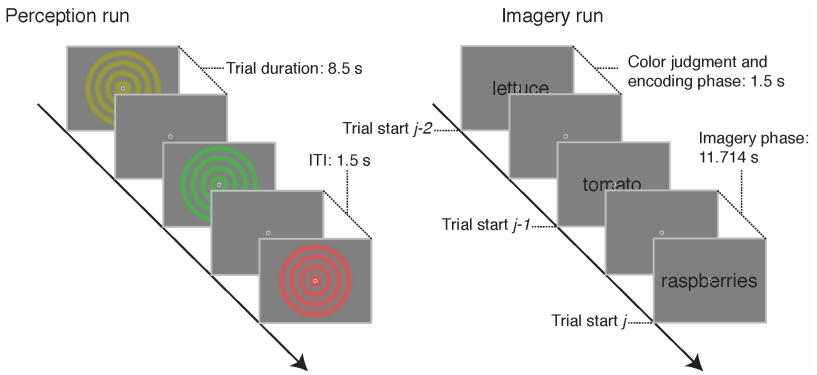

Figure 2. **Experimental Design.** Trial sequence in perception and imagery runs. In perception runs (left) participants viewed concentric rings slowly drifting outward. The rings could be one of three colors: yellow, green, or red. Observers performed a detection task requiring them to press a button every time the luminance of the stimulus changed. There could be 0, 1, or 2 target events per stimulus presentation (8.5 *s*, ITI = 1.5 *s*). In imagery runs (right), participants saw a word cue at the beginning of the trial for 1.5 *s* indicating which of the nine object images they had to imagine in the subsequent imagery phase (11.714 *s*). Each time a new word cue appeared they had to decide by pressing on of two buttons whether or not the color of the object they had to imagine in this trial matched the color of the object in the previous trial (1-back same/different color judgment task).

fMRI scan details

We measured BOLD activity with a 64-channel head coil at 3 T magnetic field strength (Siemens Prisma) using 56 slices oriented axially but slightly tilted in parallel with the AC-PC line. The sampling volume covered almost the whole brain with a slice thickness of 2 *mm* and no gap between slices. In plane resolution was 96 x 96, yielding an isotropic voxel size of 2 *mm*. We used a 4-fold GRAPPA accelerated parallel imaging sequence (GRAPPA factor 2) to measure T2*-weighted functional images. Repetition time (TR) and echo time (TE) were 0.87 *s* and 30 *ms*, respectively, with a flip angle of 57°. Anatomical images with an isotropic voxel size of 1 *mm* were measured using a T1 weighted MP-RAGE sequence and magnetic field inhomogeneities were measured with a Gradient Echo field map.

fMRI data preprocessing

The first 11 functional images recorded per run were discarded to allow the MRI signal to reach equilibrium. Functional data were realigned and unwarped using the estimated field map, slice time corrected and co-registered to the anatomical image. Finally the data were normalized to MNI space using a segmentation-based normalization of the anatomical image. No smoothing was applied to the images from the main experiment. We used SPM8 (http://www.fil.ion.ucl.ac.uk/spm) for preprocessing. The data from the retinotopic mapping session underwent the same preprocessing up to co-registration in SPM8. The resulting images were then further preprocessed with Freesurfer (http://surfer.nmr.mgh.harvard.edu/), which involved smoothing them with a 4 *mm* Gaussian kernel. Individual cortical surfaces for all participants were obtained using Freesurfer's recon-all pipeline.

fMRI data analysis

The aim of this study was to examine how color of imagined objects is represented in relation to perceived color. We estimated vectors of fMRI responses using a standard GLM approach and then carried out pattern classification analyses on these data.

Our analysis strategy followed two main steps: in a first step we trained classifiers to predict which of three colors an observer was seeing using only data from color perception runs (perceived-to-perceived color classification). To verify this training procedure, we used cross-validation leaving out data from a different run in every iteration to obtain an unbiased accuracy estimate for the classifier. In a second step, we trained classifiers on responses from all perception runs but this time tested them on responses from the imagery runs (perceived-to-imagined color classification). Crucially, this analysis tested for commonalities between the representation of color perceived in abstract color rings and the color of imagined objects.

Furthermore, we wanted to know if participants actually performed the imagery task instead of merely keeping the object color in mind. We therefore carried out another classification analysis to decode the shape of the imagined object (imagined-to-imagined shape classification). Since the three shapes were roughly matched across color categories, we could thus train classifiers to predict the shape of the object ("longish", "round", or "pile-shaped") that they imagined on a given imagery trial while balancing the number of colors across shape categories. If the shape of the imagined object could be decoded, this would indicate that participants indeed activated a neural representation of shape although the judgment task did not require participants to mentally represent this stimulus dimension.

Pattern estimation. We modeled the unsmoothed voxel time series with one boxcar regressor per trial. In the perception runs, the onset times of each color block served as regressor onsets. To avoid contamination with visual processing related to the displayed cue, we chose the cue offset times as regressor onsets in imagery runs. All regressors were shifted 5 *s* forward in time to account for the hemodynamic lag. Realignment parameters regressed out the linear dependence between head motion and voxel time series. We estimated one response vector for each of the 216 perception and 162 imagery trials. Across vectors, each component was quadratically detrended by removing, in every run, the fit of a 2$^{\text{nd}}$ order polynomial from each parameter time series to filter out low-frequency noise. Each residual time series was z-scored for each run separately. To make our analysis more robust against outliers, we set all values with a difference of more than 2 standard deviations from the mean to -2 and 2, respectively.

Classification details. We employed linear discriminant analysis (LDA) classifiers for pattern classification. Due to the low number of samples and high dimensionality of the dataset we used a shrinkage estimator for the covariance matrix to ensure that it remained non-singular (Ledoit & Wolf, 2004). Additionally, we used recursive feature elimination (RFE, De Martino et al., 2008) on training data only to select the set of voxels that optimally distinguished between the categories to be classified. The optimal set of voxels was then used to fit the classifier to the entire training set and to validate it on the test set, which was not part of the voxel selection procedure. RFE determined the optimal voxel set by repeatedly training LDA classifiers on part of the training set (i.e., leaving out one run each time) and testing it on the remaining part of the training set (i.e., the withheld run) to obtain an accuracy score. This procedure was repeated 15 times while each time dropping those 15 % of the voxels from the classification whose coefficients varied the least across discriminant functions and hence were least discriminative of the category to be predicted.

Statistical inferece. We used permutation tests to evaluate the statistical significance of our classifications results. For every time we trained a classifier to discriminate between fMRI patterns, we refit new classifiers after randomly permuting the labels in the training set 10^3 times. The reasoning for this is that under the null hypothesis of no association between fMRI patterns and category membership (e.g., color category) the labels can be randomly reassigned to fMRI vectors without changing the expected classification accuracies. We used the 10^3 classification accuracies from each participant to obtain a null distribution of mean accuracies at the group level expected under the null hypothesis (including the accuracy that was actually observed using the unpermuted dataset). From these null distributions, p values for a one-tailed test can be calculated as the number of values in the distribution that exceeded the observed accuracies divided by the number of permutations. Since we examined classification accuracies from several ROIs, we needed to correct for multiple comparisons. We controlled the family-wise error (FWE) by constructing a common null distribution for all ROIs by taking the maximum value across ROI group means in each permutation step while making sure that the same label permutations were used in every ROI (Nichols & Holmes, 2002). The resulting null distribution was then used to calculate FWE corrected *p* values.

Behavioral data analysis

We instructed our participants to perform a color judgment task in order to foster the mental representation of color. This task required them to indicate if the color of the object referred to by the word cue to was identical with the color of the object they had just been imagining. We fitted reaction times (RTs) and errors using hierarchical drift diffusion models (HDDM, Wiecki, Sofer, & Frank, 2013). Drift diffusion models provide a principled way to integrate RTs and errors in a single value called a drift rate. Drift rate quantifies how quickly evidence is accumulated in a decision-making process and thus indicates how easily a task is performed. Higher values mean shorter RTs and fewer errors. An important advantage of HDDM is that it estimates model parameters by directly modeling the dependencies between model parameters across subjects in a hierarchical Bayesian framework for improved sensitivity.

We used HDDM to probe the psychophysiological relevance of the color information decoded in the perceived-to-imagined classification for the imagery task. The hypothesis was that when participants are well engaged in the imagery task, this would improve the behavioral performance as well as the signal of the neural representation of the imagined object color. It follows that behavioral performance is better when color classifiers made correct compared to incorrect predictions for imagery trials. To test this, we fit HDDMs to our behavioral data that assumed separate drift rates for correctly and incorrectly classified trials. We then used Bayesian parameter estimation to calculate the posterior probability of the drift rate being indeed higher for correct than incorrect trials. The relationship between brain signal and behavior was computed at different time points (see . In one model different drift rates were assumed depending on whether the imagery pattern immediately following the color judgment was classified correctly ("post-judgment model") while the second model assumed different drift rates for correct versus incorrect classifications of the pattern that preceded the color judgment ("pre-judgment model"). These two patterns were thought to be relevant since the task required a comparison about the color of the previous with the following objects. An additional model that checked dependence on classification two trials before the color judgment (not shown) was also included for comparison. We performed MCMC sampling to approximate the posterior distribution over model parameters given the data using ten chains. Each of them drew 10^4 samples plus additional 10^3 for burn-in. When conducting statistical inference on the basis of posterior probabilities in a family of tests, it is noteworthy to point out that, by declaring drift rates as "different" if the posterior probabilities of their difference being larger than 0 exceeded .95, we automatically enforced a corresponding upper bound of 5 % on the false discovery rate (Friston & Penny, 2003). Note that only trials with key presses made within the 1.5 *s* response window were included in this analysis, i.e., while the cue was still on the screen.

Finally, we complemented the statistical inference based on Bayesian parameter estimation with a model comparison approach: We fit a model that assumed only a single drift rate to the data and compared it with the complex models using the Deviance Information Criterion (DIC). This measure quantifies hierarchical model fit while imposing a penalty for model complexity.

RESULTS

Real color decoding

The purpose of the first classification analysis was to validate that classifiers could predict which color a person was seeing from the multivariate pattern of fMRI responses to the colored ring stimuli. llustrates that real color could be decoded from fMRI activity in all ROIs we studied, which replicates previous findings (e.g., Brouwer and Heeger 2009; Seymour et al. 2009). Classification accuracies ranged from 41.4 % in LO2 (p = .001, FWE corrected, Cohen's d = 1.2) to 51.4 % in V1 (p = .001, FWE corrected, Cohen's d = 3.38). This shows that it is possible to construct a classifier to decode perceived color.

Decoding accuracies in the remaining areas were 51.4 % (Cohen's d = 2.72) in V2, 49.3 % (Cohen's d = 2.1) in V3, 49 % (Cohen's d = 2.2) in hV4, 42.8 % (Cohen's d = 1.74) in VO1, and 43.6 % (Cohen's d = 1.76) in LO1 (all p = .001, FWE corrected).

Predicting imagined object color

Our main hypothesis was that the neural representations of imagining object color overlap with those of veridical color perception. The crucial test for this was to train classifiers to distinguish between perceived colors and to use them to predict the color of objects that participants were imagining in imagery trials. If perceived and imagined object color representations overlap, one would expect classification accuracies above chance level. As can be seen from the color of

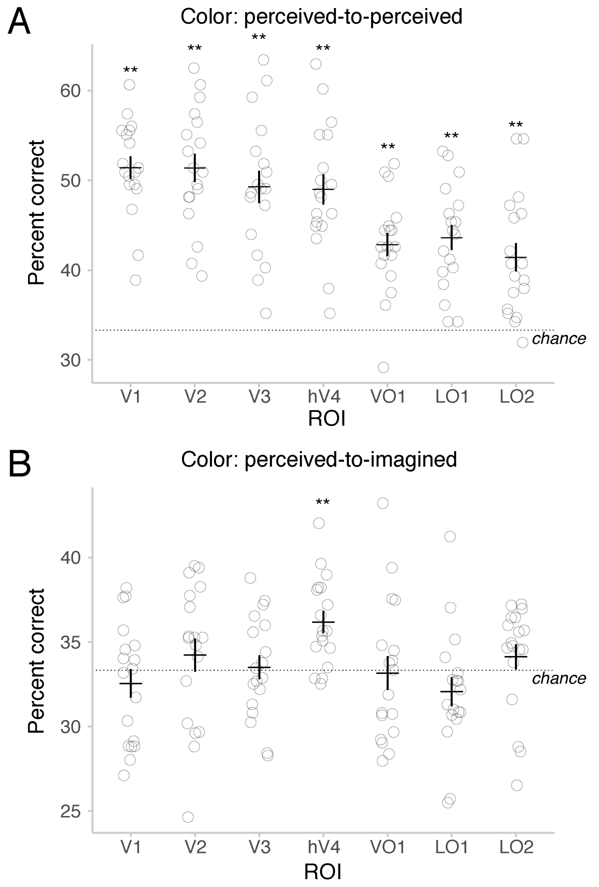

Figure 3. **Color decoding results.** Perceived-to-perceived and perceived-to-imagined color decoding. (A) Classifiers were trained to distinguish between the three color categories based on the responses in each ROI to the perceived colors. Classifier performance was cross-validated leaving out one of the six runs for testing on each iteration to obtain an average accuracy. In all ROIs classifiers could predict the color of the stimulus that participants were viewing significantly above chance. (B) Color classifiers were trained on the whole set of fMRI responses to perceived colors to predict which color observers were seeing. The learned classifiers were than used to predict on a trial-by-trial basis the color of the objects that participants were imagining in the imagery runs of the experiment. Permutation tests showed that the color of the imagined objects could be decoded significantly above chance only from activity patterns in area hV4. (A and B) Horizontal and vertical bars represent group means and SEMs, respectively. Chance level was 1/3. $^{**}p$ < .01, FWE corrected.

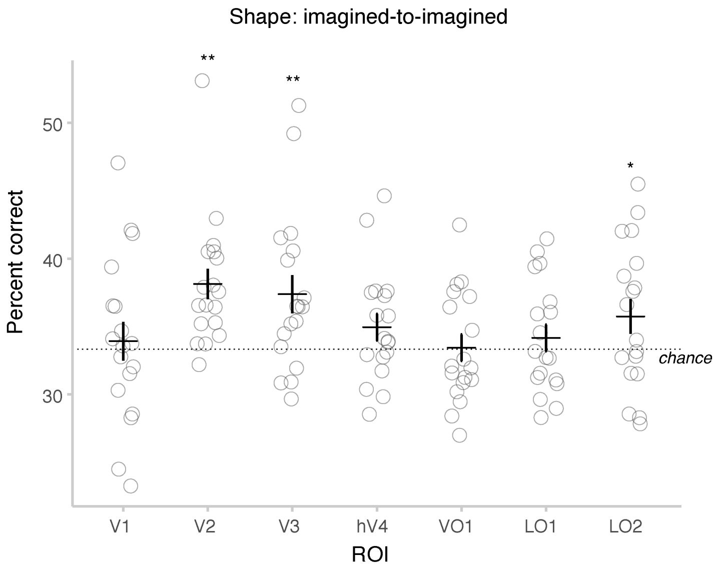

Figure 4. **Shape decoding results.** Classifiers were trained to distinguish between the three shapes (longish, round, pile-shaped) and tested on data from one of the six imagery runs that was excluded from the training procedure in a six-fold leave-one-run-out cross-validation scheme. The shape property of objects was orthogonal to the color feature dimension (see Mean decoding accuracies were significantly above chance in areas V2, V3, and LO2 according to permutation tests. Horizontal and vertical bars represent group means and SEMs, respectively. Chance level was 1/3. $^{*}p < .05$, $^{**}p < .01$, FWE corrected.

imagined objects could indeed be decoded successfully from area hV4 (36.2 %, p = .005, FWE corrected, Cohen's d = 1.08). This means that the color-specific patterns of fMRI activity elicited by object imagery resembled those measured during actual color perception.

Decoding accuracies in the remaining areas were 32.5 % (p = 1, Cohen's d = -.22) in V1, 34.2 % (p = .667, Cohen's d = .21) in V2, 33.5 % (p = .971, Cohen's d = .06) in V3, 33.2 % (p = .993, Cohen's d = -.04) in VO1, 32.1 % (p = 1, Cohen's d = -.35) in LO1, and 34.1 % (p = .757, Cohen's d = .25) in LO2 (all FWE corrected).

Since earlier visual areas V1 and V2 are known to be selective for color, we should emphasize that the fact that null hypotheses in these areas, in contrast to hV4, were retained cannot be explained by poor signal quality relative to hV4. As shown in the previous section, decoding accuracies in fact tended to be slightly larger in V1 and V2 than in hV4. Post-hoc t-tests revealed that this difference, however, marginally failed to reach significance (V1: t_{17} = 1.4189, p = .087; V2: t_{17} = 1.619, p = .0619, each one-tailed and uncorrected).

Decoding the shape of imagined objects

We sought to check that our participants did in fact imagine the objects as a whole, i.e., including object shape, which was irrelevant to the color judgment task. The objects in this experiment were chosen such that there were three identical types of shape in every color category: "longish", "round", "pile-shaped". If observers imagined the entire objects instead of just retaining a (possibly non-pictorial) representation of that object's color in mind, we would expect shape information also to be represented in the neural signal evoked during object imagery. We tested this by training classifiers on imagery responses to discriminate between the three shape categories and testing them on fMRI responses from a run that was not part of the training set (leave-one-run-out cross-validation). As shown in he imagined object shape could successfully decoded from areas V2 (38.1 %, p = .001, FWE corrected, Cohen's d = 1.01), V3 (37.4 %, p = .001, FWE corrected, Cohen's d = .68), and LO2 (35.7 %, p = .01, FWE corrected, Cohen's d = .44). This is consistent with the interpretation that participants' object imagery on average encompassed also the representation of object shape although it was irrelevant to the color judgment task.

Brain-based drift diffusion modeling

We identified hV4 as a visual area where the color of imagined objects could be decoded using color classifiers trained on responses to perceived colors. It is unclear, however, if the neuronal signal underlying this observation plays a functional role for object imagery. We therefore sought to link the predictions of the classifier with our participants' behavior. We fitted hierarchical drift diffusion models (HDDM) to our participants' behavioral data (errors and RTs) and tested if the drift rate in these models differed for imagery trials in which object color was decoded correctly relative to incorrect trials. Our convergence diagnostics including visual inspection of trace plots, autocorrelation between samples at different lags, and Gelman-Rubin statistic calculations (for all parameters, $R < 1.02$) showed that the chains had converged to their stationary distributions. They could therefore be combined and used for Bayesian inference. We found that drift rates were higher when color decoders correctly predicted the color of an imagined object in the trial that immediately preceded the color judgment. This finding held true only for area hV4. The posterior probability of a higher drift rate on correct trials was $p(v_c > v_i|D) = .9802$ (see "pre-judgment model" in . This means that observers performed the color judgment more easily when the object color could be correctly predicted from hV4 activity measured in the imagery blocks immediately preceding the color judgment task. The posterior probabilities for the models fit to data in the other ROIs did not reach the .95 threshold (V1: $p(v_c > v_i|D) = .2714$; V2: $p(v_c > v_i|D) = .219$; V3:

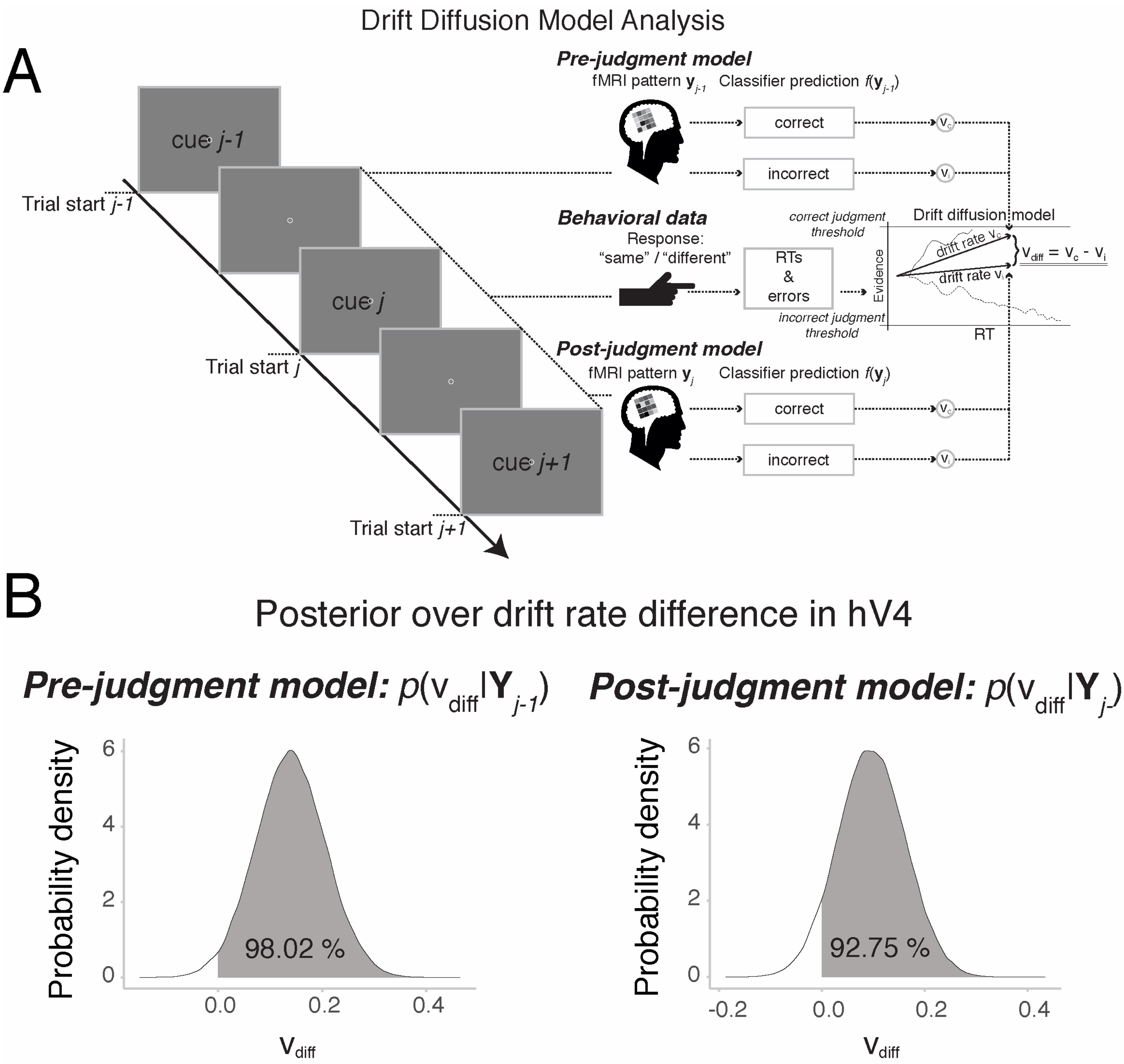

Figure 5. **Brain-based drift diffusion modeling.** Hierarchical drift diffusion models (HDDM) were used to study the relationship between the information content in ROIs and performance in the color judgment task. (A) Behavioral data (RTs and errors) on trial j were modeled with different drift rates depending on whether the classifier predicted the correct or the incorrect label either for the response pattern y_{j-1} on the previous trial $j - 1$ ("pre-judgment model"), which was recorded before the behavioral judgment, or the response pattern y_j on the same trial j, which was recorded after the behavioral judgment ("post-judgment model"). As depicted on the right, higher drift rates mean faster response times and fewer errors. (B) Posterior probability distribution over the difference v_{diff} between drift rates v_c on correctly classified and v_i on incorrectly classified trials for the two models in area hV4. The posterior probability of v_c being larger than v_i was 98.02 % in the pre-judgment model, i.e., when different drift rates were assumed depending on the classification of the trial preceding the behavioral judgment. However, when instead classifier correctness for the imagery block following the behavioral judgment was taken into account (post-judgment model), the posterior probability dropped to 92.75 %.

$p(v_c > v_i|D) = .7298$; VO1: $p(v_c > v_i|D) = .8296$; LO1: $p(v_c > v_i|D) = .2277$; LO2: $p(v_c > v_i|D) = .4801$). A comparison with a null model that assumed only one drift rate irrespective of classifier correctness showed that, in contrast to all other ROIs, the more complex hV4 model provided the better fit in terms of the Deviance Information Criterion: $DIC_{full} = -597.692 < DIC_{restricted} = 596.0735$.

When considering the classifier prediction of the imagery trial immediately following the color judgment, the posterior probability for hV4 did not to exceed the .95 threshold and dropped to $p(v_c > v_i|D) = .9275$ (see "post-judgment model" in . Likewise, drift rates were not larger on correctly than incorrectly decoded trials in any of the other ROIs (V1: $p(v_c > v_i|D) = .6$; V2: $p(v_c > v_i|D) = .0663$; V3: $p(v_c > v_i|D) = .7447$; VO1: $p(v_c > v_i|D) = .4414$; LO1: $p(v_c > v_i|D) = .3165$; LO2: $p(v_c > v_i|D) = .2921$). Also when assuming different drift rates depending on classifier accuracies two trials before the color judgment, the posterior probabilities in none of the ROIs exceeded the threshold of .95 (V1: $p(v_c > v_i|D) = .5894$; V2: $p(v_c > v_i|D) = .2907$; V3: $p(v_c > v_i|D) = .2214$; hV4: $p(v_c > v_i|D) = .8709$; VO1: $p(v_c > v_i|D) = .3923$; LO1: $p(v_c > v_i|D) = .1151$; LO2: $p(v_c > v_i|D) = .2416$).

The correlation between the correctness of the classifier predictions in hV4 and behavioral performance suggests that the information content represented in the activity patterns in that area were behaviorally relevant.

DISCUSSION

We carried out an fMRI experiment to investigate the role of low-level color representations during imagery of colored objects. Classifiers trained to distinguish between perceived colors – red, green, or yellow – could predict the color of the object our participant was imagining based on the activity in hV4. According to our results participants also imagined the shape of the object, which was orthogonal to the color judgment task, showing that they did actually represent the whole object.

Our results show that imagining a colored object elicits an activity pattern in area hV4 that is similar to the perception of the actual color matching that object. Hence, the color experience that is internally generated during imagery corresponds phenomenologically to the perception that is triggered by bottom-up color stimulation. We were able to map this phenomenological correspondence to area V4, which is in agreement with its central role in color perception (Bartels & Zeki, 2000; Bouvier & Engel, 2005; Brouwer & Heeger, 2009, 2013; Conway & Tsao, 2009; Lueck et al., 1989). The perceptual relevance of color-selective activity in hV4 is further corroborated by the psychophysiological link that we discovered using a brain-based behavioral modeling approach. Observers performed better in the color judgment task when object color could successfully be decoded from hV4 activity immediately before the behavioral decision. This finding provides an important additional clue to the perceptual nature of this neural signal because perception has evolved to guide behavior in an adaptive way (Friston, 2010; Hoffman, Singh, & Prakash, 2015; Purves, Wojtach, & Lotto, 2011).

To the extent that visual imagery and working memory rely on the same mechanisms (Albers et al., 2013; Keogh & Pearson, 2011), our results are thus consistent with theories ascribing a central role to perceptual representations in visual short-term memory and imagery (D'Esposito & Postle, 2014; Finke, 1980; Pasternak & Greenlee, 2005). Memory for color thus modulates neuronal excitability and elicits sustained firing of V4 neurons in monkeys (Ferrera, Rudolph, & Maunsell, 1994; Motter, 1994). Interestingly, simultaneous recordings from V4 and prefrontal cortex showed phase-locking of local field potentials (LFP) to correlate with the behavioral accuracy in working memory for colored objects (Liebe, Hoerzer, Logothetis, & Rainer, 2012). Since LFP is a good predictor for BOLD activity (Logothetis, 2008), this may provide evidence for a neural mechanism underlying the relationship between hV4 activity and task performance observed in our study.

In sum, the present findings hence suggest a similar function of V4 in humans as well. Importantly, it extends previous findings by showing that activity in this area was selective for the precise content of internally generated color experience, thereby signaling the color of the imagined object. Furthermore, the neural representations shared between color perception and object imagery indicate that the representational format in hV4 is sufficiently general to encode color irrespective of whether it is perceived in isolation ("void mode", Zeki, 1983) or whether it occurs as part of an imagined object. This distinguishes it from color representations in other early visual areas such as V1, which we could only show to represent perceived color. Such dissociations may reflect that color is a non-unitary attribute linked to multiple perceptual primitives (Mausfeld, 2003) and, as such, is represented at a multitude of processing stages from striate cortex (Engel, Zhang, & Wandell, 1997; Wachtler, Sejnowski, & Albright, 2003) to V4 (e.g., Beauchamp et al., 1999; Brouwer & Heeger, 2009; Conway et al., 2007; Tanigawa, Lu, & Roe, 2010; Zeki, 1980), more anterior ventral stream regions (Lafer-Sousa, Conway, & Kanwisher, 2016), and even prefrontal cortex (Bird, Berens, Horner, & Franklin, 2014).

Given that area V4 is involved in the processing of object properties such as shape (Bushnell & Pasupathy, 2012; Dumoulin & Hess, 2007; Kobatake & Tanaka, 1994; Pasupathy & Connor, 2002) and texture (Kohler, Clarke, Yakovleva, Liu, & Norcia, 2016), it is plausible that there is an overlap in the coding of color in perception and object imagery. This is because the internal generation of object percepts requires a unified representation of several different object-related features (such as shape, color, texture, glossiness, etc.), which may engage in particular those neural representations that afford a suitable degree of feature binding with color (Seymour, Clifford, Logothetis, & Bartels, 2009, 2010; Seymour, Williams, & Rich, 2015).

The present findings appear less consistent with results from an fMRI decoding study on working memory that failed to find color-specific activity patterns in extrastriate areas (Serences et al., 2009) because our results imply that imagery can lead to color-specific responses in hV4. In their behavioral paradigm, however, saturation, not hue, was task-relevant and the observed effect may thus merely reflect the incidental co-activation of hue representations while attending to saturation (instead of orientation). As noted above, direct comparisons between working memory and imagery are not unconditionally permissible because participants may have pursued different cognitive strategies to meet task demands (Keogh & Pearson, 2011).

It can thus be instructive to analyze the different roles of striate cortex and more downstream areas from a process-based perspective that distinguishes between top-down signals that can be characterized as either explicit and voluntary or as implicit and involuntary (Albright, 2012; Pearson & Westbrook, 2015). With respect to color, the current understanding of the extent to which these two categories of processes engage a common neural substrate is as yet poor. According to the picture emerging from recent findings, color-selectivity of BOLD signals in early areas seem to reflect implicit top-down influences on color processing (Amano, Shibata, Kawato, Sasaki, & Watanabe, 2016; Bannert & Bartels, 2013) while in higher areas this tends to apply rather for explicit effects (Brouwer & Heeger, 2013; Vandenbroucke et al., 2016).

From a process-based viewpoint, the present findings will therefore have to be discussed especially with respect to visual attention. There are two reasons for this: first, it plays a central role in the binding of object features (Humphreys, 2016; Treisman, 1988) and, second, because it rests on similar cognitive top-down mechanisms as working memory (D'Esposito & Postle, 2014; Gazzaley & Nobre, 2012; Kastner & Ungerleider, 2000). Both properties make it hence likely for object imagery to be accomplished by attending to internal object feature representations. The sensitivity of V4 neuronal activity to attention to color is well established by electrophysiological work in nonhuman primates (McAdams & Maunsell, 2000; Moran & Desimone, 1985) and human imaging evidence (Bartels & Zeki, 2000; Brouwer & Heeger, 2013; Saenz, Buracas, & Boynton, 2002). Feature-based attention can flexibly change the spatial tuning of V4 neurons along stimulus dimensions depending on task demands to implement a matched filter for the target stimulus in visual search (David, Hayden, Mazer, & Gallant, 2008). It is plausible that such task-dependent tuning changes may be expressed as changes in presynaptic integration processes, which can be detected with fMRI and may depend on (top-down) input from other brain regions (e.g., Liebe, Logothetis, & Rainer, 2011).

The fact that color classification did not generalize from perceived color to imagined object color in V1 or V2 does not imply that such an effect would not have been obtained with more sensitive methods or even that those areas do not partake in the representation of color in object imagery. One imaging study conducted at 7 T field strength showed that imagined pieces of art could be identified from activity in V1 and V2 (Naselaris et al., 2015), but it did not identify the unique contributions of different visual features (and none of them was color). We do not believe, however, that poor sensitivity can explain our findings because, in the present study, perceived color could be predicted at least as accurately from fMRI patterns in V1 or V2 as in hV4. We interpret the difference that we observed between early visual areas and hV4 therefore as reflecting the increase in sensitivity to top-down processing like attention for higher visual areas (Kastner & Ungerleider, 2000; Saenz et al., 2002), which may give rise to a perception/imagery gradient in the visual cortex (S.-H. Lee, Kravitz, & Baker, 2012). It is has been argued that the involvement of V1 in imagery is task-dependent such that imagery tasks requiring more detailed information (p. 596 Pearson, Naselaris, Holmes, & Kosslyn, 2015), may activate more low-level visual features as well. We aimed to foster processing of color detail in our task, by instructing our participants to imagine a particular exemplar of an object category as opposed to others and additionally had them complete a study phase before each scan to ensure that they could distinguish the target image (e.g., of a tomato) from distractor images belonging to the same basic level category. Nevertheless, we did not find an effect in V1.

To conclude, our experiment shows that in subjective experience of self-generated mental images, the color of objects is represented in a very similar way as colors that we actually see. We identified area hV4 as a neural site bridging the domains of perceived and imagined object color. The fact that neural activity in this area predicted behavioral performance highlights its role in generating color percepts, be they externally triggered or a product of our own minds.

ACKNOWLEDGMENTS

This work was supported by a grant from the German Federal Ministry for Education and Research (BMBF), grant number: FKZ 01GQ1002, by the German Excellence Initiative of the German Research Foundation (DFG) grant number EXC307, and by the Max Planck Society, Germany. The authors declare no competing financial interests.

References

Albers, A. M., Kok, P., Toni, I., Dijkerman, H. C., & de Lange, F. P. (2013). Shared representations for working memory and mental imagery in early visual cortex. *Current Biology*, *23*, 1427–31.

Albright, T. D. (2012). On the perception of probable things: neural substrates of associative memory, imagery, and perception. *Neuron*, *74*, 227–45.

Amano, K., Shibata, K., Kawato, M., Sasaki, Y., & Watanabe, T. (2016). Learning to Associate Orientation with Color in Early Visual Areas by Associative Decoded fMRI Neurofeedback. *Current Biology*, *26*, 1861–6.

Bannert, M. M. & Bartels, A. (2013). Decoding the yellow of a gray banana. *Current Biology*, *23*, 2268–72.

Bartels, A. & Zeki, S. (2000). The architecture of the colour centre in the human visual brain: new results and a review. *European Journal of Neuroscience*, *12*, 172–93.

Beauchamp, M. S., Haxby, J. V., Jennings, J. E., & DeYoe, E. A. (1999). An fMRI version of the Farnsworth-Munsell 100-Hue test reveals multiple color-selective areas in human ventral occipitotemporal cortex. *Cerebral Cortex*, *9*, 257–63.

Bird, C. M., Berens, S. C., Horner, a. J., & Franklin, a. (2014). Categorical encoding of color in the brain. *Proceedings of the National Academy of Sciences*, 1–6.

Bouvier, S. E. & Engel, S. A. (2005). Behavioral Deficits and Cortical Damage Loci in Cerebral Achromatopsia. *Cerebral Cortex*, *16*, 183–191.

Brainard, D. H. & Maloney, L. T. (2011). Surface color perception and equivalent illumination models. *Journal of Vision*, *11*, 1–18.

Brooks, J. (2012). Counterbalancing for serial order carry-over effects in experimental condition orders. *Psychological Methods*, 1–54.

Brouwer, G. J. & Heeger, D. J. (2009). Decoding and Reconstructing Color from Responses in Human Visual Cortex. *Journal of Neuroscience*, *29*, 13992–14003.

Brouwer, G. J. & Heeger, D. J. (2013). Categorical Clustering of the Neural Representation of Color. *Journal of Neuroscience*, *33*, 15454–15465.

Bushnell, B. N. & Pasupathy, A. (2012). Shape encoding consistency across colors in primate V4. *Journal of Neurophysiology*, *108*, 1299–1308.

Chang, S., Lewis, D. E., & Pearson, J. (2013). The functional effects of color perception and color imagery. *13*, 1–10.

Cichy, R. M., Heinzle, J., & Haynes, J. D. (2012). Imagery and perception share cortical representations of content and location. *Cerebral Cortex*, *22*, 372–380.

Conway, B. R., Moeller, S., & Tsao, D. Y. (2007). Specialized color modules in macaque extrastriate cortex. *Neuron*, *56*, 560–73.

Conway, B. R. & Tsao, D. Y. (2009). Color-tuned neurons are spatially clustered according to color preference within alert macaque posterior inferior temporal cortex. *Proceedings of the National Academy of Sciences*, *106*, 18034–9.

David, S. V., Hayden, B. Y., Mazer, J. a., & Gallant, J. L. (2008). Attention to stimulus features shifts spectral tuning of V4 neurons during natural vision. *Neuron*, *59*, 509–21.

De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., & Formisano, E. (2008). Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *NeuroImage*, *43*, 44–58.

D'Esposito, M. & Postle, B. R. (2014). The Cognitive Neuroscience of Working Memory. *Annual Review of Psychology*, *66*, 115–42.

Dumoulin, S. O. & Hess, R. F. (2007). Cortical specialization for concentric shape processing. *Vision Research*, *47*, 1608–1613.

Engel, S., Zhang, X., & Wandell, B. A. (1997). Colour tuning in human visual cortex measured with functional magnetic resonance imaging. *Nature*, *388*, 68–71.

Ferrera, V. P., Rudolph, K. K., & Maunsell, J. H. R. (1994). Responses of neurons in the parietal and temporal visual pathways during a motion task. *Journal of Neuroscience*, *14*, 6171–6186.

Finke, R. A. (1980). Levels of equivalence in imagery and perception. *Psychological Review*, *87*, 113–132.

Foster, D. H. (2011). Color constancy. *Vision Research*, *51*, 674–700.

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*, 127–38.

Friston, K. J. & Penny, W. (2003). Posterior probability maps and SPMs. *NeuroImage*, *19*, 1240–1249.

Gazzaley, A. & Nobre, A. C. (2012). Top-down modulation: Bridging selective attention and working memory. *Trends in Cognitive Sciences*, *16*, 129–135.

Gegenfurtner, K. R. & Rieger, J. (2000). Sensory and cognitive contributions of color to the recognition of natural scenes. *Current Biology*, *10*, 805–8.

Gould van Praag, C. D., Garfinkel, S., Ward, J., Bor, D., & Seth, A. K. (2015). Automaticity and localisation of concurrents predicts colour area activity in grapheme-colour synaesthesia. *Neuropsychologia*, *88*, 5–14.

Hansen, T., Olkkonen, M., Walter, S., & Gegenfurtner, K. R. (2006). Memory modulates color appearance. *Nature Neuroscience*, *9*, 1367–8.

Harrison, S. A. & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, *458*, 632–5.

Hoffman, D. D., Singh, M., & Prakash, C. (2015). The Interface Theory of Perception. *Psychonomic Bulletin and Review*, *22*, 1480–506.

Howard, R. J., ffytche, D. H., Barnes, J., McKeefry, D., Ha, Y., Woodruff, P. W., ... Brammer, M. (1998). The functional anatomy of imagining and perceiving colour. *Neuroreport*, *9*, 1019–23.

Hubbard, E. M., Arman, A. C., Ramachandran, V. S., & Boynton, G. M. (2005). Individual Differences among Grapheme-Color Synesthetes: Brain-Behavior Correlations. *Neuron*, *45*, 975–985.

Humphreys, G. W. (2016). Feature confirmation in object perception: Feature integration theory 26 years on from the Treisman Bartlett lecture. *Quarterly Journal of Experimental Psychology*, *69*, 1910–40.

Ishihara, S. (2011). *Ishihara's tests for colour deficiency* (38 Plates). Tokyo, Japan: Kanehara Trading Inc.

Kaiser, P. K. (1991). Flicker as a function of wavelength and heterochromatic flicker photometry. In J. J. Kulikowski, V. Walsh, & I. J. Murray (Eds.), *Limits of vision* (pp. 171–190). Basingstoke (United Kingdom): MacMillan.

Kastner, S. & Ungerleider, L. G. (2000). Mechanisms of visual attention in the human cortex. *Annual Review of Neuroscience*, *23*, 315–41.

Keogh, R. & Pearson, J. (2011). Mental imagery and visual working memory. *PLoS ONE*, *6*.

Kleiner, M., Brainard, D. H., & Pelli, D. G. (2007). "What's new in Psychtoolbox-3?"

Kobatake, E. & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology*, *71*, 856–67.

Kohler, P. J., Clarke, A., Yakovleva, A., Liu, Y., & Norcia, A. M. (2016). Representation of Maximally Regular Textures in Human Visual Cortex. *Journal of Neuroscience*, *36*, 714–729.

Kuriki, I., Sun, P., Ueno, K., Tanaka, K., & Cheng, K. (2015). Hue selectivity in human visual cortex revealed by functional magnetic resonance imaging. *Cerebral Cortex*, *25*, 4869–4884.

Lafer-Sousa, R., Conway, B. R., & Kanwisher, N. G. (2016). Color-Biased Regions of the Ventral Visual Pathway Lie between Face- and Place-Selective Regions in Humans, as in Macaques. *Journal of Neuroscience*, *36*, 1682–97.

Larsson, J. & Heeger, D. J. (2006). Two retinotopic visual areas in human lateral occipital cortex. *Journal of Neuroscience*, *26*, 13128–42.

Ledoit, O. & Wolf, M. (2004). Honey, I Shrunk the Sample Covariance Matrix. *The Journal of Portfolio Management*, *30*, 110–119.

Lee, M. & Wagenmakers, E. (2012). Bayesian cognitive modeling: A practical course.

Lee, S.-H. & Baker, C. I. (2016). Multi-Voxel Decoding and the Topography of Maintained Information During Visual Working Memory. *Frontiers in Systems Neuroscience*, *10*, 2.

Lee, S.-H., Kravitz, D. J., & Baker, C. I. (2012). Disentangling visual imagery and perception of real-world objects. *NeuroImage*, *59*, 4064–73.

Liebe, S., Hoerzer, G. M., Logothetis, N. K., & Rainer, G. (2012). Theta coupling between V4 and prefrontal cortex predicts visual short-term memory performance. *Nature Neuroscience*, *15*, 456–462.

Liebe, S., Logothetis, N. K., & Rainer, G. (2011). Dissociable effects of natural image structure and color on LFP and spiking activity in the lateral prefrontal cortex and extrastriate visual area V4. *Journal of Neuroscience*, *31*, 10215–10227.

Livingstone, M. S. & Hubel, D. H. (1984). Anatomy and physiology of a color system in the primate visual cortex. *Journal of Neuroscience*, *4*, 309–56.

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature*, *453*, 869–78.

Lueck, C. J., Zeki, S., Friston, K. J., Deiber, M. P., Cope, P., Cunningham, V. J., ... Frackowiak, R. S. (1989). The colour centre in the cerebral cortex of man. *Nature*, *340*, 386–9.

Mausfeld, R. (2003). 'Colour' As Part of the Format of Different Perceptual Primitives: The Dual Coding of Colour. In R. Mausfeld & D. Heyer (Eds.), *Colour perception: mind and the physical world* (pp. 381–430). Oxford: Oxford University Press.

McAdams, C. J. & Maunsell, J. H. R. (2000). Attention to both space and feature modulates neuronal responses in macaque area V4. *Journal of Neurophysiology*, *83*, 1751–5.

Mollon, J. D. (1989). "Tho' she kneel'd in that place where they grew..." The uses and origins of primate colour vision. *The Journal of Experimental Biology*, *146*, 21–38.

Moran, J. & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science*, *229*, 782–784.

Motter, B. C. (1994). Neural correlates of attentive selection for color or luminance in extrastriate area V4. *Journal of Neuroscience*, *14*, 2178–2189.

Naselaris, T., Olman, C. A., Stansbury, D. E., Ugurbil, K., & Gallant, J. L. (2015). A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *NeuroImage*, *105*, 215–228.

Nichols, T. E. & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping*, *15*, 1–25.

Nunn, J. A., Gregory, L. J., Brammer, M., Williams, S. C. R., Parslow, D. M., Morgan, M. J., … Gray, J. a. (2002). Functional magnetic resonance imaging of synesthesia: activation of V4/V8 by spoken words. *Nature Neuroscience*, *5*, 371–5.

Olkkonen, M., Hansen, T., & Gegenfurtner, K. R. (2008). Color appearance of familiar objects: Effects of object shape, texture, and illumination changes. *Journal of Vision*, *8*, 13.1–16.

Parkes, L. M., Marsman, J.-B. C., Oxley, D. C., Goulermas, J. Y., & Wuerger, S. M. (2009). Multivoxel fMRI analysis of color tuning in human primary visual cortex. *Journal of Vision*, *9*, 1.1–13.

Pasternak, T. & Greenlee, M. W. (2005). Working memory in primate sensory systems. *Nature Reviews Neuroscience*, *6*, 97–107.

Pasupathy, A. & Connor, C. E. (2002). Population coding of shape in area V4. *Nature Neuroscience*, *5*, 1332–1338.

Pearson, J. & Kosslyn, S. M. (2015). The heterogeneity of mental representation: Ending the imagery debate. *Proceedings of the National Academy of Sciences*, *112*, 10089–10092.

Pearson, J., Naselaris, T., Holmes, E. A., & Kosslyn, S. M. (2015). Mental Imagery: Functional Mechanisms and Clinical Applications. *Trends in Cognitive Sciences*, *19*, 590–602.

Pearson, J. & Westbrook, F. (2015). Phantom perception: Voluntary and involuntary nonretinal vision. *Trends in Cognitive Sciences*, *19*, 278–284.

Purves, D., Wojtach, W. T., & Lotto, R. B. (2011). Understanding vision in wholly empirical terms. *Proceedings of the National Academy of Sciences*, *108*, 15588–15595.

Reddy, L., Tsuchiya, N., & Serre, T. (2010). Reading the mind's eye: Decoding category information during mental imagery. *NeuroImage*, *50*, 818–825.

Rich, A. N., Williams, M. A., Puce, A., Syngeniotis, A., Howard, M. A., McGlone, F., & Mattingley, J. B. (2006). Neural correlates of imagined and synaesthetic colours. *Neuropsychologia*, *44*, 2918–25.

Rouw, R., Scholte, H. S., & Colizoli, O. (2011). Brain areas involved in synaesthesia: A review. *Journal of Neuropsychology*, *5*, 214–242.

Saenz, M., Buracas, G. T., & Boynton, G. M. (2002). Global effects of feature-based attention in human visual cortex. *Nature Neuroscience*, *5*, 631–632.

Schiller, P. H., Logothetis, N. K., & Charles, E. R. (1990). Functions of the colour-opponent and broad-band channels of the visual system. *Nature*, *343*, 68–70.

Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. (2009). Stimulus-specific delay activity in human primary visual cortex. *Psychological Science*, *20*, 207–214.

Sereno, M. I., McDonald, C. T., & Allman, J. M. (1994). Analysis of retinotopic maps in extrastriate cortex. *Cerebral Cortex*, *4*, 601–20.

Seymour, K. J., Clifford, C. W. G., Logothetis, N. K., & Bartels, A. (2009). The coding of color, motion, and their conjunction in the human visual cortex. *Current Biology*, *19*, 177–83.

Seymour, K. J., Clifford, C. W. G., Logothetis, N. K., & Bartels, A. (2010). Coding and binding of color and form in visual cortex. *Cerebral Cortex*, *20*, 1946–54.

Seymour, K. J., Williams, M. A., & Rich, A. N. (2015). The Representation of Color across the Human Visual Cortex: Distinguishing Chromatic Signals Contributing to Object Form Versus Surface Color. *Cerebral Cortex*, 1–9.

Shevell, S. K. & Kingdom, F. A. A. (2008). Color in complex scenes. *Annual Review of Psychology*, *59*, 143–66.

Tanigawa, H., Lu, H. D., & Roe, A. W. (2010). Functional organization for color and orientation in macaque V4. *Nature Neuroscience*, 1–8.

Treisman, A. (1988). Features and objects: the fourteenth Bartlett memorial lecture. *Quarterly Journal of Experimental Psychology. A*, *40*, 201–37.

van Leeuwen, T. M., Petersson, K. M., & Hagoort, P. (2010). Synaesthetic colour in the brain: Beyond colour areas. A functional magnetic resonance imaging study of synaesthetes and matched controls. *PLoS ONE*, *5*.

Vandenbroucke, A. R. E., Fahrenfort, J. J., Meuwese, J. D. I., Scholte, H. S., & Lamme, V. A. F. (2016). Prior Knowledge about Objects Determines Neural Color Representation in Human Visual Cortex. *Cerebral Cortex*, *26*, 1401–1408.

Wachtler, T., Sejnowski, T. J., & Albright, T. D. (2003). Representation of color stimuli in awake macaque primary visual cortex. *Neuron*, *37*, 681–691.

Wandell, B. A. & Winawer, J. (2011). Imaging retinotopic maps in the human brain. *Vision Research*, *51*, 718–37.

Wantz, A. L., Mast, F. W., & Lobmaier, J. S. (2015). Colors in Mind: A Novel Paradigm to Investigate Pure Color Imagery. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 1152–1161.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics*, *7*, 14.

Witzel, C., Valkova, H., Hansen, T., & Gegenfurtner, K. R. (2011). Object knowledge modulates colour appearance. *i-Perception*, *2*, 13–49.

Xiao, Y., Wang, Y., & Felleman, D. J. (2003). A spatially organized representation of colour in macaque cortical area V2. *Nature*, *421*, 535–9.

Zeki, S. (1980). The representation of colours in the cerebral cortex. *Nature*, *284*, 412–418.

Zeki, S. (1983). Colour coding in the cerebral cortex: the reaction of cells in monkey visual cortex to wavelengths and colours. *Neuroscience*, *9*, 741–65.

Zeki, S. & Stutters, J. (2013). Functional specialization and generalization for grouping of stimuli based on colour and motion. *NeuroImage*, *73*, 156–166.

7 SUMMARY AND CONCLUSION

The three experiments reported here investigated some phenomena in color vision that highlight the complexity between color experience and the spectral power distribution of the light that constitutes the input signal to the visual system. They hence extend the understanding of the functional neuroanatomy of the human brain with regard to color vision because most previous studies conceptualized color in terms of wavelength spectra. Furthermore, we used pattern recognition methods to quantify what information about color is represented in the measured brain signal.

The first experiment showed that perceived color and memory color share a common neural basis in V1, the earliest area of cortical visual processing. In the first part of the fMRI experiment participants viewed gray-scale images of color-diagnostic objects while performing an unrelated motion judgment task. In the second part they saw abstract color stimuli ("real color") that corresponded to the same color categories from which the objects had been drawn. Pattern classifiers trained to distinguish between BOLD responses measured while viewing the real colors could predict which color category an object came from while the observer viewed its gray-scale image. This shows that, if prior color knowledge about the stimulus input is available, it can be fed back to V1 and cause activity patterns similar to those elicited when seeing real color.

The second experiment localized color representations that are robust against changes in illumination to area V1 and to a brain region that is anterior to retinotopic area VO1, previously identified as V4α. Using pattern classification, these regions were found to encode surface color in a way that is invariant with respect to illumination changes. These response patterns are relevant to color constancy because they cease to encode surface color when stimuli are viewed in a reduced-cue condition. This is to be expected as the same manipulation also abolishes color constancy as measured psychophysically. On the other hand, when presented with spectrally matched stimuli that could be distinguished in terms of a difference in either illuminant color or perceived surface color, there is a decreasing illuminant bias (or, equivalently, increasing surface bias) from early to higher visual areas. This means that higher areas are more likely to attribute the chromatic input to a surface rather than to the illumination and vice versa for early areas. Taken together, this suggests that V1 and V4α both make complementary contributions to color constancy.

The third experiment demonstrated how the perception of abstract colors relate to color in visual object imagery. Like dreaming, remembering, or mind-wandering, visual imagery constitutes a highly subjective form of visual experience. The neural representations of object color in visual imagery exhibit a significant overlap with the representation of low-level color stimuli in area V4. The functional role of the activity in this area is further corroborated by its correlation with behavioral performance in a color judgment task. The involvement of area V4 in controlled top-down processing such as object imagery is consistent with its role in other forms of volitional top-down phenomena like attention and working memory.

Taken together, the research described here shows that various brain regions contribute to color vision and do so in different ways. While activity in many visual areas distinguish between stimuli that differ in wavelength composition, they make distinct contributions to color vision. This sensitivity for different colors can also be found in cases when color processing is not induced by chromatic input, as was the case in both the imagery and memory color experiments. While prior knowledge biases color processing via feedback to early visual areas, imagery, a more explicit form of top-down modulation (Albright, 2012), shares neural representations with color perception in higher visual areas. When chromatic input is present, contributions to color constancy can be found in early and late visual areas. Color constancy in primary visual cortex is possibly mediated via its sensitivity to chromatic edges, feedback, or both. The neural architecture that abstracts from wavelength-dependent color representations is distributed across the ventral pathway and not restricted to a single “color center”. Individual visual areas are thus dedicated to different aspects of color perception, depending on task demands and stimulus context, e.g., whether color specifies the property of a surface or illumination.

8 ACKNOWLEDGMENTS

I would like to express my gratitude to Dr. Andreas Bartels for his patient and knowledgeable supervision. I owe to him almost everything that I know about vision sciences and the brain. I benefited a lot from his pragmatic experimental approach and from stimulating discussions as well as his very well-meaning and motivating criticism. I also would like to thank the members of the Vision and Cognition Lab for their kind support and for the stimulating enjoyable scientific environment. Lastly, I would like to thank my friends and family, in particular my parents Barbara and Peter, for their encouragement and emotional support.

References

Albright, T. D. (2012). On the perception of probable things: neural substrates of associative memory, imagery, and perception. *Neuron*, *74*, 227–45.

Allred, S. R. & Brainard, D. H. (2013). A Bayesian model of lightness perception that incorporates spatial variation in the illumination. *Journal of Vision*, *13*, 18.

Bartels, A. & Zeki, S. (2000). The architecture of the colour centre in the human visual brain: new results and a review. *European Journal of Neuroscience*, *12*, 172–93.

Beauchamp, M. S., Haxby, J. V., Jennings, J. E., & DeYoe, E. A. (1999). An fMRI version of the Farnsworth-Munsell 100-Hue test reveals multiple color-selective areas in human ventral occipitotemporal cortex. *Cerebral Cortex*, *9*, 257–63.

Bird, C. M., Berens, S. C., Horner, a. J., & Franklin, a. (2014). Categorical encoding of color in the brain. *Proceedings of the National Academy of Sciences*, 1–6.

Bornstein, M. H. & Korda, N. O. (1984). Discrimination and matching within and between hues measured by reaction times: some implications for categorical perception and levels of information processing. *Psychological Research*, *46*, 207–222.

Bouvier, S. E., Cardinal, K. S., & Engel, S. A. (2008). Activity in visual area V4 correlates with surface perception. *Journal of Vision*, *8*, 28.1–9.

Bouvier, S. E. & Engel, S. A. (2005). Behavioral Deficits and Cortical Damage Loci in Cerebral Achromatopsia. *Cerebral Cortex*, *16*, 183–191.

Brainard, D. H. & Freeman, W. T. (1997). Bayesian color constancy. *Journal of the Optical Society of America. A*, *14*, 1393–1411.

Brainard, D. H., Longère, P., Delahunt, P. B., Freeman, W. T., Kraft, J. M., & Xiao, B. (2006). Bayesian model of human color constancy. *Journal of Vision*, *6*, 1267–81.

Brainard, D. H. & Maloney, L. T. (2011). Surface color perception and equivalent illumination models. *Journal of Vision*, *11*, 1–18.

Brewer, A. A., Liu, J., Wade, A. R., & Wandell, B. A. (2005). Visual field maps and stimulus selectivity in human ventral occipital cortex. *Nature Neuroscience*, *8*, 1102–9.

Broadbent, D. E. (1958). *Perception and Communication*. London: Pergamon.

Brouwer, G. J. & Heeger, D. J. (2009). Decoding and Reconstructing Color from Responses in Human Visual Cortex. *Journal of Neuroscience*, *29*, 13992–14003.

Brouwer, G. J. & Heeger, D. J. (2013). Categorical Clustering of the Neural Representation of Color. *Journal of Neuroscience*, *33*, 15454–15465.

Bushnell, B. N. & Pasupathy, A. (2012). Shape encoding consistency across colors in primate V4. *Journal of Neurophysiology*, *108*, 1299–1308.

Conway, B. R. (2003). Colour Vision: A Clue to Hue in V2. *Current Biology*, *13*, R308–R310.

Conway, B. R., Chatterjee, S., Field, G. D., Horwitz, G. D., Johnson, E. N., Koida, K., & Mancuso, K. (2010). Advances in Color Science: From Retina to Behavior. *Journal of Neuroscience*, *30*, 14955–14963.

Conway, B. R., Hubel, D. H., & Livingstone, M. S. (2002). Color contrast in macaque V1. *Cerebral Cortex*, *12*, 915–925.

Conway, B. R., Moeller, S., & Tsao, D. Y. (2007). Specialized color modules in macaque extrastriate cortex. *Neuron*, *56*, 560–73.

Conway, B. R. & Tsao, D. Y. (2009). Color-tuned neurons are spatially clustered according to color preference within alert macaque posterior inferior temporal cortex. *Proceedings of the National Academy of Sciences*, *106*, 18034–9.

Dacey, D. M. & Lee, B. B. (1994). The 'blue-on' opponent pathway in primate retina originates from a distinct bistratified ganglion cell type. *Nature*, *367*, 731–735.

De Valois, R. L., Cottaris, N. P., Elfar, S. D., Mahon, L. E., & Wilson, J. A. (2000). Some transformations of color information from lateral geniculate nucleus to striate cortex. *Proceedings of the National Academy of Sciences*, *97*, 4997–5002.

De Valois, R. L. & Jacobs, G. H. (1968). Primate color vision. *Science*, *162*, 533–40.

Denison, R. N., Vu, A. T., Yacoub, E., Feinberg, D. A., & Silver, M. A. (2014). Functional mapping of the magnocellular and parvocellular subdivisions of human LGN. *NeuroImage*, *102*, 358–369.

Derrington, A. M., Krauskopf, J., & Lennie, P. (1984). Chromatic mechanisms in lateral geniculate nucleus of Macaque. *The Journal of Physiology*, *357*, 241–65.

Desimone, R. & Schein, S. J. (1987). Visual properties of neurons in area V4 of the macaque: sensitivity to stimulus form. *Journal of Neurophysiology*, *57*, 835–868.

DeYoe, E. A. & Van Essen, D. C. (1985). Segregation of efferent connections and receptive field properties in visual area V2 of the macaque. *Nature*, *317*, 58–61.

Engel, S., Zhang, X., & Wandell, B. A. (1997). Colour tuning in human visual cortex measured with functional magnetic resonance imaging. *Nature*, *388*, 68–71.

Fechner, G. T. (1860). *Elemente der Psychophysik*. Leipzig: Breitkopf & Härtel.

Federer, F., Williams, D., Ichida, J. M., Merlin, S., & Angelucci, A. (2013). Two Projection Streams from Macaque V1 to the Pale Cytochrome Oxidase Stripes of V2. *Journal of Neuroscience*, *33*, 11530–11539.

ffytche, D. H., Howard, R. J., Brammer, M. J., David, A., Woodruff, P., & Williams, S. (1998). The anatomy of conscious vision: an fMRI study of visual hallucinations. *Nature Neuroscience*, *1*, 738–42.

Field, G. D. & Chichilnisky, E. J. (2007). Information processing in the primate retina: circuitry and coding. *Annual Review of Neuroscience*, *30*, 1–30.

Friedman, H. S., Zhou, H., & von der Heydt, R. (2003). The coding of uniform colour figures in monkey visual cortex. *Journal of Physiology*, *548*, 593–613.

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*, 127–38.

Gegenfurtner, K. R. (2003). Cortical mechanisms of colour vision. *Nature Reviews Neuroscience*, *4*, 563–72.

Gegenfurtner, K. R., Kiper, D. C., & Fenstemaker, S. B. (1996). Processing of color, form, and motion in macaque area V2. *Visual Neuroscience*, *13*, 161–172.

Goddard, E., Mannion, D. J., McDonald, J. S., Solomon, S. G., & Clifford, C. W. G. (2011). Color responsiveness argues against a dorsal component of human V4. *Journal of Vision*, *11*, 1–21.

Gould van Praag, C. D., Garfinkel, S., Ward, J., Bor, D., & Seth, A. K. (2015). Automaticity and localisation of concurrents predicts colour area activity in grapheme-colour synaesthesia. *Neuropsychologia*, *88*, 5–14.

Hansen, K. A., Kay, K. N., & Gallant, J. L. (2007). Topographic organization in and near human visual area V4. *The Journal of Neuroscience*, *27*, 11896–911.

Hansen, T., Olkkonen, M., Walter, S., & Gegenfurtner, K. R. (2006). Memory modulates color appearance. *Nature Neuroscience*, *9*, 1367–8.

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. *Annual Review of Neuroscience*, 435–456.

Haynes, J.-D. & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, *7*, 523–34.

Helmholtz, H. (1867). *Handbuch der Physiologischen Optik.* Leipzig: Leopold Voss.

Hendry, S. H. C. & Reid, R. C. (2000). The Koniocellular Pathway in Primate Vision. *Annual Review of Neuroscience*, *23*, 127–153.

Hering, E. (1920). *Grundzüge der Lehre vom Lichtsinn.* Berlin: Springer.

Horiguchi, H., Winawer, J., Dougherty, R. F., & Wandell, B. A. (2013). Human trichromacy revisited. *Proceedings of the National Academy of Sciences*, *110*, E260–9.

Horwitz, G. D. & Hass, C. A. (2012). Nonlinear analysis of macaque V1 color tuning reveals cardinal directions for cortical color processing. *Nature Neuroscience*, *15*, 913–9.

Howard, R. J., Ffytche, D. H., Barnes, J., McKeefry, D., Ha, Y., Woodruff, P. W., ... Brammer, M. (1998). The functional anatomy of imagining and perceiving colour. *Neuroreport*, *9*, 1019–23.

Hubbard, E. M. & Ramachandran, V. S. (2005). Neurocognitive mechanisms of synesthesia. *Neuron*, *48*, 509–20.

Hubel, D. H. & Livingstone, M. S. (1987). Segregation of form, color, and stereopsis in primate area 18. *Journal of Neuroscience*, *7*, 3378–3415.

Johnson, E. N., Hawken, M. J., & Shapley, R. (2001). The spatial transformation of color in the primary visual cortex of the macaque monkey. *Nature Neuroscience*, *4*, 409–16.

Kiper, D. C., Fenstemaker, S. B., & Gegenfurtner, K. R. (1997). Chromatic properties of neurons in macaque area V2. *Visual Neuroscience*, *14*, 1061–1072.

Knill, D. C. & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*, 712–719.

Knill, D. C. & Richards, W. (1996). *Perception as Bayesian Inference*. Cambridge: Cambridge University Press.

Koenderink, J. J. (2010). *Color for the Sciences*. Cambridge, MA: The MIT Press.

Koida, K. & Komatsu, H. (2007). Effects of task demands on the responses of color-selective neurons in the inferior temporal cortex. *Nature Neuroscience*, *10*, 108–16.

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*, 408–24.

Kuriki, I., Sun, P., Ueno, K., Tanaka, K., & Cheng, K. (2015). Hue selectivity in human visual cortex revealed by functional magnetic resonance imaging. *Cerebral Cortex*, *25*, 4869–4884.

Kusunoki, M., Moutoussis, K., & Zeki, S. (2006). Effect of background colors on the tuning of color-selective cells in monkey area V4. *Journal of Neurophysiology*, *95*, 3047–59.

Lafer-Sousa, R. & Conway, B. R. (2013). Parallel, multi-stage processing of colors, faces and shapes in macaque inferior temporal cortex. *Nature Neuroscience*, *16*, 1870–8.

Lafer-Sousa, R., Conway, B. R., & Kanwisher, N. G. (2016). Color-Biased Regions of the Ventral Visual Pathway Lie between Face- and Place-Selective Regions in Humans, as in Macaques. *Journal of Neuroscience*, *36*, 1682–97.

Lennie, P., Krauskopf, J., & Sclar, G. (1990). Chromatic mechanisms in striate cortex of macaque. *Journal of Neuroscience*, *10*, 649–69.

Leventhal, A. G., Rodieck, R. W., & Dreher, B. (1981). Retinal ganglion cell classes in the Old World monkey: morphology and central projections. *Science*, *213*, 1139–1142.

Leventhal, A. G., Thompson, K. G., Liu, D., Zhou, Y., & Ault, S. J. (1995). Concomitant sensitivity to orientation, direction, and color of cells in layers 2, 3, and 4 of monkey striate cortex. *Journal of Neuroscience*, *15*, 1808–1818.

Li, M., Liu, F., Juusola, M., & Tang, S. (2014). Perceptual Color Map in Macaque Visual Area V4. *Journal of Neuroscience*, *34*, 202–17.

Liu, J. & Wandell, B. a. (2005). Specializations for chromatic and temporal signals in human visual cortex. *Journal of Neuroscience*, *25*, 3459–68.

Livingstone, M. S. & Hubel, D. H. (1984). Anatomy and physiology of a color system in the primate visual cortex. *Journal of Neuroscience*, *4*, 309–56.

Livingstone, M. S. & Hubel, D. H. (1988). Segregation of form, color, movement, and depth: anatomy, physiology, and perception. *Science*, *240*, 740–9.

Lu, H. D. & Roe, A. W. (2008). Functional organization of color domains in V1 and V2 of Macaque monkey revealed by optical imaging. *Cerebral Cortex*, *18*, 516–533.

MacEvoy, S. P. & Paradiso, M. A. (2001). Lightness constancy in primary visual cortex. *Proceedings of the National Academy of Sciences*, *98*, 8827–8831.

Maloney, L. T. (1999). Physics-based approaches to modeling surface color perception. In K. R. Gegenfurtner & L. T. Sharpe (Eds.), *Color vision: from genes to perception* (pp. 387–416). Cambridge, UK: Cambridge University Press.

Marr, D. (1982). *Vision*. San Francisco: Freeman.

Martin, P. R. & Lee, B. B. (2014). Distribution and specificity of S-cone ("blue cone") signals in subcortical visual pathways. *Visual Neuroscience*, *31*, 177–187.

Martin, P. R., White, A. J. R., Goodchild, A. K., Wilder, H. D., & Sefton, A. E. (1997). Evidence that blue-on cells are part of the third geniculocortical pathway in primates. *European Journal of Neuroscience*, *9*, 1536–1541.

Mausfeld, R. (2003). 'Colour' As Part of the Format of Different Perceptual Primitives: The Dual Coding of Colour. In R. Mausfeld & D. Heyer (Eds.), *Colour perception: mind and the physical world* (pp. 381–430). Oxford: Oxford University Press.

McKeefry, D. J. & Zeki, S. (1997). The position and topography of the human colour centre as revealed by functional magnetic resonance imaging. *Brain*, *120*, 2229–2242.

Mendola, J. D., Dale, A. M., Fischl, B., Liu, A. K., & Tootell, R. B. (1999). The representation of illusory and real contours in human cortical visual areas revealed by functional magnetic resonance imaging. *Journal of Neuroscience*, *19*, 8560–72.

Michael, C. R. (1978). Color vision mechanisms in monkey striate cortex: dual-opponent cells with concentric receptive fields. *Journal of Neurophysiology*, *41*, 572–88.

Monnier, P. & Shevell, S. K. (2003). Large shifts in color appearance from patterned chromatic backgrounds. *Nature Neuroscience*, *6*, 801–802.

Murphey, D. K., Yoshor, D., & Beauchamp, M. S. (2008). Perception matches selectivity in the human anterior color center. *Current Biology*, *18*, 216–20.

Nasr, S., Polimeni, J. R., & Tootell, R. B. H. (2016). Interdigitated Color- and Disparity-Selective Columns within Human Visual Cortical Areas V2 and V3. *Journal of Neuroscience*, *36*, 1841–57.

Neisser, U. (1967). *Cognitive Psychology.* New York: Appleton.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*, 424–30.

Nunn, J. A., Gregory, L. J., Brammer, M., Williams, S. C. R., Parslow, D. M., Morgan, M. J., … Gray, J. a. (2002). Functional magnetic resonance imaging of synesthesia: activation of V4/V8 by spoken words. *Nature Neuroscience*, *5*, 371–5.

Olkkonen, M., Hansen, T., & Gegenfurtner, K. R. (2008). Color appearance of familiar objects: Effects of object shape, texture, and illumination changes. *Journal of Vision*, *8*, 13.1–16.

Olkkonen, M., Saarela, T. P., & Allred, S. R. (2016). Perception-memory interactions reveal a computational strategy for perceptual constancy. *Journal of Vision*, *16*, 38.

Pasupathy, A. & Connor, C. E. (2002). Population coding of shape in area V4. *Nature Neuroscience*, *5*, 1332–1338.

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature Neuroscience*, *16*, 1170–8.

Radonjic, A., Cottaris, N. P., & Brainard, D. H. (2015). Color constancy supports cross-illumination color selection. *Journal of Vision*, *15*, 1–19.

Reichert, D. P., Seriès, P., & Storkey, A. J. (2013). Charles Bonnet Syndrome: Evidence for a Generative Model in the Cortex? *PLoS Computational Biology*, *9*, e1003134.

Rich, A. N., Williams, M. A., Puce, A., Syngeniotis, A., Howard, M. A., McGlone, F., & Mattingley, J. B. (2006). Neural correlates of imagined and synaesthetic colours. *Neuropsychologia*, *44*, 2918–25.

Rodieck, R., Binmoeller, K., & Dineen, J. (1985). Parasol and midget ganglion cells of the human retina. *The Journal of Comparative Neurology*, *233*, 115–132.

Rüttiger, L., Braun, D. I., Gegenfurtner, K. R., Petersen, D., Schönle, P., & Sharpe, L. T. (1999). Selective color constancy deficits after circumscribed unilateral brain lesions. *Journal of Neuroscience*, *19*, 3094–106.

Schein, S. J. & Desimone, R. (1990). Spectral properties of V4 neurons in the macaque. *The Journal of Neuroscience*, *10*, 3369–89.

Schiller, P. H., Logothetis, N. K., & Charles, E. R. (1990). Functions of the colour-opponent and broad-band channels of the visual system. *Nature*, *343*, 68–70.

Schneider, K. A. (2011). Subcortical mechanisms of feature-based attention. *The Journal of Neuroscience*, *31*, 8643–53.

Seymour, K. J., Clifford, C. W. G., Logothetis, N. K., & Bartels, A. (2009). The coding of color, motion, and their conjunction in the human visual cortex. *Current Biology*, *19*, 177–83.

Seymour, K. J., Clifford, C. W. G., Logothetis, N. K., & Bartels, A. (2010). Coding and binding of color and form in visual cortex. *Cerebral Cortex*, *20*, 1946–54.

Seymour, K. J., Williams, M. A., & Rich, A. N. (2015). The Representation of Color across the Human Visual Cortex: Distinguishing Chromatic Signals Contributing to Object Form Versus Surface Color. *Cerebral Cortex*, 1–9.

Shapley, R. & Hawken, M. J. (2011). Color in the cortex: single- and double-opponent cells. *Vision Research*, *51*, 701–17.

Shevell, S. K. & Kingdom, F. A. A. (2008). Color in complex scenes. *Annual Review of Psychology*, *59*, 143–66.

Shipp, S., Adams, D. L., Moutoussis, K., & Zeki, S. (2009). Feature binding in the feedback layers of area V2. *Cerebral Cortex*, *19*, 2230–2239.

Shipp, S. & Zeki, S. (1985). Segregation of pathways leading from area V2 to areas V4 and V5 of macaque monkey visual cortex. *Nature*, *315*, 322–5.

Shipp, S. & Zeki, S. (2002). The functional organization of area V2, I: specialization across stripes and layers. *Visual Neuroscience*, *19*, 187–210.

Sincich, L. C., Jocson, C. M., & Horton, J. C. (2007). Neurons in V1 patch columns project to V2 thin stripes. *Cerebral Cortex*, *17*, 935–941.

Sincich, L. C., Jocson, C. M., & Horton, J. C. (2010). V1 Interpatch Projections to V2 Thick Stripes and Pale Stripes. *The Journal of Neuroscience*, *30*, 6963–6974.

Solomon, S. G. & Lennie, P. (2007). The machinery of colour vision. *Nature Reviews. Neuroscience*, *8*, 276–86.

Solomon, S. G., Peirce, J. W., & Lennie, P. (2004). The impact of suppressive surrounds on chromatic properties of cortical neurons. *Journal of Neuroscience*, *24*, 148–60.

Song, J.-H., Rowland, J., McPeek, R. M., & Wade, A. R. (2011). Attentional Modulation of fMRI Responses in Human V1 Is Consistent with Distinct Spatial Maps for Chromatically Defined Orientation and Contrast. *Journal of Neuroscience*, *31*, 12900–12905.

Stockman, A. & Sharpe, L. T. (2000). The spectral sensitivities of the middle- and long-wavelength-sensitive cones derived from measurements in observers of known genotype. *Vision Research*, *40*, 1711–1737.

Stoughton, C. M. & Conway, B. R. (2008). Neural basis for unique hues. *Current Biology*, *18*, R698–9.

Szmajda, B. A., Buzás, P., Fitzgibbon, T., & Martin, P. R. (2006). Geniculocortical relay of blue-off signals in the primate visual system. *Proceedings of the National Academy of Sciences*, *103*, 19512–7.

Tanigawa, H., Lu, H. D., & Roe, A. W. (2010). Functional organization for color and orientation in macaque V4. *Nature Neuroscience*, 1–8.

Tootell, R. B. H., Nelissen, K., Vanduffel, W., & Orban, G. A. (2004). Search for Color 'Center(s)' in Macaque Visual Cortex. *Cerebral Cortex*, *14*, 353–363.

Ts'o, D. Y. & Gilbert, C. D. (1988). The organization of chromatic and spatial interactions in the primate striate cortex. *Journal of Neuroscience*, *8*, 1712–27.

Wachtler, T., Sejnowski, T. J., & Albright, T. D. (2003). Representation of color stimuli in awake macaque primary visual cortex. *Neuron*, *37*, 681–691.

Wade, A. R., Augath, M., Logothetis, N. K., & Wandell, B. A. (2008). fMRI measurements of color in macaque and human. *Journal of Vision*, *8*, 6.1–19.

Wade, A. R., Brewer, A. A., Rieger, J. W., & Wandell, B. A. (2002). Functional measurements of human ventral occipital cortex: retinotopy and colour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *357*, 963–973.

Ware, C. & Cowan, W. B. (1982). Changes in perceived color due to chromatic interactions. *Vision Research*, *22*, 1353–1362.

Westheimer, G. (2008). Was Helmholtz a Bayesian? *Perception*, *37*, 642–650.

Winawer, J., Horiguchi, H., Sayres, R. A., Amano, K., & Wandell, B. A. (2010). Mapping hV4 and ventral occipital cortex: The venous eclipse. *Journal of Vision*, *10*, 1–1.

Witzel, C., Valkova, H., Hansen, T., & Gegenfurtner, K. R. (2011). Object knowledge modulates colour appearance. *i-Perception*, *2*, 13–49.

Xiao, Y. & Felleman, D. J. (2004). Projections from primary visual cortex to cytochrome oxidase thin stripes and interstripes of macaque visual area 2. *Proceedings of the National Academy of Sciences*, *101*, 7147–7151.

Xiao, Y., Wang, Y., & Felleman, D. J. (2003). A spatially organized representation of colour in macaque cortical area V2. *Nature*, *421*, 535–9.

Young, T. (1802). On the theory of light and colors. *Philosophical Transactions of the Royal Society*, *91*, 12–49.

Yuille, A. & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, *10*, 301–8.

Zeki, S. (1980). The representation of colours in the cerebral cortex. *Nature*, *284*, 412–418.

Zeki, S. (1983). Colour coding in the cerebral cortex: the reaction of cells in monkey visual cortex to wavelengths and colours. *Neuroscience*, *9*, 741–65.

Zeki, S. (1990). A century of cerebral achromatopsia. *Brain*, *113*, 1721–77.

Zeki, S. (1996). Are Areas TEO and PIT of Monkey Visual Cortex Wholly Distinct from the Fourth Visual Complex (V4 Complex)? *Proceedings of the Royal Society B: Biological Sciences*, *263*, 1539–1544.

Zhang, P., Zhou, H., Wen, W., & He, S. (2015). Layer-specific response properties of the human lateral geniculate nucleus and superior colliculus. *NeuroImage*, *111*, 159–166.

Zhang, X., Qiu, J., Zhang, Y., Han, S., & Fang, F. (2014). Misbinding of color and motion in human visual cortex. *Current Biology*, *24*, 1354–1360.

Picture Credits

Figure 3 on page 10 was modified from Bizhan Sharopov's work, https://commons.wikimedia.org/wiki/File:%D0%97%D0%BE%D1%80%D0%BE%D0%B2%D0%B8%D0%B9_%D1%88%D0%BB%D1%8F%D1%85.jpg, published under "CC BY-SA 3.0".