

STEVEN HORST

# MODELING, LOCALIZATION AND THE EXPLANATION OF PHENOMENAL PROPERTIES: PHILOSOPHY AND THE COGNITIVE SCIENCES AT THE BEGINNING OF THE MILLENNIUM

ABSTRACT. Case studies in the psychophysics, modeling and localization of human vision are presented as an example of "hands-on" philosophy of the cognitive sciences. These studies also yield important results for familiar problems in philosophy of mind: the explanatory gap surrounding phenomenological feels is not closed by the kinds of investigations surveyed. However, the science is able to explain some sorts of phenomenological facts, such as why the human color space takes the form of the Munsell color solid, or why there is a phenomenologically-pure yellow but not a phenomenologically-pure orange.

## 1. INTRODUCTION: PHILOSOPHY OF COGNITION AT THE TURN OF THE MILLENIUM

Future historians of philosophy may very well see important parallels between the 20th century and the 17th.[1] In both centuries, philosophers were much engaged with the philosophical understanding of new scientific developments. At the midpoint of each century, the most influential philosophical view of science was modeled upon deduction and construction in logic or mathematics; and there were grand philosophical ambitions, motivated in no small measure upon *a prioristic* grounds, for a unification of all knowledge in the form of something like a single axiomatic-deductive system. And by the end of the century, these ambitions were challenged on two important fronts, in spite of (and in part as a result of) actual scientific progress. On the one hand, both 17th century rationalist and 20th century positivist models of science became problematic for their apparent unsuitedness to incorporate the things we count as most distinctive about ourselves: consciousness, freedom, moral and semantic normativity. And on the other hand, actual scientific progress often took forms that looked significantly unlike the kind of

Synthese (2005) 147: 477–513
DOI 10.1007/s11229-005-8365-5

deductive system envisioned by Descartes and Hobbbes, or by Carnap and Nagel. In the late 17th century and into the 18th, Newton's overthrow of contact mechanism and his cryptic remark "*hypotheses non fingo*" on the question of the nature of gravitational force led many of his admirers to an even broader rejection of the reductionist program in favor of a view of science that concentrated upon finding quantitative laws that describe the phenomena and are useful for prediction and control without seeking for hidden causes. In the late 20th century, "naturalistic" philosophers of science followed the trail blazed by the history of science movement in the 1960s in concentrating upon describing the actual variety of explanatory strategies found in a number of sciences, particularly a number of comparatively recent entries in the life sciences, rather than trying to force them into a preconceived framework imported from logic and mathematics.

We are now in a period that has variously been heralded as the "decade of the brain" or "century of neuroscience". While writers as early as Newton's friend John Locke aspired to become "Newtons of the mind" by producing a science of mind upon Newtonian principles, it is really only much more recently that anyone can seriously lay claim to successfully laying down even the beginnings of such a project. The sciences of cognition have recently made rapid gains along a number of fronts: the physiology and functional anatomy of the brain, beginning with investigations like the study of brains of trauma patients by Broca in the 19th century and up through modern imaging techniques; the more fine-grained studies of neuroanatomy, both at the level of single-cell function and anatomy, and of the connections between networks of cells; the collection and rigorous systematization of psychophysical data stemming from the works of Weber and Fechner; and the development of specialized mathematical and computational modeling techniques. It would be an exaggeration to say that anyone has yet made the kinds of fundamental and unifying breakthroughs in the cognitive sciences that Newton is credited with in mechanics. But we might at least be at a point comparable with that of physics and celestial mechanics in the early 17th century.

It is, as a result, both an exciting and a trying time to be a philosopher of mind/psychology/neuroscience. It is trying and exciting, not only because of the daunting task of trying to keep up with the wealth of new discoveries about mind and brain, but also because the explosion of knowledge in the sciences of cognition

raises questions about the very nature of the philosopher's role in understanding the mind.

One question that looms large is that of how the philosophy of the cognitive sciences should be in dialog with the actual sciences themselves on the one hand, and with more traditional and mainstream philosophical problems in epistemology and metaphysics, like Levine's (1983) explanatory gap on the other. My own view is that it should make every effort to be in significant dialog with both. On the one hand, it is important to be guided by our best understanding of what explanations of mind and brain actually look like, and not by some armchair notion, such as that they should look like mathematical deductions in an axiomatic system. Moreover, we should be attentive to developments elsewhere in philosophy of science to see how the sciences of cognition are similar to or different from other sciences such as physics, chemistry and biology. But on the other hand, the cognitive sciences are comparatively young and, compared with physics, quite immature. The dominant paradigm changes in the space of years or decades. (When I was an undergraduate, there was almost no support available for research in neural network modeling, for example.) So philosophers will need to take a long-term perspective rather than assume that the latest big news will have lasting status. Moreover, until proven otherwise, we would do well to think that long-lived philosophical problems, such as Descartes' "real distinction" between mind and body, Levine's "explanatory gap" and Chalmers' "hard problem of consciousness", may have some real and lasting intellectual bite.

## 2. THREE APPROACHES TO THE PHILOSOPHY OF MIND AND COGNITION

Clark Glymour has pointed out, in a review of Jaegwon Kim's (1998) *Mind in a Physical World*, that philosophers of mind/psychology/neuroscience tend to fall into two camps. One camp (in which he locates Kim's book) prefers to treat metaphysical problems about the mind in isolation from the sciences of mind. The other (to which he himself subscribes) approaches philosophy of mind/psychology/neuroscience as a kind of running dialog with, and commentary upon, the sciences of cognition. While there is a wide spectrum of attitudes lying between Kim's and Glymour's, in practice there is in fact a kind of division among philosophers around what group of problems they are interested in addressing.

Using Chalmers' (1996) terminology, some are interested in "the hard problems" of consciousness: epistemic and metaphysical issues about whether phenomena like consciousness and intentionality are metaphysically supervenient upon brain states. Others are interested in what Chalmers calls the "easy problems" – which of course Chalmers himself admits are not easy except by comparison – that are more continuous with empirical and theoretical work in the sciences, in the spirit of "naturalistic" philosophy of science. This can easily lead to a stark conclusion: the natural sciences, including neuroscience, can tell us a great deal – perhaps everything – about the structure and function of mind and brain, but can tell us nothing about the phenomenological properties of experience. Some empirically minded philosophers and philosophically minded scientists are happy to embrace this conclusion, and treat questions about phenomenology as uninteresting, irrelevant, or even philosophical illusions. Others may be inclined to the opposite attitude: to view the explanatory gap itself as an illusion that can be cured by actual explanation in the cognitive sciences, perhaps even explanations that are already at hand.

We might problematize these attitudes by making two questions quite explicit:

(1) When we look at actual case studies in the explanation of mental states having a phenomenology – seeing colors, for example – do these explanations confirm or disconfirm the philosophical intuition that there is a principled and abiding explanatory gap?
(2) If such a gap remains after explanation, does this entail the conclusion that the sciences can explain *nothing* about phenomenology at all?

In this article, I hope to model a way of playing it down the middle, as it were. The sciences of cognition yield some very powerful explanations of features of cognition, *including features of its phenomenology*; but they do so in spite of leaving exactly the sort of explanatory gaps that have been claimed on the basis of philosophical thought-experiments. To motivate this conclusion, and to model how case studies in real explanations of the mind can help us see such issues more clearly, I will give an introductory-level overview of some fairly basic work in one sample area: the study of vision, particularly color vision in humans. On the basis of this, I will ask just *what* is explained thereby, and what *kind* of explanation is given. In this case, there are actually extremely impressive

and robust explanations of psychophysical data by properties of cells employed in *early* visual processing, out of which some of the "shape" of the psychophysical data simply "falls out". On the other hand, we are not in a comparable position to talk about exactly what is going on further down the neural pathways that process color information. I shall illustrate this with respect to the area of the visual cortex called V4, which seems to be particularly implicated in color vision, but it could be made with respect to other cortical areas as well. Nor, more fundamentally, are we able to explain the presence or character of visual qualia. There really does seem to be a robust explanatory gap just where it was argued on philosophical grounds, and examining the science seems to reinforce rather than belie this conclusion. However, this does not mean that neuroscience explains *nothing* about color vision, *even if we mean by that the qualitative space of color vision.*

I intend this article to be accessible to a wide range of readers, including on the one hand readers who are familiar with the classic philosophical articles on the explanatory gap by Levine (1983), Nagel (1974), Jackson (1982) and Chalmers (1996) but are not at all conversant with any actual explanations in psychology or neuroscience, and on the other hand readers who may only have heard of the explanatory gap but have never taken it seriously because it is usually cast in terms far removed from actual scientific explanation. I shall thus assume a background knowledge of the idea of an explanatory gap between mind and brain, but shall present a very elementary walk-through of several stages, historically arranged, of the understanding of color vision, and ask at each stage what is and is not explained. In the course of these I shall, on several occasions, stress how the explanatory gap is not crossed. The article thus practices a form of "naturalistic" philosophy of mind/psychology/neuroscience – working from real case studies in explanation rather than attempting to constrain scientific practice on aprioristic, extrascientific grounds – and then relates these to issues usually raised independently, in the context of aprioristic philosophical thought-experiments.

## 3. PRELIMINARIES: THE FEEDBACK CYCLE OF UNDERSTANDING MIND AND BRAIN

It is useful to begin with a general and schematic overview of the process of understanding mind and brain. Three important aspects

of this project are: psychophysics, localization, and mathematical modeling. *Psychophysics* (or more exactly, what Fechner called *outer* psychophysics) gathers data about the relationships between percepts and the stimuli that cause them. In the case of the modeling of perception, psychophysics supplies a goodly part of the data that theoretical modeling must explain. *Localization* is the process of finding areas (and perhaps global patterns of activity, though these are dubiously labeled "localized") that are specially implicated in particular perceptual or cognitive processes. Such data come from a variety of sources: study of the brains of patients who have lost particular cognitive functions due to injuries or strokes, various types of brain scans, single-cell studies of surgical patients, anatomical and lesion studies of animal models. If psychophysics tells us what needs to be explained in the study of perception, localization studies tell us something about where to look for the machinery that does the explaining. But of course "because something happened over there" is not yet a very good explanation. We need, additionally, a *model* that shows how the palpable properties of the data (e.g., the stimulus-to-percept curves of the psychophysical data) can be accounted for by neurally-reasonable assumptions about specific parts of the brain. This is the job of *mathematical modeling of the mind.* In practice, there is an ongoing and iterated cyclic relationship between these endeavors, as well as other more peripheral activities like attempts at implementing models in artificial agents. The rough form of this relationship is depicted in Figure 1.

## 4. THE STUDY OF COLOR VISION – A SELECTIVE HISTORY

Much of the modern problematic in color vision originated with Newton's discovery, with the help of a refracting prism, that white light is in fact composed of a mixture of colored lights. It was later discovered that the spectral colors are typified by different *wavelengths* of light, and that visible light is in fact just a small portion of the electromagnetic spectrum. As psychophysics and psychometrics gained momentum in the nineteenth century, researchers began to discover some familiar (and some not-so-familiar) facts about our perceptions of color. Among them:

- We are more sensitive to some wavelengths of light than to others. In general, we are more sensitive to light in the middle of the spectrum than at the ends. (This is actually slightly

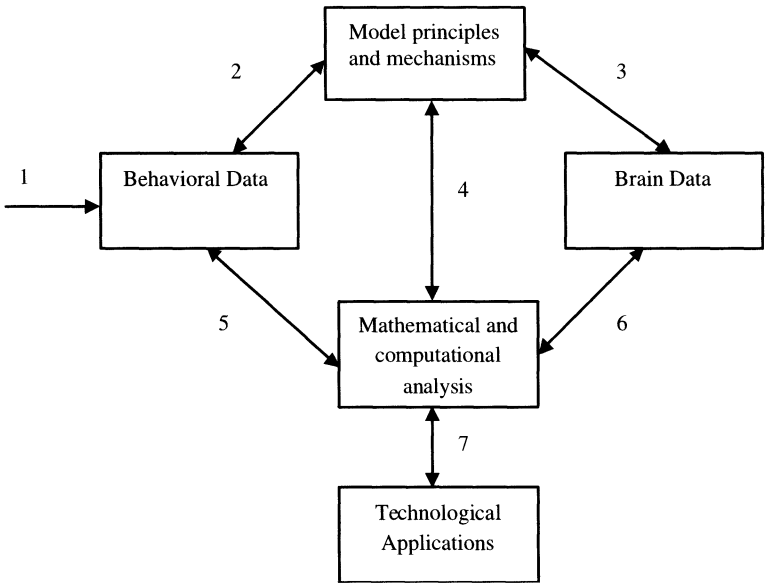


*Figure 1.* Flowchart of the theoretical process in cognitive modeling (derived from personal communication with Stephen Grossberg). Process begins with the collection of behavioral data (1). From this, theoreticians are able to derive new model principles and mechanisms (2) in the form of neural networks that explain brain data (3) in a new functional way. Once this connection is established, it is possible to work top-down from behavioral data and bottom-up from brain data to further refine the model principles and mechanisms in a continuing modeling cycle that explains its explanatory scope with successive theoretical cycles. Model principles and mechanisms get modeled through mathematical and computational analysis (4) which can generate data predictions for both behavioral data (5) and brain data (7). Finally, models can be tested through technological application (7).

complicated by the fact the different kinds of photoreceptors in the eye – the rods and cones – have different photoreceptive curves.) See Figure 2.

- The wavelengths to which we have greatest sensitivity change depending on whether our vision is dark-adapted. See Figure 3.
- Experienced color does not map neatly on to spectral color or wavelength. A pure wavelength in the yellow portion of the spectrum will produce an experience of yellow. But one can also produce an indistinguishable yellow sensation by carefully mixing red and green light. This is called *metameric matching*. In fact, Thomas Young (1773–1829) showed that the entire range of spectral light can be generated by combining three pure lights (e.g., red, green and blue) in different combinations.

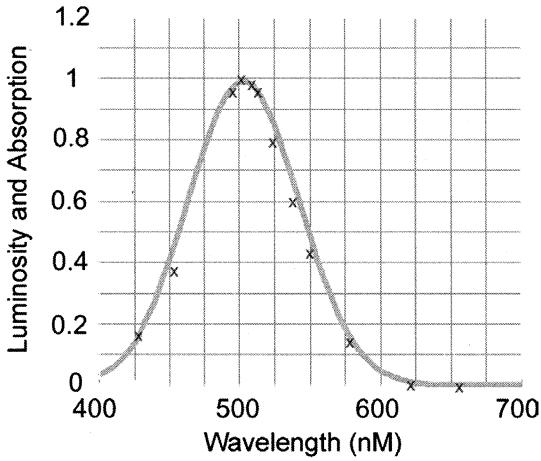


*Figure 2.* The chemical basis of vision. The curve represents the photosensitivity of the (dark-adapted) human eye. The gray *x*'s represent the amount of light absorbed at the same wavelengths by the pigment rhodopsin (based on Gregory 1978).

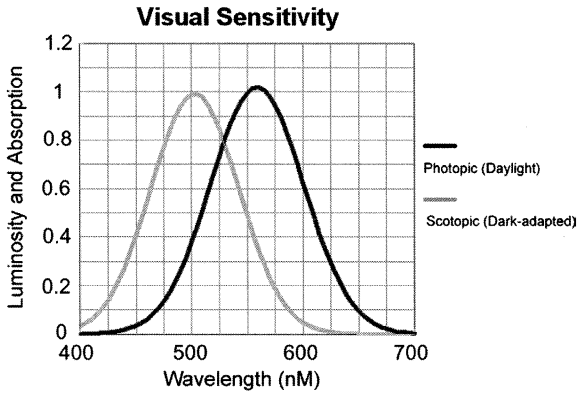


*Figure 3.* Photoreceptivity curves for daylight and dark-adapted vision.

There are not, however, three canonical colors that must be used for this effect – any three wavelengths suitably separated in the spectrum will do the trick. See Figure 4.

- The rules for mixing *lights* are different from those for mixing *pigments*.
- Not everything we see as a "color" can be produced by this three-color process. No mixture of pure spectral lights, for example, can produce browns. These only appeal amidst certain *context and contrast effects*, such as when an area has a

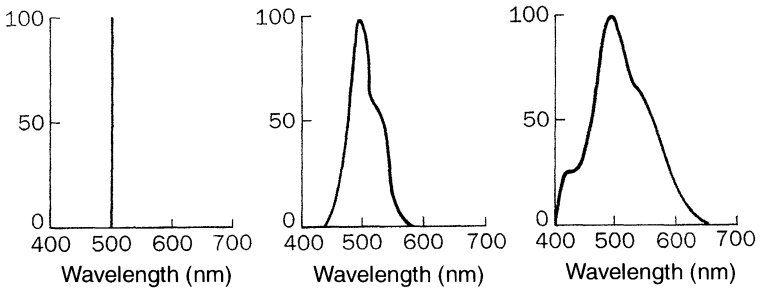


*Figure 4.* Light distributions and matameric matches. Each of these wavelength distributions will produce a sensation of a unique green – one that is not perceived as having any mixture of yellow or blue. From Hurvich (1981, p. 78), reproduced from Clark (1993, p. 43).

particular pigment and is surrounded by an area with a particular contrasting pigment.

- There are several kinds of colorblindness. Some people are unable to distinguish red and green, others blue and yellow. And a very small percentage of the population does not experience color at all, while being normally sighted in other respects.
- Sensitivity to color seems to differ greatly across species. Other than humans and perhaps some other primates, the mammals are largely colorblind, while many fish and birds seem to be very sensitive to color, and in some cases to portions of the electromagnetic spectrum beyond our own perception.
- Edwin Land (1959) noted that the perceived color of an area is sometimes affected by whether the area is construed as an object.

Some of these observations were systematized early on into explicit models. The "color space" of humans was modeled in the "color solid" of Munsell. (See Figure 5.) This kind of formal model is not *explanatory*, but can be viewed, like the curves of the Weber laws, as a systematization of data – the final stage of psychophysics.

On the basis of such information, visual theorists like Helmholtz (of physics fame) began to formulate *theoretical* models of how we might be able to see color. In an important respect, color vision turns out to be very different from how the auditory system enables us to hear tones. When two distinct tones are played together, we hear the result as a chord, and not as a pure tone. The reason for this is that in the inner ear, there are a very great number of transducers that respond to particular distinct tones. But there are

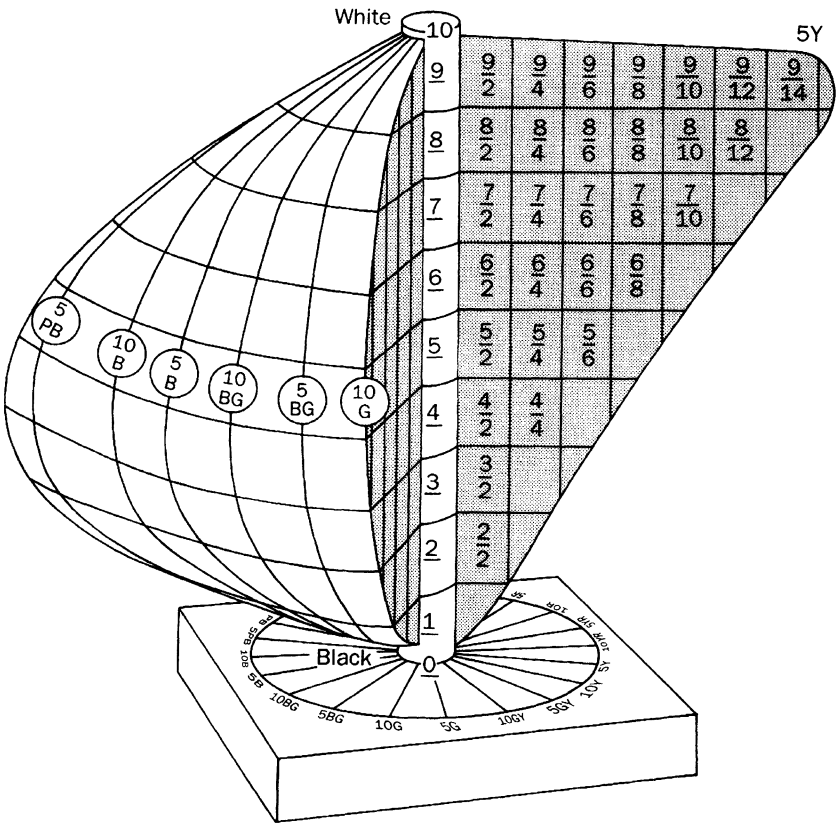


*Figure 5.* The Munsell color solid, a geometric representation of trichromatic human color space. A portion of the solid has been cut away to reveal the interior. From Hurvich (1981, p. 274). Reproduced from (Clark 1993, p. 122).

not hundreds of different kinds of color receptors in the eye, corresponding to the different hues that we can distinguish. And combinations of pure chromatic frequencies – say, a pure red and a pure green – are *not* perceived as visual "chords", but as a different color entirely, such as yellow. In hearing a chord; we hear two tones, in seeing a mixture of red and green light, we are aware of only one color, yellow. Helmholtz (1867) suggested that Young's evidence that any experienced color can be produced by the combination of three pure chromatic frequencies pointed to a color-detection mechanism that has three elements that are responsive to different frequencies. This *three-color process* model was originally presented as a theoretical model, without a localization in the nervous system, though of course the model does a great deal to guide research into the visual

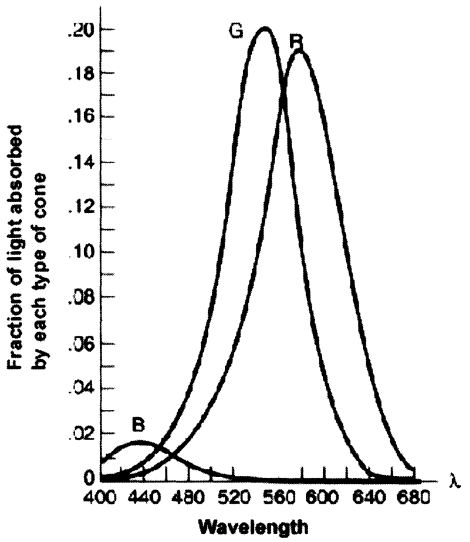


*Figure 6.* Response curves of the three cone systems in humans. The horizontal axis represents wavelength, the vertical axis the fraction of light absorbed by each type of cone.

system. It tells us what to look for if the theoretical model is to be confirmed: namely, something that can underwrite the functionality needed for the three-color theory.

The three-color theory proposed by Helmholtz turned out to have a straightforward neural substrate – i.e., a structure in the nervous system that matches its functionality. Investigation of the human retina reveals that most humans possess three kinds of cone cells, each of which is differentially sensitive to light of different wavelengths. In Figure 6, above, we see the response curves of the three different receptors. The vertical axis represents the fraction of light absorbed by each type of cone while the horizontal axis represents wavelength.

Each cone cell, taken alone, is "colorblind" – that is, it cannot discriminate between different colors. What the cone cell does is send spikes down the visual cascade. But it is not sensitive to just one wavelength of light: a little bit of yellow light, to which a particular cone cell responds strongly, will cause it to spike at a certain rate, but so will a larger amount of red light, to which it responds less strongly. So, if you are a cell at the other end of the optic nerve "listening" to that cone cell's output, you cannot tell whether it is "saying" that there is a little bit of yellow light or a lot of red light.

And of course, the problem is not limited to a choice between two different frequencies, but the entire range of the spectrum to which that cell responds.

However, mathematical investigation of the properties of the different types of cone cells reveals that the fact that we have more than one type of cone cell allows us to extract more chromatic information than we could with just a single type. In particular, chromatic information is encoded in the *differences* and *ratios* between the response curves of the different cone cells. Specifically, the ratios between curves will often determine a unique chromatic frequency, regardless of the intensity of the light. (See Table I.) The chromatic information carried in the ratios between the sensitivities of different cones is not perfect: there are still combinations of frequencies that can mimic a pure signal, for example, even in normally sighted people. But from a standpoint of psychological explanation this is *good*, as these ambiguities of the stimulus are also found in the psychophysical data, and you *want* your theoretical model to have the same idiosyncracies you find in the system you are attempting to model. Individual cone cells do not provide an explanation for how such information is extracted by the nervous system – this takes place at a later point in the visual cascade. But the facts that such encoding preserves information distributed over the cone system, and that it matches the peculiarities of the psychophysical data (e.g., that there are metameric matches just where the model shows information to be ambiguous), both confirms this part of the model and helps guide further research.

TABLE I

Sample data of the absorption percentages, differences and ratios of the M(edium) and L(ong wavelength) cone systems

| Wavelength (nm) | Quanta incident | Absorbed by M | Absorbed by L by L | Difference | Ratio |
|---|---|---|---|---|---|
| 560 | 1000 | 192 (19.2%) | 165 (16.5%) | 27 | 1.16:1 |
| 520 | 1000 | 165 (16.5%) | 62 (6.2%) | 103 | 2.66:1 |
| 560 | 3814 | 732 (19.2%) | 629 (16.5%) | 103 | 1.16:1 |

From Clark (1993, p. 34) Note that the ratio of M/L preserves information about wavelength regardless of intensity.

This simple neural system turns out to yield fairly strong explanations of a surprising amount of the psychophysical data about color vision. First, the sensitivity curves of the individual color receptors can be used to derive the psychophysical data of the different luminance-to-brightness functions of different portions of the spectrum. In Figure 2 earlier, for example, we see a curve plotted for the photoreceptivity of human vision over the spectrum. Across that curve are also data points representing the photoresponse of the chemical photoreceptor rhodopsin extracted from a frog's eye. These data points fall along the human photosenstivity curve, and the presence of rhodopsin in one kind of cone cell explains the photometric response of that type of cell.

Cell behavior also explains why a single internal state can be ambiguous between a number of different stimuli (e.g., light in the green band or a combination in blue in yellow bands): namely, because the information that is transmitted from the stimulus to the cone receptors literally is ambiguous between two (or more) possible environmental states that would produce the same effects in the eye. The firing of a cone cell is driven by the number of light quanta that it absorbs. And this, in turn, is determined by the product of (a) the number of quanta at a given wavelengh striking it, and (b) the percentage of quanta absorbed for that particular wavelength. Since a given receptor will absorb a greater portion of light at some frequencies than at others, adding more light of a less sensitive frequency will produce the same overall absorption as less light of a more sensitive frequency. You can also add wavelengths algebraically to obtain matches, as shown in Table II below. It is, in fact, metaphysically necessary that a system thus configured would result in ambiguities exactly where such ambiguities have discovered by psychophysicists.

Finally, physiological data about the cone cells explain the phenomenon of colorblindness, at least if we interpret "colorblindness" to mean an absence of sensitivity to chromatic information. Normally sighted individuals have three different types of cone cells. But a small portion of the population has only two. As a result, the cone system of these dichromats carries less information about the chromatic features of their environment, and there is a greater range of stimuli that they are unable to distinguish. Depending upon which cells are missing, they will either be unable to distinguish reds from greens or be unable to distinguish blues from yellows. The very rare individuals who have only one kind of cone

cell, or are lacking in them altogether, are "monochromatic" – they do not distinguish colors at all, but only differences in brightness. All of these phenomena can be *generated* from the neural model, through largely mathematical techniques.

Are we, then, licensed to say that we have found a localization of color qualia in the cone cell system of the eye? Tempting though this may seem, the answer has to be *no*. First, there are a wide variety of psychophysical data about color perception that are not explained by properties of the cone system, and indeed which need to be idealized away from for the cone system to explain the properties it does explain. For example, we have already mentioned the problem of the non-spectral colors, such as brown. There are also several kinds of idealizations that have been made in the above explanation: for example, that the results apply only to stimuli presented in the central 4 degrees of the visual field, phenomena that arise when contrasting colors are set alongside one another, dark and light-adaptation, and the effect of priming the eye with one color before exposing it to another. Some of these effects can themselves be predicted from the particulars of the cone cells: e.g., cells communicate through chemical neurotransmitters, and sustained activity of one receptor (say, the long-frequency cones) can lower its level of transmitter ions, so that its response to a new stimulus will be proportionally decreased relative to the others (and hence the hue will seem off). But others (like contrast effects) cannot be explained simply by appeal to the cone system. These may be explained by other features of the visual system, but they have

TABLE II

Metameric matches and absorption of light by M and L cone cells. Predictions of a Match

| Wave-length (nm) | Quanta incident (count) | Percent absorbed by M | Quanta absorbed by M | Percent absorbed by L | Quanta absorbed by L |
|---|---|---|---|---|---|
| 560 | 1576 | 19.7 | 310 | 16.5 | 260 |
| 515 | 2100 | 13.1 | 275 | 5.7 | 120 |
| + | | | + | | + |
| 615 | 1166 | 3.0 | 35 | 12.0 | 140 |
| Total | | | 310 | | 310 |

The combinations of wavelengths in the second and third rows will match the stimulus in the first row, producing the same number of absorptions in both the L and the M systems. From (Clark 1993, p. 39).

not been explained in what we have said so far, and the fact that our explanations idealize away from features that matter *in vivo* is a fact we should note well.

Second, the fact that the cone cells are responsible for chromatic discrimination does not entail that they are the part of the nervous system specially associated with color *experiences*. Many people, for example, can dream and visualize in color, even though this is not caused by retinal stimulation. Indeed, if the optic nerve is severed or the eyes are lost, such dreams and visualization need not be immediately affected, while stimulation to the retina will cease to cause color qualia. (And conversely, surgically removed eyes presumably experience no color on their own, even while cells are still able to fire.) If we are to look for a special location associated with all color experiences, we will have to look deeper in the brain. Finally, this the kind of explanation we have presented thus far does not explain qualia as such at all: what it does is to *presume* qualia as one of the relata of the psychophysical data, and then to explain facts *about* the quality-space such as the "shape" of the color solid. There is, to be sure, a more complete explanation of color *discrimination capacities*. But why these should be accompanied either by their *particular* qualitative counterparts, or indeed by *any* qualia at all, has gone unaddressed. A Martian scientist, unacquainted with qualia, investigating the human visual system, could derive psychophysical phenomena like metameric matches with complete assurance from its knowledge of the functional anatomy of the visual system. But nothing in this stage of visual processing would remotely suggest to it the conclusion that humans experience visual qualia.

## 5. COMPLICATING THE MODEL: COLOR OPPONENCY

While the properties of the cone system explain a surprising amount of the psychophysical data, there is also a great deal that they do not explain. One problem was noted very early on by Hering (1878): As one might predict on the basis of the Young–Helmholtz three-color theory, we can perceive a "pure red" – a red in which we experience no admixture of yellow or blue – and likewise a "pure blue" and a "pure green". And, again consistent with this theory, we perceive many hues as "mixed" – aqua, for example, as a mixture of blue and green. However, Hering points out, we also experience yellow as a pure color – one in which we do not perceive red and

green components, as predicted by the theory. The *sensation* of yellow seems to be simple or unmixed. In fact, it seems *impossible* to experience a hue that would be described as a "reddish green" or a "bluish yellow".

Hering's interpretation of this was not that the Young–Helmholtz three-receptor theory was wrong, but that it was not the whole story. He suggested that the basis for hue sensations lies in a process in which there is *competition* between red and green and between blue and yellow. He proposed that all of our sensations of color can be accounted for by combinations of these "primary" hues, arranged along two axes generated by the opponent processes: a red–green axis and a yellow–blue axis. Hering's theory, known as the "*opponent process theory*," has gained widespread acceptance, and has itself found a neural correlate early in the visual system. Early visual processing in the eye turns out to have a number of stages prior to the transmission of information to the brain by way of the optic nerve. Information in the cone system is influenced by horizontal cells, bipolar cells, and retinal ganglion cells before it passes down the optic nerve. (See Figure 7.) In the ganglion cells (with the help of the horizontal and bipolar cells) we find the kind of opponent process postulated in Hering's theoretical model. The connections between ganglion cells and cones are in what is called a *center-surround architecture*, an architecture that is also found in many other parts of the nervous system.

In a center-surround architecture, we are dealing with the relations between two layers of cells, L1 and L2. In Figure 8, X is a sample cell in L2. X samples (has inputs from) a great number of cells in L1, covering an area of its surface. The connections are of two types. In the center of X's receptive field, there are *excitatory* connections: when cells in this area of L1 are activated, they "excite" X, which is to say they raise its likelihood of spiking. The other connections, in the periphery of X's receptive field, behave in just the opposite way: when cells in the periphery of the field are firing, they *inhibit* X, or make it less likely that it will fire. The question of whether X will in fact fire is then governed by a weighted summation of excitatory and inhibitory inputs. In this particular example, the center of the field is excitatory (or ON) and the periphery is inhibitory (or OFF). This kind of center-surround structure is called "ON-center, OFF-surround." But the center-surround architecture can take other forms as well: it can have an inhibitory center and an excitatory periphery (OFF-center, ON-surround), or if the

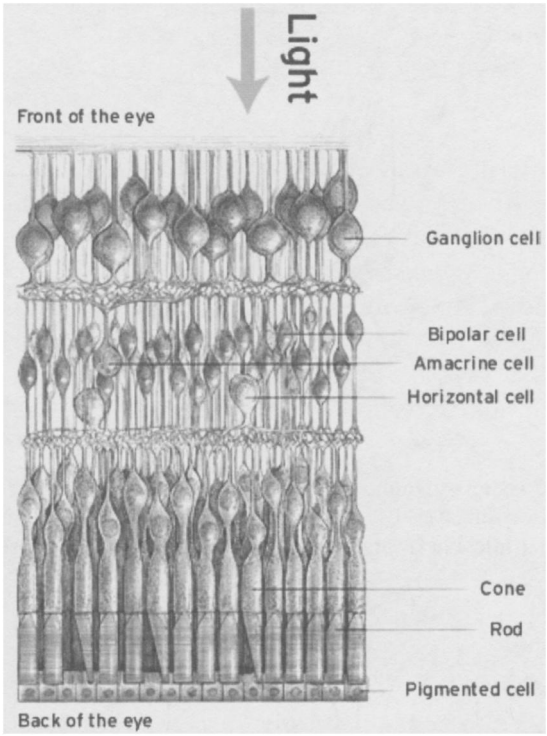


*Figure 7.* Layers of cells in the retina. Light passes through several layers of cells before it is detected by rod and cone cells, located at the back of the eye. These pass information on to further layers of processors, such as the horizontal, amacrine, bipolar and ganglion cells.

center and periphery are different *kinds* of cells, the architecture can also be implemented as an ON/ON or an OFF/OFF function.

Center-surround architectures are extremely useful. They are, among other things, the basis for detecting the contrasts of light and dark that signal edges and boundaries. Figure 9 illustrates organizations that compare information from the different kinds of cone cells, here designated by the wavelengths they respond to: S(hort), M(edium) and L(ong). The center-surround cells are of three types, (only two of which are shown in Figure 9). The first type involves opponency between the M and L cones. The greater number of such cells have excitatory centers, either in the M or L systems, though some also have inhibitory centers. Such cells have peaks for both excitatory and inhibitory responses and are called "red-green opponent" cells. The second type of cell compares the S and the combi-

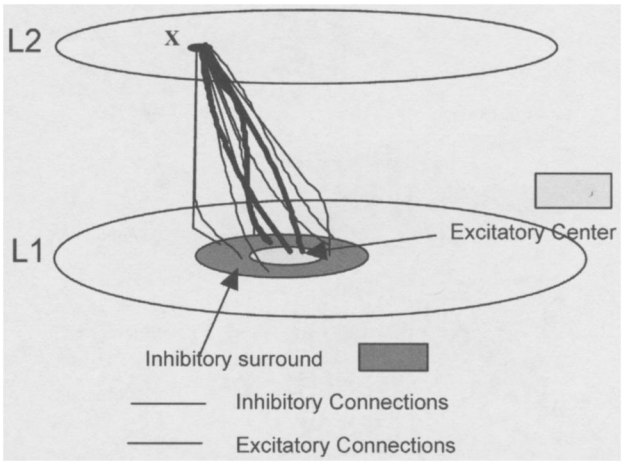


*Figure 8.* Basic center-surround diagram. A cell X in layer L2 has connections to cells in a region of layer L1. X is excited by activity of cells in the center of the region in L1 (bold lines), and inhibited by the activity of those that surround it (lighter lines).

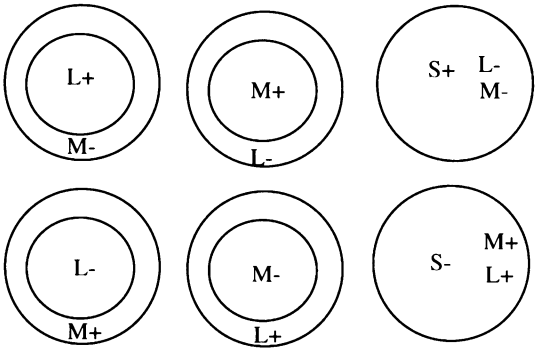


*Figure 9.* Spatial structure and frequency of incidences of the six most common varieties of color-opponent ganglion cells.

nation of the M and L functions. This process is called "yellow-blue opponent." A third type of ganglion cell does not appear to make chromatic distinctions, but follows the photopic luminosity function, and hence seems to code brightness and darkness. (See Figure 10.)

In these first two types of cells we have a neurological basis for the color opponency called for by Hering. Again, a number of psychophysical data can be predicted or demonstrated from the model, such as the fact that there is a phenomenologically pure yellow but not, say, a phenomenologically pure orange. The spectral luminance

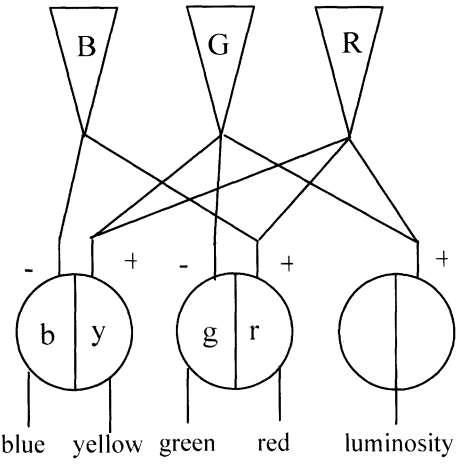


*Figure 10.* Color opponency in the ganglion cells. Inputs from the three types of cone cells interact competitively to activate ganglion cells, producing four chromatic channels (blue, yellow, red and green) and one for luminosity.

contrast sensitivities of the antagonist architectures also produce a curve that approximates the data for human chromatic sensitivity for test spots on a white background.

Historically, both Helmholtz and Hermann suggested theoretical models that could explain particular psychophysical data before the neural localization for the model was known. The data provided a formal "shape" that the underlying mechanism needed to explain. The theoretical model showed a structure capable of producing that formal shape. And later investigations into cell physiology revealed candidates that had the requisite properties located at a plausible point in the visual cascade. (In study of higher cognition, we are often in the opposite position: things like trauma studies reveal candidates for the gross localization of a capacity such as speech comprehension or face recognition before we have a formal model of how such functions might be achieved.)

But we must be very careful to specify what we have explained thus far. We have explained the color opponency phenomena postulated by Hering. But are we now in a position to localize color vision entirely in the retinal ganglion cells, or perhaps in their systemmatic cooperation with the cone system? The answer is still *no*, and for the very same reasons as before: First, there are still purely empirical phenomena that are not explained at this level. Notably, the contrast effects that produce the nonspectral hues, and inter-

actions with cues of depth and object boundaries noted by Edwin Land (1959) are not yet explained. Second, our localization thus far is confined to the retina, and this seems the wrong place to localize color qualia, since these can occur during dreams and visualization, which can persist after loss of one or both eyes, while stimulation of live retinal cells will not cause qualia (or, for that matter, discriminative abilities) if the optic nerve or the geniculate body are too damaged to carry information. And finally, our explanation is still in the business of explaining things *about* qualitative states (e.g., why there are more "pure" colors than kinds of cone cells, and why there is not a hue that is experienced as a greenish red) without explaining their qualitative character in its own right. The first two issues – accounting for additional data – can be pushed farther if we follow visual information further down the perceptual cascade. The third will remain intractable, but we will save the more careful examination of it until we have pushed as far as we can.

## 6. FROM EYE TO BRAIN

Among the more surprising 20th century discoveries about color vision was Edwin Land's claim that there are color vision effects that are not dependent simply upon the properties of the visual field, but also upon whether patches of color are interpreted as objects. This strongly suggests that there is feedback from whatever system(s) in the brain play a role in object groupings to some point in the causal stream that eventuates in color perception. And since there is no such feedback to the retina, we apparently need to look further into the brain before we are done with our localization of color sensation.

The ganglion cells in the retina are connected to the brain by the optic nerve. (Indeed, some are inclined to view the retina as a part of the brain that happens to extend into the eye, but the difference is not important for our purposes.) The signal passes through the *optical chiasm*, where signals from the left side of the visual field in both eyes are routed to the right side of the brain, and signals from the right side of both visual fields to the left side of the brain. These connections project (provide input to) a small body called the lateral geniculate nucleus (LGN), and from there to the visual cortex, located at the back of the brain. There are also feedback projections from parts of the cortex to the LGN, and indeed it seems to be the

general rule that when there is a projection from one part of the brain A to another part B, there are usually feedback channels from B to A as well.

The visual cortex (also called the striate cortex) is an area of brain in which much of our visual processing takes place. It is divided into areas, V1–V5, and each of these areas is itself divided internally into layers. Projections from the LGN enter V1 in the middle layers. Past that point, visual information seems to divide itself into three different streams: one for color, one for shape, and one for movement, location and spatial relations. The visual cortex also projects to other parts of the cortex that seem to be involved in yet more complex functions. Studies by Mishkin and associates (e.g. Ungerleider and Mishkin 1982) suggest that information from the visual cortex splits into two further streams. A dorsal stream (one projecting to the top of the brain) goes into the parietal lobe and seems to be responsible for perception of location and orientation to objects. This is sometimes called the "where stream." Patients who have had damage to these parts of the brain are often able to identify objects, but unable to grasp them properly, or report on their spatial relations. A second, ventral stream (one that projects to the underside of the brain) goes to the temporal lobe, and seems to be responsible for various sorts of recognition of objects. This is sometimes called the "what stream." One sub–area of the temporal lobe, in both monkeys and humans, seems to have the highly specialized function of recognizing faces of conspecifics. It was a portion of this "what" area that was damaged in Oliver Sacks's (1985) famous "man who mistook his wife for a hat", who was as a consequence unable to identify faces – in this extreme case, the patient was indeed not only unable to tell one face from another, but even unable to recognize a face as a face. (Here we have a case where the localization has preceded the theoretical model.) The features of visual information flow are illustrated in Figure 11.

In the case of color vision, however, when we pass beyond the optic chiasm, the type and quality of our explanations begin to change. In contrast to the strikingly rich, exact, quantitative and generative explanations of a number of psychophysical data that came from looking at two cell groups in the back of the eye, we at present know only that certain regions of the brain are selectively sensitive to chromatic data, but do not know just what they do or how they do it. We know, for example, that the cortical area V4 seems to be involved in higher processing of color information,

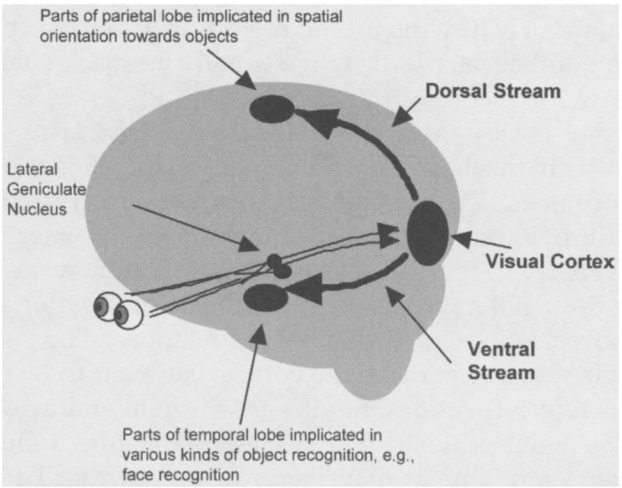


*Figure 11.* Schematic diagram of the flow of visual information. Information passes from the retina through the LGN to the visual cortex, from which it splits into ventral "where" and dorsal "what" streams.

and that there seems to be a color pathway that passes through the parvocellular areas of the LGN into the blob cells in V1 and from there into the thin stripes of V2, which then project to V4. (See Figure 12.)

What happens in these areas is still a matter of speculation. And there are, indeed, difficult methodological problems in proceeding further here. Whereas the explanations of psychophysical data that could be read off the responses of cones and retinal ganglion cells could be determined from examinations of single cells (or at worst from the feed-forward behavior of single cells and their projections), higher cortical activity seems to be typified by more highly *global* behavior, involving complicated feedback relations, both in the form of competition between cells at a single level and in the form of resonance phenomena between populations of cells in different systems (say, the LGN and particular areas of the visual cortex; Grossberg 1987). Moreover, in many useful models, the cortical encoding of information does not take place in single cells at all, but in activity patterns distributed over groups of cells, or in the connection strengths between them. Indeed, Land's discoveries suggest that there is important interaction between the color system and systems for shape and object recognition, so that it may be impossible to adequately model color vision just by under-

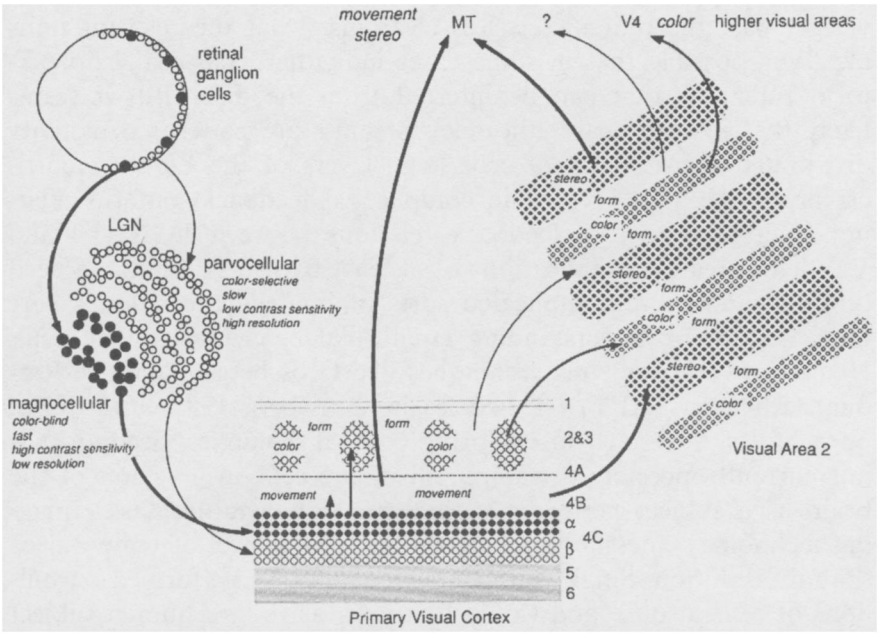


*Figure 12.* Schematic diagram of the flow of visual information through layers of LGN and visual cortex. Figure 23 and accompanying text from Spillman and Werner (1990, p. 195).

standing the so-called "color pathways". Because of the complexity of cortical structures and feedback relations among them, it is unlikely that the explanation of psychological phenomena residing in the cortex will be so closely linked to cell physiology as it was in the retina. Because of these problems, the *distance* between our formal models and our neurophysiology is far greater in the cortex than it is in the retina. As a result, it is far more difficult to test the neural plausibility of rival models. Likewise, because of the somewhat opaque, distributed nature of the coding of neural networks, it is difficult to guess or verify what functional task a cortical module is performing when we do not even know the units of the "code".

Here we arguably have an important issue for the philosophy of neuroscience. In early vision (processing within retinal cells), the relevant units for explaining the psychophysical data are localized in specific cells that can studied in isolation, much as one might study a particular mechanical structure or a simple electronic circuit. And indeed, it is the structural, chemical and electrical properties of individual cells that do much of the explaining. But once

we get past the retina, it is less clear just what the relevant units are. It is possible that in some cases individual cells really do perform functions that can be inferred from the data. But it seems likely that in many cases the relevant units are patterns of activity distributed across fields of cells (e.g., layers of the LGN or cortical areas like V4), or even in complicated feedback patterns relating several areas (e.g., feedback relations between LGN, V4 and V2). This creates complications of at least two sorts. First, we need different and more complicated sorts of modeling techniques here than we do for understanding circuit-like behavior in cone cells. Hence network modeling techniques like Grossberg's Adaptive Resonance Theory (ART) (cf. articles in Grossberg 1987) have taken on a life of their own in exploring cortical dynamics. Second, it is not currently possible to sample all of the cells in a region of the brain as a subject performs a perceptual or cognitive task. Imaging technology does not provide the necessary level of temporal or spatial resolution. Single-cell sampling cannot be performed on millions of cells at once and would be too invasive in a human subject in any case. And EEGs, which do provide good global information with a high degree of temporal resolution, do not provide the spatial resolution necessary to distinguish spatially distributed patterns of activity within a particular region. These are limitations of our current experimental technology which may or may not be insuperable. Moreover, on at least some possible scenarios – such as that it is patterns of activity in cortical areas that are the significant units, rather than single-cell activations – these limitations might keep us from being able to discern the physical properties that correspond to the significant units. As a result, for at least some types of problems, modeling of cortical dynamics must often proceed at a fairly global level, in abstraction from the details of the implementing system. Nor is this really comparable to the software-level vs. hardware-level distinction in a digital computer. There we might know that a number is represented by some pattern of bits in some discrete location, even if we do not know if it is represented by 8, 16 or 32 bits, or whether these are implemented in vacuum tubes, transistors, or integrated circuit boards, or in a Pentium 3 or G4 chip. With the brain, we do not know whether the significant units are localized in discrete areas or are patterns distributed over a population of cells. (E.g., the encoding of two perceptual data or two concepts may be implemented in the very same population of cells, and

factorable through some kind of vector algebra rather than stored in separate cells.)

## 7. PROJECTING A MORE COMPLETED MODEL

Unlike lab scientists, philosophers and theoreticians can take the liberty of projecting what things might look like if we were to transcend our current experimental limitations. We might therefore look at the overall shape that a completed account of color vision might take. Suppose we were to find that some particular area, such as V4, has all of the right properties to be the localization of color experience. Suppose, for example, that we were to find that the state-space of V4 was isomorphic to a space of color solids for different areas of the visual field, and that all effects in the psychophysics of color had parallel mappings from stimulus to V4-state. Suppose, moreover, that selective damage to V4 were to prove to cause cortical colorblindness (perhaps without total loss of discriminative abilities performed by "upstream" systems), and that stimulation to V4 with a neural probe caused predictable color sensations, regardless of traumas to upstream modules. Then we would have a plausible *localization* of color experience in V4. Would we then have an *explanation* of color *experience* as well?

If, by 'explanation', we mean an explanation of *everything about* color experience, the answer is surely no. Figure 13 schematizes the sort of explanation that we would have. On the basis of psychophysical experiments, we are able to construct models of visual color space (and discrimination space) such as the color solid. On the basis of neurological experiments, we are able to examine the properties of various parts of the visual system: the retina, the LGN, parts of the visual cortex. In our projected scenario, this culminates in revealing that color vision "all comes together" in some specific part of the cortex – for purposes of our musings, V4 – where we have a state-space that (a) is isomorphic to the state-space arrived at by our psychophysics, and (b) covaries with it: e.g., you experience a particular shade of red in a particular position when, and only when, a portion of V4 is in state $V4_i$. This convergence of formal shape and covariation is enough to provide the basis for calling V4 the localization of color experience, and an occurrence of $V4_i$ the localization of a particular experience of that shade of red in that portion of the visual field, if all we mean by "localization"

is "part of the brain (or pattern of brain activity) that is specially implicated in the experience".[2] And this is a very empirically powerful kind of relation, as it licenses all sorts of predictions and interventions that could be useful in, say, diagnosing and treating forms of cortical blindness. It would also provide the important scientific virtue of relating two previously disparate variables in a lawlike way, even though it is not a reduction of visual phenomenology to neural activation patterns.

But there is an important difference between what we would have here and what we had in the case of discrimination of chromatic information by the cone and ganglion cells. There, once we have seen how the retinal cells respond differentially to different chromatic conditions, it would be nonsensical to ask, "Yes, but how do they discriminate different wavelengths?" Discrimination just is differential response, and once we have the right sort of circuit, its functional behavior follows necessarily. By contrast, once we have mapped out the discriminative abilities of the visual system, and shown that some downstream area is isomorphic to the color space arrived at through our psychophysics, it makes perfect sense to ask, "Yes, but why do things *look red* when cells in my V4 area are firing that way?" "Looking red" does not just emerge out of differential response the way that discrimination or functional description does. It is an additional explanandum.[3]

Indeed, there are two additional explananda here. (Compare Chalmers 1996, Jacobson 1997.) One is the very presence of qualia. These do not simply emerge out of lower-level explanations the way discriminative abilities do: color qualia cannot be viewed as a *construction* out of neural states. A second is the association of *particular* qualitative characters with *particular* brain states: why does the presence of $V4_i$ co-occur with my seeing this particular reddish hue rather than, say, what I call a bluish hue? We may put this problem more pointedly thus: the person who is red-green colorblind will have both a different color space than mine and a different state-space for V4. We can, in our projected scenario, identify the V4 states that she is in when exposed either to red or to green. But we cannot, on the basis of this alone, predict the particular qualitative character of the experiences she has on those occasions. For example, might they be like my experiences of red? Might they be like my experiences of green? Or perhaps they are neither? There is simply nothing in the neurophysiology that lets us generate an answer to this. (We can, perhaps, predict it *given* a pre-existing knowledge of

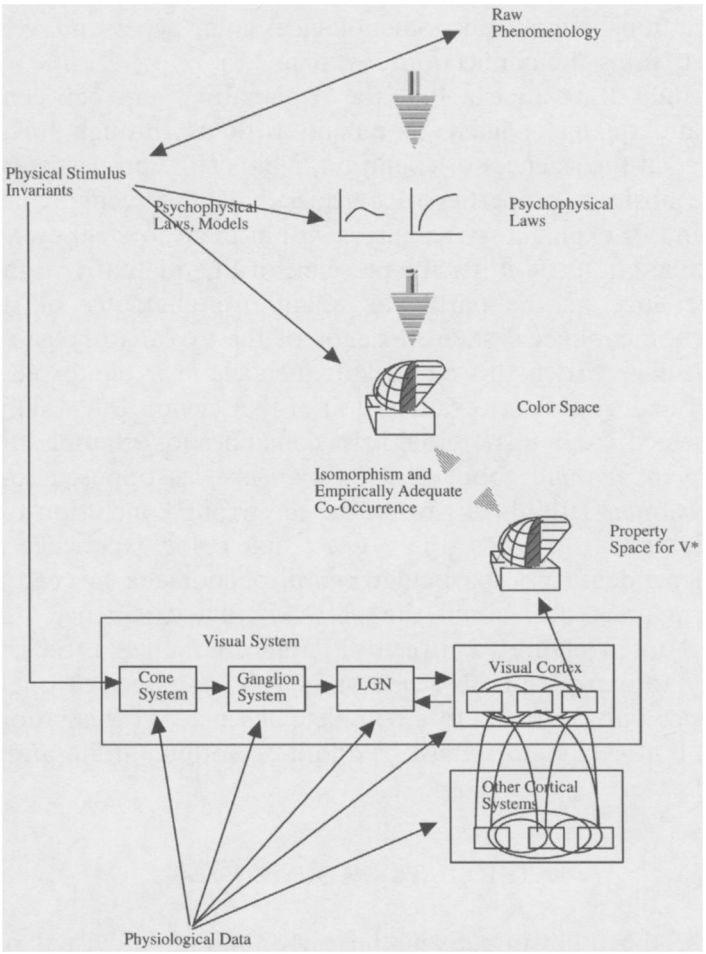


*Figure 13.* Diagram of projected completed explanation of color vision. Psychophysics yields data about the relationships between stimuli and percepts (top of diagram). On the basis of these data, it is possible to construct geometric or topological models of phenomenological color space like the Munsell color solid. Such models provide criteria for testing the adequacy of any hypothesized localization of color experience: the properties of the localizing system must have the same form we find in the psychophysical data. We can study the properties of regions like V4 through a variety of means, such as single-cell sampling, EEG and imaging techniques, as well as in vitro studies of animal cells. A plausible localization is found when the known properties of the location in question are isomorphic to those produced by the psychophysics (middle of diagram).

a correlation between phenomenological color types and V4 states, but that leaves the connection unexplained.)

It would thus appear that the explanatory gap between brain states and qualia is likely to remain with us through foreseeable advances in the sciences of cognition. Neither the properties of cells, nor the abstract properties of cortical dynamics seem to have the right kind of explanatory resources within them to yield even *candidate* explanations of either the presence of the qualitative dimension of experience or the particular qualitative character of individual phenomenological states. Friends of the explanatory gap would seem to have gotten this part right: their claim is reinforced rather than refuted by a closer examination of the science. But what should we conclude from this? Should we conclude, for example, that we can explain *nothing* about visual *experience*, as opposed to visual *discrimination?* I think this would be the wrong conclusion to draw, on at least two grounds. First, *given* that color experience is specially dependent upon particular neural phenomena, we can in fact explain a great deal about the *shape* of that experience – in this case, the inter-relations between different color qualia. Second, the inability to explain the *presence* or *specific character* of qualitative experience only amounts to a complete absence of explanation if we assume that explanation must be an all-or-nothing affair, and this is not the case.

## 8. THE SHAPE OF QUALITY-SPACE

Consider the following: Even the early findings of visual psychophysics present us with some problems that are intuitively puzzling. Why should it be that there is a phenomenologically-pure yellow but not a phenomenologically-pure orange? Why should very different chromatic patterns (different combinations of frequencies of light) be able to produce the selfsame color sensations? Why does human color-space take the form it does and not some other form? Why are some people's color-spaces different from others? These can be cast as questions about qualitative space, and not just about discriminative abilities. And there is nothing about visual qualia as such that entails that we should experience them in these ways rather than alternative ways. Psychophysics consists in empirical discoveries, not *a priori* necessities, and many of these discoveries were indeed quite surprising.

And there are ways to answer such questions. The basic form of such an explanation requires two distinct components:

(E1) The qualitative properties in question are specially related to the activation of particular neural states (for purposes of discussion, we are assuming color qualia are specially related to states of V4, normally activated through retinal and LGN activity), and

(E2) The neural mechanisms leading to these states have the right properties to explain the formal shape of the problems to be explained. For example, the nature of the cone and ganglion system explains metamers, and (one projects) a full description of the visual system would result in a model of V4 state-space isomorphic to the phenomenological color space.

*Given that* color phenomenology has this specific strong relationship to V4 activity, the *peculiarities* of color-space can thereby be explained. Indeed, they simply fall out of the model. This is by no means a trivial sort of explanation.

There is, of course, an abiding philosophical puzzle in how to understand element (E1) of such an explanation. What is the nature of this "special relationship"? Is it one of causation, as Descartes would have it? Or is it better captured by notions like type or token identity, supervenience, property dualism, or even some form of reduction that we do not yet have the conceptual machinery to work out? Or is it perhaps better cast in epistemic than metaphysical terms: as an artifact of our having to simply *associate* elements of two different models of the same processes without being able to reduce them to a single common denominator? We do not, at the moment, have a conclusive answer to such questions; and at present, at least, it does not appear that the science has supplied an answer, and the questions may thus be trans-empirical.

It seems to me, however, that having such philosophical puzzlement is not any kind of barrier to the science itself. A pragmatic association of elements from different models is not an unusual move in science. Sometimes such identifications later turn into something stronger, more on the order of reductions or ontological identities. But the assumption that they will *always* do so (at least in the successful cases) seems more like a methodological principle to guide persisting inquiry rather than either an empirical discovery or a sound metaphysical principle.

However, let us remain neutral on just how many cases there are in the sciences in which our explanations require an element of the form (E1), and look for a moment at just how such explanations differ from explanations that lack such an element, such as the explanation of certain discriminative abilities by features of the retinal cells. The latter sort of explanation has a sort of *epistemic transparency* to it: given a description of the mechanisms in the retinal cells, certain discriminative properties simply "fall out" – that is, they can be *deduced* or *constructed* from the properties of the cells. This is an example of the type of explanation that was seized upon by 17th century Rationalists and 20th century Positivists as the paradigm case of explanation: deduction and construction in mathematics or logic. I have (Horst 1996) characterized such explanations as *conceptually adequate*: we can treat the explaining system as the definitions and axioms of a deductive system and demonstrate or construct the corresponding properties of the system to be explained without the addition of any new (non-formal) conceptual content. (We might sometimes need additional formal resources, such as the statistical machinery needed to derive the gas laws from classical interactions of gas molecules, or an independent math-functional description of a circuit that is not itself constructible from the physical properties of the circuit. However, these are presumably epistemically and ontologically innocuous, as they add nothing fundamentally new, at least on the assumption that we are entitled to help ourselves to formal resources for free.)

Naturalistic philosophy of science has criticized the Rationalist and Positivist assumption that *all* explanations must be conceptually adequate. Yet the philosophical preoccupation with conceptually adequate explanations (CAEs) is not solely an artifact of misguided apriorism or math-envy. CAEs are of particular philosophical interest for a very good reason: CAEs guarantee metaphysical necessity as well. If we can derive phenomenon A from phenomenon B, $B \rightarrow A$ is true in every possible world, and hence $B \rightarrow A$ is metaphysically necessary and A is metaphysically supervenient upon B. An *explanatory* reduction (a CAE of A to B) guarantees an *ontological* reduction (that A is nothing over and above B) as well. We might put this in the form of the following principle:

*Positive Explanation-to-Metaphysics Connection Principle (Positive EMC):* If there is a CAE of A in terms of B, then $B \rightarrow A$ is metaphysically necessary.

For philosophers interested in the metaphysics of mind, particularly those interested in establishing a materialist metaphysics, Positive EMC is a very powerful principle, at least when it can be applied. The error of many reductive naturalists, in my view, is in making the assumption (often on *a priori* grounds) that CAEs are always to be had, and that their absence implies cause for methodological or even ontological suspicion.

It is much more contentious whether the *un*availability of CAEs has any metaphysical consequences. Dualists from Descartes to Chalmers and Jackson (2000) have supposed that a principled unavailability of CAEs entails a failure of necessity and supervenience as well. They thus employ a second principle:

*Negative Explanation-to-Metaphysics Connection Principle* (*Negative EMC*): If A cannot be explained by B by way of a CAE, $B \rightarrow A$ is not metaphysically necessary and A is not metaphysically supervenient upon B.

Such a principle, however, is only persuasive if one assumes that the world in its entirety – or at least metaphysical necessities – should be epistemically transparent to creatures like us. But once one articulates this assumption, it becomes clear that there is also a reasonable alternative view, articulated by "mysterians" like Colin McGinn (1991): that there are features of the world that are either entirely incomprehensible to us, or at least not susceptible to complete explanation in terms of something else. McGinn and other mysterians have argued in particular that there might be problems in human minds understanding themselves – a failure of "cognitive closure". But I think that the general failure of the reductionist programme in philosophy of science – even in the biological and physical sciences – suggests that there might be abiding and principled explanatory gaps elsewhere as well. (Horst, forthcoming a, b) (Note that if this is true, Negative EMC should lead, not to dualism, but to a much more radical ontological pluralism, in which chemical and biological facts are not metaphysically supervenient upon facts of basic physics.)

This is a juncture at which philosophers of mind and the sciences of cognition might do well to pay heed to the works of post-reductionist, naturalistic philosophers of science such as Simon (1977), Darden and Maull (1977) and Bechtel and Richardson (1993), who have begun the important project of cataloguing important sorts of ways, short of reduction, that elements in two scientific domains can be linked. Of particular importance are what Simon (1977) calls

"non-decomposable" systems, in which we have only *partial* explanations of one system A in terms of the relations of its known material parts B. In cognitive science and cognitivist philosophy of mind, it is common practice to speak of the relation between a larger system (particularly a functionally characterized system such as a digital computer) and its parts as one of "instantiation" or "realization". One system – say, a computer program – is said to be "instantiated" or "realized" by a particular arrangement of components (say, the activation states of circuits in the hardware). The words 'instantiation' and 'realization', however, are given a variety of meanings by philosophers and scientists who employ them, and hence it is useful to stipulate a more exacting usage.

Robert Cummins (1983) proposes the notion of an "instantiation analysis" in the following way. An instantiation analysis of a property P in a system S has the following form:

(6i) Anything having components $C_1 \ldots C_n$ organized in manner O – i.e, having analysis [$C_1 \ldots C_n$, O] – has property P;
(6ii) S has analysis [$C_1 \ldots C_n$, O];

---

(6iii) S has property P. (Cummins, 1983, p. 17, numbering preserved from original text)

One should be able to *derive* a proposition of the form (6i) from a description of the properties of the components of the system, and that when we can do this we can "*understand how* P is *instantiated* in S." (p. 18, italics in original, underscoring emphasis added) That is, from a specification of the properties of the components of the system in the form

(6a) The properties of $C_1 \ldots C_n$ are <whatever>, respectively;

we should be able to *derive* (6i):

(6i) Anything having components $C_1 \ldots C_n$ organized in manner O – i.e., having analysis [$C_1 \ldots C_n$, O] – has property P;

Horst (1996) contrasts this with a weaker form of explanation called a "realization account".

A realization account provides a specification of how a property P is realized in a system S through the satisfactions of some set of conditions $C_1 \ldots C_n$ – but *without* any implication that the satisfaction of $C_1 \ldots C_n$ provides a metaphysically sufficient condition for the presence of P. (Horst 1996, p. 242.)

Horst gives the example of a *generous act*. A generous act must be realized through some overt action or other – say, giving money to the needy. However, a characterization of the act itself – say, writing a check to a charitable organization – is not sufficient to guarantee that the act is generous: it might, for example, been done purely as a tax write-off or as an attempt to impress one's friends. Generosity is always expressed through some palpable behavior or other, and the performance of *some* such action is necessary for a generous act to have been performed. But it is not a sufficient condition. Similarly, suppose some economist solved the problem of world hunger by way of an economic model that required machine computation. A total explanation of the historical event of providing such an solution would require appeal to an explanation of how the computation was performed by the economist's computer, but this computational process alone would not explain why the overall event was a solution to world hunger, as that would need to appeal to additional facts as well. The computation is a vital part, but only a part, of that story. The solving of the problem of world hunger might be said to be realized through a computational process; but the computation alone is not a sufficient condition for finding a recipe for the abolition of hunger.

In these paradigm cases, it is clear both that the explanations in question are not CAEs and that specifiable additional conditions (such as the intentions of the giver or the application or potential application of the model to a specific real-world problem) are required for the type of event in question to actually take place. However, the notion of a realization account itself can be regarded simply at the level of explanation, while remaining neutral on the Negative EMC: there are *incomplete* part-whole *explanations* that fall short of instantiation analyses (or, more generally, of CAEs). They still have explanatory power, even though it falls short of the explanatory power of a CAE. There are probably a number of distinguishable types of realization accounts, but it is most useful for present purposes to treat realization accounts as a broad category, and as neutral with respect to the proper metaphysical interpretation of a given instance of their application.

What is important about realization accounts is that properties of the realizing system can accrue to the realized system as well. If Jones wrote the check on December 31, then Jones's generous act was performed on December 31. If the computation was performed using the Jones algorithm, the production of a remedy for

world hunger was abolished by way of the Jones algorithm as well. Likewise, if human color vision is accomplished through retinal processes and V4 processes, relevant features of those processes accrue to color vision as well, even if some aspects of color vision (such as its involving color qualia) are not thereby explained. *If color vision is realized through such processes, and these processes result in a state-space like that described by the Munsell color solid, then phenomenological color-space will take a form described by the color solid as well.*

This is a non-trivial form of explanation. It is one that could not be reached by introspection or *a priori* reasoning. And it is an explanation of phenomenological facts: namely, the abstract shape of phenomenological color space, and more local facts such as that there is a phenomenologically pure yellow but not a phenomenologically pure orange. Likewise, it explains why the phenomenological color-space of dichromats is different from that of trichromats in ways wholly predictable from what neurocience can tell us about the process of seeing, despite the fact that the neuroscience cannot explain why there is this special realization relationship between qualia and the brain.

What this broad notion of realization allows us to do in cognitive science is to keep the explanatory gap relatively well contained. There is *one* explanatory gap in color experience: color experience is *realized* through some brain states – there is a special and empirically robust relation between them – but we know not what that special relation is nor can we explain it. We do not, therefore, need a separate account of, say, the opponent color process or contrast effects for *qualia* once we have one for *discriminative abilities* in the nervous system. The structure of the visual system explains the topology of discriminative abilities, which eventuate in some state space of brain activity (in our fanciful speculations, in states of V4), and color experience in us is realized through these very states. (More accurately, for each of us at a given time *t* there is such a state space through which our color qualia are realized. It may be different across individuals and across times in a given individual.) The topography of the realizing system then accrues to the realized system as well. (Compare Horst 1996, p. 357.)

We might thus distinguish three types of problems in terms of whether they are explained by neuroscientific explanation, alone or in combination with other assumptions, and whether the features thus explained are phenomenological. (See Table III.) This last cat-

TABLE III

| Phenomenon | Explained by | Phenomenological? |
|---|---|---|
| Functional and discriminative abilities | Neural anatomy and cortical dynamics | No |
| Presence of qualia | Not explained | Yes |
| Particular qualitative characters | Not explained | Yes |
| Topology of visual qualitative space | Neural anatomy plus cortical dynamics plus assumption of realization of phenomenological seeing through visual system | Yes |

egory of explanation has gone largely unremarked-upon, and shows that neuroscientific explanation can be applied to some phenomenological facts even though it cannot explain the presence or precise character of qualia.

## 9. CONCLUSION

A "hands-on" investigation of case studies in the sciences of cognition is philosophically productive. It allows our philosophy of psychology to proceed beyond armchair speculation by confronting us with the different types of explanatory situations we actually encounter in different cases. (For example, differences between the local mechanism-like explanations of retinal systems and the more global and distributed explanations of cortical dynamics, or between the cases in which localization leads the way and those in which modeling does so.) It also enables us to see more clearly whether problems first posed by philosophers, such as the explanatory gap, are dissolved in the face of empirical investigation and theory. (I argue that they have not been thus dissolved, and seem unlikely to be dissolved in the future.) But perhaps most interestingly, it helps us to see how empirical and philosophical problems really intersect rather than treating them in isolation. For example, there *does* seem to be an abiding explanatory gap, whose metaphysical or epistemic interpretation is still in doubt. But this does not mean that nothing about qualitative phenomenology is explained by the sciences of cognition.

## NOTES

[1] In much of the material presented here about vision, I am indebted to Stephen Grossberg and the Center for Adaptive Systems at Boston University, where I

spent a sabbatical in 1993. The work on the mathematical properties of retinal cells is highly indebted to Austen Clark's (1993) *Sensory Qualities*. In the concerns about the history of philosophy of science in the introductory section, I was much influenced by conversations with Paul Humphreys, Bas van Fraassen and Joseph Rouse, which collectively awoke me from the illusions of reductionist philosophy of science. Previous versions of this paper were written while I was on sabbatical in 1997-8 at Princeton University and the Center for the Study of Language and Information at Statford University under the auspices of an NEH Fellowship. Any errors in my presentation are, of course, wholly my own.

[2] It is important to note that the color solid is not a model of global color-perception activity, either phenomenologically or in the brain, but a model of color-perception possibilities in one region of the visual field. The model of either visual color phenomenology or V4 would have to posit such a structure for each portion of the visual field capable of responding to color. As Clark (1993) points out, this would require a model of at least six dimensions.

[3] On this point, I believe my analysis is contrary to that of Clark (1993). While Clark does not equate qualia with capacities to distinguish objective phenomena, and employs a "methodological solipsism" in which the data for phenomenological spaces such as color space are constrained by the phenomenology of what the subject distinguishes, he believes the resulting relational description which locates, say, colors with respect to one another provides a complete account of qualia. He admits that this might generate problems in the case of a symmetrical color space. I agree that qualia are best identified in the way he describes, but deny that this exhausts their nature.

## REFERENCES

Baker, Lynne Rudder: 1995, *Explaining Attitudes*, Cambridge University Press.

Bechtel, William and Robert C. Richardson: 1993, *Discovering Complexity: Deconmposition and Localization as Strategies in Scientific Research*, Princeton University Press

Chalmers, David: 1996, *The Conscious Mind*, Oxford University Press.

Chalmers, D. J. and Jackson, F.: 2001. 'Conceptual analysis and reductive explanation', *Philosophical Review*.

Clark, Austen: 1993, *Sensory Qualities*, Oxford: Clarendon Press.

Cummins, Robert: 1983, *The Nature of Psyhcological Explanation*, MIT Press.

Darden, Lindley and Nancy Maull: 1977. 'Interfield Theories', *Philosophy of Science* **44**, 43–64.

Glymour, Clark: 1999, 'A mind Is a Terrible Thing to Waste–Critical Notice: Jaegwon Kim, "Mind in a Physical World"', *Philosophy of Science* **66** (3): 455–471.

Gregory, R. L: 1978, *Eye And Brain: The Psychology of Seeing*, 3rd ed., McGraw Hill/World University Library, New York.

Grossberg, Stephen: 1987, *The Adaptive Brain*, North-Holland, Amsterdam.

Helmholtz, Hermann von: 1867, *Handbuch der Physiologischen Optik*, 1st ed., Voss, Leipzig.

Hering, E: 1878, *Zur Lehre vom Lichtsinn*, Gerald u. Söhne, Wien.

Horst, Steven: 1996, *Symbols, Computation and Intentionality: A Critique of the Computational Theory of Mind*, University of California Press, Berkeley and Los Angeles.
Horst, Steven: Forthcoming-b, 'Beyond Reduction: What Can Philosophy of Mind Learn from Post-Reductionist Philosophy of Science?'
Horst, Steven: Forthcoming-b, *Beyond Reduction*.
Hurvich, L. M.: 1981, *Color Vision*, Sinauer Associates, Inc., Sunderland, MA.
Jackson, F.: 1982, 'Epiphenomenal qualia,' *Philosophical Quarterly* **32**, 27–136.
Jacobson, John.: 1997, "The Mysteries of Consciousness and the Foundations of the Broad Approach." Undergraduate thesis, Wesleyan University, Middletown, CT.
Kim, jaegwon: 1998, *Mind in Physical World: An Essay on the Mind-Body Problem and Mental Causation*, MIT Press, Cambridge, Massachusetts.
Land, Edwin: 1959, 'Experiments in Colour Vision', *Scientific American* **5**, 84 (1959).
Levine, J.: 1983, 'Materialism and qualia: The Explanatory Gap', *Pacific Philosophical Quarterly*, **64**, 354–361.
Livingstone, Segregation of Color, Form, Motion and Depth: Anatomy, Philosophy and Perception, *Science* **240**, 740–750.
McGinn, Colin: 1991, *The Problem of Consciousness: Essays Toward a Resolution*, Blackwell.
Nagel, Thomas: 1974, 'What is it like to be a bat?', *Philosophical Review* **4**, 435–450.
Sacks, Oliver: 1985, '*The Man who Mistook his Wife for a Hat and Other Clinical Tales*', Summit Books, New York.
Simon, Herbert: 1977, *Models of Discovery*, D. Reidel, Dordrecht.
Spillman, Lorthar and John S. Werner (eds.): 1990, *Visual Perception: The Neurophysiological Foundations*, Academic Press, San Diego, CA.
Ungerleider L. G. and Mishkin, M.: 1982, 'Two Cortical Visual Systems', in *Analysis of Visual Behavior*, Ingle, D.J.
Young, Thomas: 1801/1961, 'On the Theory of Light and Colours', Reprinted in R.C. Teevan and R.C. Binney (eds.), *Colour Vision*, Van Nostrand (1961).

Department of Philosophy
Wesleyan University
Middletown, CT 06457
USA
E-mail: shorst@wesleyan.edu