PSYCHE

Cortical Feedback and the Ineffability of Colors

Mark F. Sharlow
Department of Chemistry and Biochemistry
University of California, Los Angeles
Los Angeles, California 90095
USA

msharlow *at* usermail *dot* com

PSYCHE 11 (7), October 2005

KEYWORDS: Ineffability, Sensation, Color, Perception, Vision, Visual System, Feedback

ABSTRACT: Philosophers long have noted that some sensations (particularly those of color) seem to be ineffable, or refractory to verbal description. Some proposed neurophysiological explanations of this ineffability deny the intuitive view that sensations have inherently indescribable content. The present paper suggests a new explanation of ineffability that does not have this deflationary consequence. According to the hypothesis presented here, feedback modulation of information flow in the cortex interferes with the production of narratives about sensations, thereby causing the subject to assess as inadequate his or her own verbal descriptions of sensations.

1. Introduction

The task of describing an experience of a color to a person who never has seen that color is a proverbially difficult undertaking. The difficulty of this task, and similar difficulties with the description of non-visual sensations, have attracted the attention of philosophers over the centuries (see, for example, Locke 1689: Bk. II, Ch. II, par. 2 and Bk. III, Ch. IV, pars. 4-7; Hume 1739: Bk. I, Pt. I, Sec. I; Dennett 1991: pp. 49-50, 382-383). Reflection on these difficulties has given rise to the well-known claim that at least some sensations are *ineffable*—in other words, that it is truly impossible to describe the phenomenal aspects of sensory experiences. The view that sensations (particularly those of color and tone) are ineffable has become part of the folklore of the philosophy of mind, though certainly not a part free from controversy. (For a general discussion of this doctrine and some challenges to it, see, for example, Dennett 1991: pp. 49-50, 382-383.)

The idea of the ineffability of sensations has played important roles in arguments in the philosophy of mind. In earlier centuries, John Locke and David Hume maintained that one cannot form a correct idea of a color without first having had a sensation of that color (Locke 1689: Bk. II, Ch. II, par. 2 and Bk. III, Ch. IV, pars. 4-7; Hume 1739: Bk. I, Pt. I, Sec. I). Hume allowed only a very limited class of exceptions to that claim. Both Locke and Hume used this impossibility to support certain aspects of their theories of mind. More recently, Frank Jackson (1982) argued that one cannot describe experiences of color in a complete way using physical descriptions alone. Jackson's argument is not about ineffability as such, but it coheres with the traditional view that color sensations are, in some sense, indescribable. Even more recently, Daniel C. Dennett (1991: pp. 49-50, 382-383) critiqued some traditional views of ineffability in the light of current neurobiological knowledge.

Some recent investigators have proposed neurophysiological explanations for the reported ineffability of sensations. Dennett (1991: pp. 375-383) has suggested one such explanation. According to Dennett's hypothesis, the great difficulty of describing colors arises from a kind of complexity in the features detected during color recognition in the brain. This explanation equates ineffability to a type of practical impossibility -- the impossibility of describing experiences that are, in a sense, too complex to describe in practice. Paul M. Churchland (1996) has suggested a different neurobiological explanation of ineffability. According to Churchland's account, some discriminated features are not analyzed by the brain in a way that would permit their conscious perception as complex features. Feature discriminations having this characteristic, Churchland notes, could cause a subject to experience what seem to be unanalyzable (and hence ineffable) qualia. Both Dennett and Churchland note that the bounds of what is ineffable may be subject to change.

Both Dennett's and Churchland's explanations of ineffability are deflationary in the following sense. According to Dennett's proposal, a subject finds a sensation indescribable because that subject cannot get a grip on all the intricacies of the feature being discriminated. This suggests that if the subject were able to notice a lot more things at once, then the subject might be able to describe the feature after all. According to Churchland's view, a subject finds a sensation indescribable only because that subject is not equipped to analyze, in a consciously accessible way, certain discriminated features. This implies that if the subject were able to analyze the features a little more thoroughly, then the ineffability would be reduced. Thus, these explanations imply that sensations which seem ineffable do not really have anything inherently indescribable about them—and specifically, that such sensations do not have any phenomenal content that is inherently indescribable. (We note in passing that this "deflationary" character of Churchland's and Dennett's explanations of ineffability contributes to, but does not depend logically upon, those authors' deflationary attitudes toward qualia. The ineffability of sensations involves, at very least, a behavioral phenomenon: the inability to produce narratives of certain kinds. This inability is real whether or not there exist qualia, or even sensations. Hence the phenomenon called "the ineffability of sensations" is an appropriate target for explanation regardless of one's views on qualia or one's account of sensations.)

In this paper I will suggest another possible neurobiological explanation for the ineffability of sensations. This explanation is compatible with our present knowledge of neuroscience, and is of philosophical interest because it does not have the deflationary consequences of the earlier explanations discussed above. The new account takes the impossibility of fully describing sensations to be a rather strong kind of impossibility—not merely a simple kind of practical impossibility resulting from insufficient information processing capacities. According to the hypothesis presented here, ineffability is a kind of side effect of the feedback modulation of information flow in the cortex. Specifically, ineffability arises when this feedback modulation interferes, on a relatively short time scale, with the production of verbal narratives about conscious experiences.

The explanation presented here is not intended as a polemic against other ideas about the neural origin of ineffability. Rather, it may be viewed as one among several mechanisms that could contribute simultaneously to the reported ineffability of sensations. Dennett's and Churchland's accounts of ineffability appear to be mutually compatible, in the sense that the mechanisms described by both of these accounts could act in the same brain and contribute jointly to the difficulty of describing sensations. The present hypothesis appears to be compatible, in the same way, with both Dennett's and Churchland's hypotheses. All three of these sources of ineffability could exist and act within the same brain.

The data that I will adduce in support of my explanation of ineffability do not provide conclusive evidence for that explanation. Although the explanation is compatible with our present knowledge of neuroscience, it is underdetermined by the facts known today. Nevertheless, the explanation is of interest because it shows that a neural explanation of ineffability need not be as deflationary as some previous explanations of this sort have been.

For concreteness' sake, I will motivate and support this hypothesis mainly with arguments about color. I will begin by developing an explanatory hypothesis about the origin of ineffability in the visual system. By doing this, I do not mean to suggest that the ineffability of colors is a phenomenon different from the ineffability of sensations in general. The explanation of visual ineffability that I will develop makes use of processes and structures that have analogues in other sensory systems as well, and hence should be equally applicable to those other systems (more on this later). I will study the visual system first because it is convenient to study, and also because philosophers' examples of ineffability tend to be color examples.

The main thread of my argument runs as follows. Feedback connections between components of the visual system are known to be widely distributed in primate brains (see, for example, Felleman & Van Essen 1991; Van Essen et al. 1992). These connections are thought to be involved in the modulation of data flow through the visual system (Felleman & Van Essen 1991; Van Essen et al. 1992). The hypothesis I will suggest rests on the assumption that this feedback modulation can influence the content of verbal narratives about visual experiences. Beginning with this assumption (in a more refined and detailed form) and with published data about the timing of certain neural events, I will argue that a subject who attempts to describe his or her color experiences usually will judge the resulting descriptions to be inadequate. This perceived inadequacy of most phenomenal color descriptions is, I will suggest, a possible cause of the reported

ineffability of colors. Further, I will argue that a color experience that is ineffable for this reason is indescribable in a stronger sense than is possible with Churchland's and Dennett's hypotheses. I will spell out what I mean by this last claim later in the paper.

2. Ineffability: Old and New Views

Philosophers have held conflicting views about the nature of ineffability. At bottom, ineffability is an observed psychological phenomenon; conscious subjects who try to describe sensations often find that they cannot do so, or at least cannot do so in any way that the subject assesses as adequate. Philosophical claims about ineffability grow out of attempts to understand this phenomenon. Traditionally, philosophers have linked the concept of ineffability to the idea of the simplicity of sensations. In the seventeenth century, Locke characterized ideas of color as "simple ideas" (Locke 1689: Bk II, Ch. II, par. 2). In the eighteenth century, Hume classified color sensations as "simple impressions" (Hume 1739/1983: Bk. I, Pt. I, Sec. I). Today, the connection between ineffability and simplicity is part of the folklore of the philosophy of mind. Often, this connection seems intuitively compelling. People who reflect upon their sensations, and especially upon their sensations of color, often feel very strongly that some sensations are unanalyzable. For example, when one sees a particular shade of blue, the color that one sees does not seem to have any internal structure that would allow one to analyze the color into simpler qualities.

The traditional belief that sensations are simple has come under fire in recent decades. Scientists and philosophers alike have cast serious doubt upon this belief. The work of color scientists has made it quite clear that colors are not entirely unanalyzable (see Hardin 1988, especially pp. 40-44, and Kaiser & Boynton 1996: pp. 40-45, 304-308). Perceived colors can be assigned positions in an abstract color space whose dimensions correspond to features of colors. In one version of color space (see Hardin 1988: pp. 113-116 and Kaiser & Boynton 1996: pp. 41-45), the coordinates are hue, brightness and saturation. Perceived similarities and differences in colors can be analyzed successfully within color space models of this sort. Jonathan Westphal (1984) has criticized the notion of the simplicity of colors. His line of argument makes use of the fact that colors can be analyzed in terms of qualitative features; he points out that the belief in the simplicity of colors rests on certain identifiable conceptual mistakes. Dennett has pointed out that experiences which people assume are unanalyzable often turn out to be analyzable when the subject acquires new perceptual skills (Dennett 1991: pp. 49-50). Taken together, these considerations strongly suggest that the belief in the qualitative simplicity of colors should be abandoned.

There is at least one proposed explanation of ineffability that has no need at all for the supposed simplicity of sensations. This is the explanation proposed by Dennett (1991: pp. 375-383). According to Dennett's view, the discrimination of a color requires a highly complex discriminatory state -- one so complex that a subject who tries to describe the sensation cannot describe exactly what is being recognized. In other words, the subject has trouble describing colors because the features being discriminated are so complex as to be impossible, in practice, to describe. Dennett (1991: pp. 376-383) likens color discriminations to the mutual recognition of two persons by means of the two halves of a torn piece of cardboard. The edges of the two halves match each other exactly, allowing

for effective recognition—yet any attempt to describe what is being matched will lead to an intricate, and in practice unfinishable, description of tiny zigs and zags on the cardboard. One torn edge can "discriminate" the other torn edge immediately—yet it is impossible to explain in words the feature that is being discriminated. According to Dennett's explanation, the ineffability of discriminated colors has a similar origin. Color experiences are ineffable, but are not simple; rather, they are ineffable because they are too complex in a certain way. Dennett's explanation portrays ineffability as a practical impossibility rather than as a conceptual one.

Dennett's view of ineffability differs sharply from traditional views that link ineffability to simplicity. One might ask whether Dennett's explanation really addresses the same notion of ineffability found in the traditional views. Should we say that there is a "too-complex" notion of ineffability, distinct from the traditional "too-simple" notion? I would argue that there are not two separate notions of ineffability, just two separate routes of explanation. Ineffability is not primarily a philosophical notion, but is an observed psychological phenomenon that people encounter when they try to describe their sensations. The claim that this phenomenon depends on the simplicity of sensations is not a definition of a notion of ineffability, but is a claim about the origin of ineffability. Similarly, the claim that ineffability is a result of complexity is not a definition of ineffability, but is a claim about the origin of ineffability. Philosophers may disagree about the true explanation of ineffability (simplicity versus complexity), but this does not imply that the two disagreeing camps are talking about two different phenomena called "ineffability." They are talking about the same observed phenomenon; they are only disagreeing about the explanation of that phenomenon. Thus, instead of distinguishing between too-complex and too-simple notions of ineffability, we should distinguish between complexity-based and simplicity-based *explanations* of ineffability.

According to Dennett's complexity-based explanation, the subject of a color sensation cannot describe that sensation exhaustively. The subject may be able to say something about the color, but cannot (in practice) say enough to completely specify the phenomenal content of the color. No matter what the subject manages to say, most of the information that makes the color experience what it is will go unsaid. In practice, this amounts to the ineffability of the color.

Another modern explanation of ineffability is the one offered by Churchland (1996). According to Churchland's view, the indescribability of features of experience is a result of the perceptual system's inability to discriminate structure within those features. Even if a feature being discriminated is composite in some way, the system does not discriminate the individual components that make up the feature. Hence the feature seems, to the subject, to be indescribable and also simple. One can regard Churchland's explanation of ineffability as a simplicity-based explanation, even though it reduces the simplicity of sensations to a sort of illusion.

In this paper I am going to present a new explanation of the neural origin of ineffability. This explanation is complexity-based, but it involves a kind of complexity quite different from anything found in Dennett's approach. In the next section I will lay the necessary groundwork for the new explanation.

3. A Thought Experiment: Describing Red

Before laying out the new explanatory hypothesis, I will present a gedankenexperiment that provides some motivation for that hypothesis. First I will introduce some necessary background information from neurophysiology.

The primate visual cortex consists of several distinct areas that have been differentiated through a variety of methods. There exists a fairly extensive literature on these areas and their connections (see, for example, Van Essen et al. 1992; Felleman & Van Essen 1991; Maunsell & Newsome 1987, and some of the references therein). Among these visual cortical areas are the primary striate visual cortex and a number of extrastriate areas (see especially Van Essen et al. 1992; Maunsell & Newsome 1987). These areas are extensively interconnected by feedback and feedforward connections (see, for example, Van Essen et al. 1992; Felleman & Van Essen 1991; Maunsell & Newsome 1987). Most of the feedforward connections occur in pairs with feedback connections that join the same cortical areas (see especially Felleman & Van Essen 1991; Van Essen et al. 1992). The feedback connections are thought to modulate and direct the flow of information through the visual system (see especially Van Essen et al. 1992: p. 255; Felleman & Van Essen 1991: p. 41).

The characteristic time scale for control processes involving such modulation has been estimated to be of the order of 100 msec (Van Essen et al. 1992: p. 255). More recent experimental studies (Bullier et al. 1996) appear to be consistent with this order of magnitude estimate. I should emphasize that not all of the pathways used in the studies just cited affect color perception in a way relevant to my thought experiment below. However, one can use the figure of 100 msec as an educated guess for the order of magnitude of the time scales of typical visual cortical feedback processes. One would expect this order of magnitude to be applicable across a wide variety of primates, since the basic physics of neurons is the same throughout that group.

The gedanken-experiment that I wish to consider is quite simple. In it, a human subject views a large red dot. While the subject is doing this, an experimenter asks the subject to say what it is like to see that particular shade of red. Specifically, the experimenter requests a complete description of that shade of red—a description so thorough that a person who has not seen red would be able to tell, from the description, what it is like to experience that shade. Each time that the subject produces a description of any sort, the experimenter asks the subject to extend the description so as to make it more complete, or, if the subject finds this impossible, to describe again what it is like to see the color, this time in words other than the ones the subject already has used.

This experiment admittedly is unconstrained, but it will be sufficient for my purposes. I am deploying this experiment, not to compare the consequences of different hypotheses, but to point up some things that might happen when a subject who has visual cortical feedback loops tries to describe a sensation of color.

According to our present knowledge of the visual system (as presented particularly in Van Essen et al. 1992), the sequence of neural events precipitated by this experiment must run, in broadest outline, as follows. First, most of the outputs from the subject's retinas become inputs to a structure called the lateral geniculate nucleus, or LGN. Then outputs from the LGN reach the visual regions of the cortex and act as inputs for various sorts of processing in those regions. A small fraction of the retinal data is

transmitted to the cortex via another pathway involving the superior colliculus and the pulvinar complex. Outputs from the visual cortical regions reach various other regions of the brain (see Van Essen et al. 1992; Felleman & Van Essen 1991). Outside the visual system, other events of a highly complex nature occur, finally resulting in activation of the language system. This eventually results in the subject's uttering a verbal narrative when stimulated by the experimenter's instructions.

In the experiment we are considering, the semantical content of the final narrative will depend upon the information contained in the visual areas of the cortex, as well as upon other variables. Presumably, some of the visual areas will exercise a greater influence than others upon the content of the narrative. Although the precise identities of these latter regions are unimportant to the present hypothesis, it is plausible to suppose that these areas will be ones strongly involved in color recognition, such as V2 and V4 (see Felleman, Xiao, & McClendon 1997; Maunsell & Newsome 1987).

The production of the spoken narrative occurs on a time scale of seconds. Various experimental data suggest that the production of a verbal report on a subjective experience requires a time interval of at least the order of 300 msec (see Dennett 1991, particularly the summaries and interpretations of experimental results in Chapter 6). Hence the figure of 300 msec may be taken as an approximate lower bound for the time required to produce a verbal narrative in response to a stimulus. The production of a highly complex and detailed narrative will, of course, take longer; clearly, the attempted qualitative description of a color by the perceiver of the color requires at least a few seconds.

During the time interval occupied by production of the narrative, feedback processes occur among the visual cortical regions. Comparing the above mentioned time scale for this feedback (~100 msec; see above) with the time scale of ~300 msec for narrative production, we find that substantial changes in the information content of cortical areas can occur during the production of a verbal report about a perceived color. If the report is sufficiently complex and reflective ("Well, that shade of red is sort of like..."), then because of the narrative's temporal length, the changes in visual cortical data during the report's production may possibly be extensive. Note that this argument does not depend upon the time scale of the verbal report being much greater than that for the feedback effects. The conclusion also follows if these two time scales are approximately equal.

These considerations lead up to the first assumption of our proposed explanation of color ineffability.

> (A1) Corticocortical feedback leads to ongoing changes in visual cortical data over the time scales typical of that feedback.

This assumption is an empirical hypothesis. However, it appears plausible in view of the timing considerations just mentioned. A possible objection to this assumption is that feedback normally tends to stabilize systems rather than to promote ongoing changes. There are two replies to this objection. The first is that feedback does not always stabilize a system. Positive feedback, for example, tends to destabilize systems. Such feedback can be functionally useful when signal amplification is required; Erich Harth (1993) has

proposed that positive feedback plays an important role in the visual system and performs a kind of amplifying function there (1993: pp. 67-73). Unless we know that the feedback in the visual system always is purely negative, we cannot safely assume that it always stabilizes the system. The second reply to the objection rests on the fact that even systems well-stabilized by feedback can undergo oscillatory transient behavior when their inputs vary in time. Since the input of the visual system is time-varying (for example, due to saccades; see Dennett 1991: p. 111; Kaiser & Boynton 1996: p. 95), it seems likely that some part of the system undergoes such transient oscillations more or less continually. Hence the temporal variation that our assumption requires might exist even if feedback does stabilize the visual system.

More generally, one can argue that the feedback in the visual system, whether positive or negative, probably will produce temporal variations of some sort on the time scale of the feedback loops. If the feedback is negative, then there will be transient oscillations on that time scale. If the feedback is at least partly positive, then it will tend to produce amplification effects on the same time scale. The mere presence of feedback loops tells us nothing about the details or extent of the temporal variation. However, the presence of feedback loops in a system with rapidly changing inputs would tend to produce either transients or amplification effects, neither of which directly follows the input signal as the signals in a feedforward system might do.

Because of the wide distribution of feedback pathways in the visual system, it seems reasonable to suppose that feedback effects can causally influence the informational content of the signals that the visual system sends to other parts of the brain. In particular, it is plausible to suppose that feedback effects in the visual system have an impact upon the language system; after all, the language system responds strongly to input from the visual areas whenever a narrative about the visual world is generated. If feedback effects play major and widespread roles in vision, then it would be surprising indeed if such effects had only an insignificant impact upon what we report that we see. Thus, it is plausible to suppose that feedback-induced temporal variations within the visual system can affect the semantical content of a subject's narratives about a color experience. This supposition, like our first assumption, is an empirical hypothesis, but as we have seen, it has some support from existing evidence.

We will make this supposition the second assumption of our proposed explanation of color ineffability.

> (A2) The semantical content of narratives about color experiences frequently is influenced by feedback-induced temporal variations in the state of the visual cortex.

If we assume (A2), then it is reasonable to suppose further that the temporal variations in the visual cortex influence the outputs of the language system over the time scale of the variations. This supposition is plausible because the language system can respond to visual stimuli of duration ~50 msec, which is slightly below the time scale of the feedback-induced variations. (For facts about flicker that support this claim, see Kaiser & Boynton 1996: pp. 384-385; I will discuss these facts below.) Of course, the fact that the language system can respond to one kind of 50-ms change in the visual system does not imply that the language system will respond to another kind of 50-ms change there. Indeed, the language system's *in*sensitivity to certain changes of this kind will soon turn

out to be important to my argument. But the fact that the language system can respond to some changes on this time scale makes our further supposition tolerable from a physiological standpoint.

We will adopt this further supposition as the third assumption of our proposed explanation of color ineffability.

> (A3) If feedback-induced variations in the state of the visual cortex influence the semantical content of narratives about color experiences, then the resulting changes in the narratives frequently will occur over the time scale of the variations in the visual system.

If assumptions (A1) and (A2) are true (with or without (A3)), then feedback-induced variations in a subject's visual system often will be reflected in changes in that subject's narratives about color experiences. However, there are good biological reasons to suppose that the subject will not report these variations as variations in the perceived color, or even in the brightness, of the stimulus. An organism's ability to respond to colors in its environment—that is, to spectral characteristics of its surroundings—clearly has positive survival value (see (Kaiser & Boynton 1996: p. 37)) on the related survival value of color constancy). This value would be sharply reduced if the organism responded to routine internal fluctuations as if those fluctuations represented changes in observed colors. An organism that did respond in this way would find its visual world to be in a continual (and uninformative) condition of flicker. Evolution probably would not place organisms in such a confused state. Hence the feedback-driven temporal variations required by our assumptions are unlikely to be noticed as changes in color. The brain's color perception system has evolved to help the organism keep track of the external world. That system is not likely to mistake the internal noise arising from its own routine processing for blatant changes in the external world. The fact that the system is physically capable of changing its state on the time scale of the noise does not alter this fact.

Another argument for this same conclusion is a variant of Dennett's argument about the imperceptibility of ocular saccades (Dennett 1991: pp. 355-356, 361-362). Dennett points out that a temporal change in sensory input may escape conscious notice if the brain has no effective way to detect the change itself. As an example, Dennett notes that ocular saccades do not cause perceived breaks in visual experience; he attributes this lack of sensitivity to the brain's failure to detect the breaks in visual input (Dennett 1991: pp. 355-356). By a similar argument, if there is no detector capable of responding selectively to the ongoing revisions of data in the visual system and of making the subject aware of those revisions, then there is no reason why the subject should notice the revisions at all.

These considerations lead up to the next assumption of our proposed explanation of color ineffability:

> (A4) Feedback-produced alterations in visual information usually do not cause changes in the identities of perceived colors as reported by the perceiver.

An objection to this assumption arises from the fact that chromatic flicker can be perceived at frequencies much higher than those corresponding to the time constants of the corticocortical feedback pathways. Under certain conditions, the frequency threshold

for the perception of chromatic flicker can be as high as 20 Hz or greater (Kaiser & Boynton 1996: pp. 384-390). This implies that the visual system can detect 50-msec variations in visual input. This experimental fact seems to conflict with the claim that variations in visual data over ~100 msec need not be experienced as variations.

My reply to this objection is that these two kinds of variation differ from each other both physically and functionally, and that because of these differences they are likely to have very different perceptual effects. The variation in visual information in a chromatic flicker experiment is simply a variation in the input signal of the visual system. The feedback-driven variation required by my hypothesis originates much farther inboard, and is generated by mechanisms quite different from those acting in the retina. Functionally speaking, it is noise and not signal. We should not expect the visual system to handle this kind of variation in the same way in which it handles variations in the signal from the retina. Certainly there is no physical reason why variations of these two kinds must be handled similarly. Indeed, it would be quite surprising if they were, since all sensory systems manage to separate the signals that they are designed to process from a large amount of noise due to irrelevant neuronal firings. Thus, the above-mentioned facts about flicker do not, in themselves, count against the assumption that the feedback-caused variations are not noticed as flicker. For the same reason, my earlier argument about the noticeability of visual stimuli of duration ~50 msec also does not count against this assumption, even though that argument was based on the perceptibility of chromatic flicker.

The assumption that feedback-caused variations do not alter a perceiver's judgments about the identities of perceived colors implies that these variations do not alter visual information in an entirely random fashion. Some kinds of information remain stable; according to (A4), the stable set of data includes the information that directly underpins judgments about which color one is seeing. The feedback-driven variations will not make the subject describe a stimulus first as red, then as yellow—or, for that matter, first as a bright red and then as a less vivid red.

This restriction against variations in reported identities of colors cannot be extended to cover all verbal responses to color stimuli. Not all kinds of judgments about colors offer the organism substantial evolutionary advantages; those that do not can have more variability than those that do. The only judgments that seem likely to share the stability of color identity judgments are those that concern relative degrees of resemblance between perceived colors. These resemblances are tied to the positions of colors in color space (Hardin 1988: pp. 113-121) and hence to the perceived identities of colors. (On color resemblances in general, see Hardin 1988, especially pp. 113-121.) But this constraint does not rule out all conceivable variations in a perceiver's descriptions of perceived colors. Later I will point out some kinds of variation that are allowed.

The assumptions (A1) — (A4) introduced in this section are the premises of the explanation of ineffability of color that I am proposing in this paper. (Later I will suggest that this explanation can be extended to other sensations besides those of color.) All four of these assumptions are empirical, and as the preceding arguments show, all of them are plausible from physiological and evolutionary standpoints. Of these four assumptions, (A1) is the best supported by the evidence; the distance from known neuroscience to (A1) is not very long. (A4) also seems quite plausible in view of what we know, while (A3) is

conjectural but is consistent with what we know, and (A2) is a bit less conjectural than (A3). Note that these assumptions are not the only empirical assumptions that I will use in this paper. I will utilize some facts from neuroscience and experimental psychology as well. However, I think that these four assumptions are the only ones conjectural enough to require special mention.

4. From Cortical Feedback to Ineffability

Let us now try to predict the outcome of the gedankenexperiment described in the previous section, while keeping in mind the assumptions presented there.

At a time *t* just after the subject finishes uttering a report, the informational state of the subject's visual system is substantially different from what it was at the time when the subject began the utterance. (This follows from (A1) and the time scale orders of magnitude presented earlier.) According to the design of the experiment, at *t* or shortly thereafter, the experimenter will initiate a request for a more thorough description. This request then will provoke a new report from the subject.

The first and second reports are "about" the same object (the dot). Furthermore, according to (A4), if the subject is asked whether the dot's color has changed, he will reply in the negative. But in spite of this, the subject's new report is likely to differ in content from the old report, since the new report is generated with the help of new data that were not available earlier. This follows from (A1) - (A3).

During the course of the experiment, the sequence of events just outlined will be repeated several times, resulting in a series of reports on the stimulus. Some of these reports will differ in content from their predecessors in the series. Because of the ongoing changes in the visual system, this production of new and varying reports may continue as long as the subject tries to describe the color (that is, until the experiment ends) or until fatigue intervenes at some level. (This latter outcome may occur after just a few reports or after many, depending upon the details of the physiology of language production and upon the state of the subject.)

We can summarize the outcome of our thought experiment informally as follows: Even after the subject has uttered some verbal reports about what it is like to see the color, he will find himself with more to say. Because of the way in which the informational states of his visual areas are changing, the subject will be able to continue producing substantively new reports. However, the subject does not notice any change in the identity of the perceived color, so it will not seem to him that he is describing different colors from moment to moment. Instead, it will seem to him that he is describing a single, unvarying color, which for some reason he cannot quite describe completely, no matter how hard he tries.

In other words, the subject will be afflicted with a kind of ineffability.

5. The Dynamic Narrative

The preceding discussion leaves open the question of the precise nature of the alterations in the subject's narrative. If descriptions of colors are altered but identifications of colors are unchanged, then in what way do the descriptions change?

One answer to this question (though not the only possible answer) is that new reports about the color appear in the narrative. There are many different kinds of reports that might appear; a few of them are:

> (1) Reports expressing intermodal comparisons (such as "Red is a warm color"). Comparisons of this sort are discussed by C.L. Hardin (1988: pp. 128-133) and mentioned by Thomas Nagel (1974: p. 449).
>
> (2) Descriptions that express associations between observed colors and the subject's past experiences. (A particular shade of red may remind the subject more of a certain past sunset now, but more of a certain past rose later.)
>
> (3) Statements about the emotional impact of a color. ("That shade of red is a cheerful color, but when you look at it a while, it seems somehow unsettling.")

There is no compelling reason why descriptions of these sorts, and especially of the last two sorts, should be entirely stable to feedback-driven fluctuations in visual information. According to our assumptions, they may well not be stable. As a result of feedback effects, the subject might report on a color experience differently at different times in each of these three ways.

Utterances of these three sorts will not crop up in a completely random fashion. A subject who has said "Red is a warm color" is not very likely to say "Red is a cool color" later, because certain intermodal comparisons, like "Red is a warm color" or "Red is exciting," have some universality grounded in human physiology (Hardin 1988: pp. 128-131). However, it is possible that the feedback-driven fluctuations might help to trigger different intermodal comparisons that the subject already is physiologically capable of making. Thus, when the subject in our experiment is asked to describe the perceived color at greater length, the ongoing turnover of visual data may lead to the color being described intermodally in multiple ways, even if the resulting descriptions are ones we might have expected (for example, "Red is warm" and "Red is exciting"). The fluctuations may even cause the subject to notice some of these physiologically grounded intermodal resemblances for the first time. The situation is different for the second and third kinds of descriptions listed above. For these, the subject has more leeway and is more likely to discover associations that did not exist before the experiment.

Another way in which the fluctuations can affect the narrative is more complex, and probably more important, than the triggering of new kinds of utterances in the narrative. Recall that the time scale for the fluctuations is shorter than the time scale for production of a narrative. This suggests that the subject might begin to say something about the perceived color and then simply be unable to complete the utterance. Alternatively, the subject might complete the utterance, but then judge that the utterance is not quite correct or somehow misses the mark. The subject might experience either of

these incapacities as a feeling of being unable to find the right words to describe the color, or as an inability to say anything about the color without having the description seem inadequate. This seems, at least to me, to resemble what really happens when one tries very hard to describe a color to oneself.

By altering the narrative in any or all of these ways, the feedback effects may make color experiences impossible to describe in any way that seems satisfactory to the describer.

In real life, of course, it is possible to describe colors to oneself or to others up to a point. The subject in our experiment could have described the perceived color by saying, "It's red," or "It's almost the same color as a ripe apple." The experimenter would have understood these descriptions. Obviously, the phenomenon that philosophers call ineffability does not involve the impossibility of descriptions like these. The explanation of ineffability proposed here leaves plenty of room for descriptions of this sort. The feedback effects invoked in that explanation would not render all the perceived features of a color unstable. As I pointed out earlier, there are good biological reasons to think that some of those features would be stable—in particular, the reported identities of colors and the locations of colors in color space. (This stability is partially captured by assumption (A4).) People's ability to describe experiences to one another is largely based upon a certain kind of intersubjective understanding (see Nagel 1974: p. 442). The stable features of color experience, as portrayed by our explanation, are sufficient to anchor the kind of intersubjective understanding that makes statements like "It's red" useful for communication. For example, suppose that the subject in our thought experiment says to the experimenter, "That is red." The experimenter can understand this utterance immediately because both people have learned how to use the word *red* by way of past contacts with rather similar stimuli—above all, with surfaces illuminated by light of similar wavelengths. For both brains, the production of the word *red* is causally connected with processing of certain kinds of visual information. Some of this information may be fluctuating quite independently in both brains. Nevertheless, this information (or some part or feature of it) is stable enough to allow either person alone to consistently identify red things as red. Thus, the two people end up applying the word *red* to approximately the same things, and can understand each other when they describe things as being red. If they tried separately to describe the color as thoroughly as the subject was asked to do, then their descriptions might differ because of the fluctuations. (Perhaps one would first find the shade of red cheerful and then realizes that it is a warm color, while the other would begin by saying that the color seems unsettling.) But these differences are not of a kind that would unmoor intersubjective agreements about what is red. Thus, our explanation can accommodate the fact that color is, up to a point, describable.

One might ask whether the mechanisms described above are related in any way to the fact, discussed by Jonathan Schooler (2002), that experiences can be modified by efforts to describe them. The changes in content that our hypothesis invokes do not arise from attempts to describe, but from fluctuations that are not consciously noticed by the subject. Nevertheless, one might wonder whether the subject's attempts to describe the experience could affect the feedback-driven fluctuations, and thereby affect the content of

the experience. I will not address this question in the present paper, but it is something to consider.

Our new explanatory hypothesis also may shed some light on problems connected with the inverted spectrum, which is widely discussed in the philosophical literature (see, for example, Hardin 1988: pp. 134-142). According to our hypothesis, the exhaustive description of a color is impossible. Hence, two subjects who compare their reports of color experiences can never be certain that they are having exactly the same color experience. However, partial agreement between the two unfinished reports would constitute evidence that the two experiences are similar in some respects. Thus, it would be possible for the two subjects to establish that the color experiences they are having are similar in some respects. The degree of established similarity would increase if the reports agreed with each other on more points as the reports became longer. Thus, if our hypothesis is correct, we cannot entirely rule out the inversion of spectra between subjects, but subjects can know something about the similarities and differences between their spectra.

It is interesting to note that our explanation of ineffability also may help to explain one of the most notable features of intermodal comparisons—namely, their perceived inadequacy. Such comparisons fit, in a rough-and-ready way, the experiences they purport to describe. Nevertheless, these comparisons are quite inadequate and strained as color descriptions go (see the example in Nagel 1974: p. 449). We discover this easily when we reflect a little upon such comparisons. Most color descriptions of this sort are what Nagel once called "loose intermodal analogies" (Nagel 1974: p. 449). In the same paper, Nagel remarked that such analogies would be "of little use" (1974: p. 449) for a certain possible project that he described in that paper. A similar unfavorable assessment of these analogies seems apposite here. Nevertheless, the fact remains that there is something intuitively right about these intermodal comparisons, despite their weaknesses. The perceived unsatisfactoriness of such comparisons can be explained by our hypothesis about ineffability—for the hypothesis implies that a perceiver's description of a perceived color, if not of a stable sort like color identity statements, will seem unsatisfactory to the describer, especially if some time (>> 100 msec) has elapsed since the description was uttered. If our hypothesis is correct, then by the time one asks oneself whether a particular intermodal description of a color is adequate, one may well find that it is not.

A possible objection to the arguments in this and the previous section arises from the possibility that the production of new reports could be limited by repetitiveness in the output of the visual system. For example, could the stimuli in the experiment cause the subject to utter a series of reports that repeats itself every several seconds? Such repetition is conceptually possible, but it is not likely to limit very effectively the production of new reports. To see why, consider the familiar fact that prolonged concentration on a given stimulus produces an experience different from that caused by a brief exposure to the same stimulus. The brain's response to a stimulus depends, not only upon the intensity and other instantaneous characteristics of the stimulus, but also upon the duration of the stimulus. Through fatigue and familiarization, the response to a prolonged stimulus alters over time. Further, the outputs of the visual system, and especially of the language system, are influenced by things going on elsewhere in the

brain—things that are not all temporally repetitive. Thus, even if the feedback loops in the visual system were physically capable of sending fairly repetitious signals to the language system, the ultimate output from the language system would likely be very imperfectly repetitious at best. A verbal report that resembles an earlier report, but is not a *precise* repetition of that earlier report, actually amounts to a report of something new—perhaps of a new shade of meaning in the earlier report, or of some details not noticed before. Differences like these suffice to make the later report new for our present purposes.

6. Simplicity and Complexity Again

So far, I have described a way in which known neural mechanisms might render a subject incapable of describing color experiences in a manner that seems adequate to that subject. Of course, the mere possibility that these mechanisms might act in this way does not guarantee that they do act in that way. The sort of ineffability proposed here might or might not be the same as the ineffability that people really report.

One objection to the identification of these two sorts of ineffability arises from the feeling that experiences of color, or other putatively ineffable experiences, are simple or entirely unanalyzable. I discussed this feeling in Section 2. A proponent of the simplicity of color sensations might find the emphasis on complexity to be a flaw in our hypothesis.

My chief reply to this objection is that color sensations are not simple. Developments in philosophy and in color science have undermined the traditional belief in the unanalyzability of colors. (I covered this point in Section 2.) Thus, the fact that a hypothesis contradicts the simplicity of color sensations is not a flaw in the hypothesis. Also, our proposed mechanism for ineffability might actually be able to explain why the idea of the simplicity of colors is so intuitively compelling. According to our hypothesis, a perceived color is analyzable to the perceiving subject, but immediately frustrates any attempt by the subject to analyze the color very thoroughly. It seems likely that the subject would describe this insurmountable difficulty in terms of the unanalyzability of the color. The subject, unaware of the neural fluctuations underlying the indescribability, could not describe the situation with a statement like this: "I can begin to describe what this color is like, but something about the color keeps changing before I can finish describing." Instead, the subject would tend to say things like: "I can't describe this color thoroughly, no matter how hard I try." The subject might then spontaneously judge that the color is impossible to analyze. This psychological mechanism might account for the prevalence of intuitions about the simplicity of colors. A similar mechanism is found in Dennett's explanation of ineffability (1991: pp. 382-383), where a subject finds a color experience to be, in a sense, too complex to describe, and hence finds the experience to be ineffable.

It is worth noting that objections based on simplicity are not peculiar to the explanation of ineffability presented here. Dennett's explanation of ineffability also traces ineffability to a certain kind of complexity of data, and therefore is potentially vulnerable to analogous objections. Dennett's explanation already contains ways of rebutting some such objections (see (1991: pp. 49-50, 382-383)); the rebuttal I have given here owes much to these ways.

It is interesting to ask whether an illusion similar to the simplicity illusion could be partly responsible for the counterintuitive character of the hypothesis presented in this paper. Consider the following three probable consequences (all mentioned earlier) of our hypothesis: (1) Upon noticing the ineffability of color experiences, a human subject tends to attribute that ineffability to simplicity. (2) A human subject's color world seems to that subject to be stable and non-fluctuating. (3) A human subject cannot notice the temporal fluctuations that underlie the ineffability of colors. Now imagine yourself trying to convince a subject of this kind that the real reason colors seem ineffable is the complexity (not the simplicity) of colors—and worse yet, a kind of complexity rooted in instability and fluctuations that cannot be noticed. An explanation of this kind would go sharply against the grain of the subject's intuitions about the simplicity, unanalyzability and stability of color experience. For this reason, if our explanation of ineffability is true, then that explanation itself may be strongly counterintuitive for us humans—and not through any fault of our own.

Earlier I pointed out that the hypothesis about ineffability presented in this paper is compatible with Dennett's and Churchland's explanations, in the sense that any combination of these three mechanisms could operate in the same brain and contribute to the ineffability of color experiences. Nevertheless, our new explanation might provide a more natural account of certain features of ineffability, in two respects. First, Churchland's explanation, taken by itself, does not contain an account of the impact of an experience's complexity upon its perceived ineffability. Our explanation does provide such an account. Second, Dennett's explanation leads to an account of the apparent simplicity of color experiences that is less natural than the account arising from our explanation. An experience that embodies high complexity from the beginning is less likely to be judged to be too simple than an experience whose complexity does not become evident until after repeated attempts at description. According to Dennett's explanation, the subject describing an experience is confronted (so to speak) with a lot of complexity all at once. According to our explanation, the complexity never becomes fully evident all at once; this circumstance makes it easier for the subject to judge that the experience really was simple to begin with. Thus, although all three explanations of ineffability could be right, the hypothesis presented in this paper might account for some features of the ineffability of color experiences slightly better than do the other hypotheses.

7. Some Philosophical Implications

Earlier in this paper, I claimed that the explanation of ineffability that I would propose does not have the deflationary philosophical implications of the explanations of ineffability suggested by Churchland and Dennett. Here I will explain what I meant by this, and will try to make this claim plausible.

The explanations of ineffability offered by Churchland and Dennett appear to be capable of accounting for the observed fact that subjects sometimes cannot describe their sensations. These explanations also deny that there is anything genuinely indescribable about the content of sensations. On Dennett's view, a subject who finds a sensation ineffable does so only because of a kind of inability to keep track of all the details of a discriminated feature. If the subject acquired the ability to process more information

about the feature, then the subject might well find that there is nothing ineffable about the sensation. On Churchland's view, a subject who finds a sensation ineffable does so only because of an inability to consciously discriminate other, component features. If the subject acquired a little more discriminatory ability, then the subject might find the ineffable feature to be quite describable.

According to the explanation I am suggesting, a subject finds a sensation to be ineffable because it always is possible, under suitable conditions, for the subject to find something more to say about the sensation. This explanation, like those of Churchland and Dennett, is reductionistic; it reduces the ineffability of sensations to the behavior of neural circuitry whose state, at any given moment, is completely describable in physical terms. However, the new explanation implies that a sensation has, in a certain sense, inexhaustible content. No matter how much the subject talks, there is nothing that the subject can say to describe a sensation completely. The fact that the information content of the subject's brain is limited does not change this.

This explanation of ineffability does not deflate the notion of ineffability to the extent that the other two explanations do. The present explanation does not say that ineffability is only a result of insufficient information processing capacity, or of insufficient feature-discriminating powers. Instead, it suggests that ineffability is a result of the impossibility of describing sensations in any exhaustive manner. This is an obstacle that would not vanish in the face of increased discriminatory powers on the part of the subject. To make this kind of ineffability go away entirely, one would have to remove the relevant feedback loops—and in view of the crucial role that these loops play in sensation, such an operation would severely degrade (or perhaps obliterate) the subject's sensations, too.

If a subject has ineffability of this sort, then even an objective, external observer who knows all the details of the subject's brain would not be able to describe completely how the sensation seems to the subject. This would be the case, not because there is anything mysterious about the sensation, but simply because the required description never can exist all at one time.

Our hypothesis may be able to explain ineffability without denying that sensations really are as ineffable as they seem. This possibility makes the hypothesis philosophically interesting, even though the hypothesis is just as unproven as other proposed explanations of ineffability.

One might ask whether our explanation of ineffability really implies that color experiences are ineffable, or only implies that color experiences seem ineffable. The answer to this question has two parts. If by *ineffable* one means completely unanalyzable, or totally inaccessible to conscious analysis, then the hypothesis says that color experiences only seem ineffable. However, as I pointed out in Section 2, this is not a correct definition of ineffability. By this definition of *ineffable*, no color experience is ineffable, regardless of whether our hypothesis is right—because color experiences are not entirely unanalyzable. If, on the other hand, *ineffability* refers to the behavioral phenomenon of ineffability as described in Section 2, then our hypothesis implies that color experiences really are ineffable.

Note that the mere presence of fluctuations in visual or semantical information does not imply the ineffability of any experience. Only certain kinds of fluctuations will do—such as the kind invoked in our hypothesis.

8. Beyond the Visual System

The explanation of ineffability presented in this paper may be generalizable to other sensations besides those of color. So far, we have concentrated on color for the sake of concreteness and convenience, and because color sensations frequently crop up in philosophical writings about ineffability. However, the mechanism which gives rise to ineffability of colors may also occur in other neural systems besides the visual system. Any neural system in which feedback plays a significant functional role is a potential target for such extension. This includes, for example, the somatosensory system, in which reciprocal corticocortical connections are known to exist (Felleman & Van Essen 1991). If we can generalize the present hypothesis to that system, then we can predict that somatosensory experiences will seem ineffable to subjects. However, it is important to note that not every system which has feedback and feedforward pathways, or which resembles the visual system in other structural respects, has to share the ineffability associated with the visual system. This ineffability depends upon a number of features of the system; in particular, the feedback in the system must be of the right sort to produce significant rapid temporal variations, and the system must interface with the language system in such a way that these variations can influence the formation of narratives. There is no guarantee that all neural systems with feedback and feedforward pathways will have these features. Systems that lack these features will not produce ineffable experiences via the mechanism described in this paper, though they might still produce ineffable sensations by way of other mechanisms. Nevertheless, our hypothesis may be applicable to a wide range of sensations.

We might also consider extending this hypothesis to cover the feedback mechanisms postulated by certain theories about the neurobiological basis of consciousness. A prime example of such a theory is that of Dennett (1991). The "Joycean machine" (1991: p. 280 and elsewhere) central to Dennett's theory is, in essence, an intricate feedback mechanism whose operation ties into ongoing revisions of informational content in various regions of the brain. If a Joycean machine were indeed involved in the causation of consciousness, then by adapting our hypothesis to the feedback occurring in that machine, we might be able to explain the apparent ineffability of a wide spectrum of conscious experiences, even without invoking Dennett's explanation of ineffability.

Harth's views about the neural basis of consciousness (Harth 1993) deserve particular mention here. According to Harth's ideas, feedback loops, including visual feedback loops, are central to the neurophysiology of consciousness. For example, on Harth's view, feedback loops amplify signals and noise in the visual system (1993: pp. 63-73, 83-87). Harth has asked whether such loops might be able to give rise to the subjective character of conscious experience, or, as Harth more specifically puts it, "the subjective *feeling* of being conscious" (Harth 1993: p. 147; italics in original). If the hypothesis suggested in the present paper is true, then feedback loops in the visual system give rise to ineffability, which often is regarded as one of the hallmarks of the subjective

character of conscious experience. Hence the present hypothesis appears to be compatible with Harth's views on consciousness, although it does not depend upon the correctness of those views.

Acknowledgments

I am pleased to thank Prof. Eric Scerri for valuable discussions, and Prof. James W. McAllister for helpful comments, on earlier versions of the paper. Also, I wish to thank a number of anonymous readers for their comments on earlier versions of the paper. To mention some particulars, I am indebted to various readers for the objections concerning feedback stabilization, flicker, simplicity, repetitiveness, and intersubjectivity; for part of the evolutionary argument; for raising the issue of a double notion of ineffability; and for advice about the relevance of the Hardin, Westphal, and Schooler references. My opinions in this paper should not be assumed to be those of any of the above mentioned persons, or of my institution.

References

Bullier, J., Hupé, J.M., James, A., & Girard, P. (1996). Functional interactions between areas V1 and V2 in the monkey. *Journal of Physiology (Paris)*, 90, 217-220.

Churchland, P.M. (1996). The rediscovery of light. *Journal of Philosophy*, 93, 211-228.

Dennett, D.C. (1991). *Consciousness explained.* Boston: Little, Brown and Co.

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

Felleman, D.J., Xiao, Y., & McClendon, E. (1997). Modular organization of occipito-temporal pathways: cortical connections between visual area 4 and visual area 2 and posterior inferotemporal ventral area in macaque monkeys. *The Journal of Neuroscience*, 17 (9), 3185-3200.

Hardin, C.L. (1988). *Color for philosophers: unweaving the rainbow.* Indianapolis: Hackett Pub. Co.

Harth, E. (1993). *The creative loop.* Reading, Mass.: Addison-Wesley.

Hume, D. (1739). *A treatise of human nature.* 2nd edition (L.A. Selby-Bigge, Ed.; revised by P.H. Nidditch). Oxford: Clarendon Press. (This edition published 1983.)

Jackson, F. (1982). Epiphenomenal qualia. *The Philosophical Quarterly*, 32, 127-136.

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

Locke, J. (1689). *An essay concerning human understanding.* Paperback edition. (P.H. Nidditch, Ed.) Oxford: Clarendon Press. (This edition published 1979.)

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

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review,* 83, 435-450.

Schooler, J.W. (2002). Re-representing consciousness: dissociations between experience and meta-consciousness. *Trends in Cognitive Sciences*, 6, 339-344.

Van Essen, D.C., Anderson, C.H., & Felleman, D.J., (1992). Information processing in the primate visual system: an integrated systems perspective. *Science*, 255, 419-423.

Westphal, J. (1984). The complexity of quality. *Philosophy*, 59, 457-471.