A NEW PERSPECTIVE ON REPRESENTATIONAL PROBLEMS

Chris Eliasmith

Dept. of Philosophy
Dept. of Systems Design Engineering
University of Waterloo
celiasmith@uwaterloo.ca

PENULTIMATE VERSION. SEE

Eliasmith, C. (2005). A new perspective on representational problems. Journal of Cognitive Science. 6: 97-123

A New Perspective on Representational Problems

Abstract

I argue that current flaws in the methodology of contemporary cognitive science, especially neuroscience, have adversely affected philosophical theorizing about the nature of representation. To highlight these flaws, I introduce a distinction between adopting the animal's perspective and the observer's perspective when characterizing representation. I provide a discussion of each and show how the former has been unduly overlooked by cognitive scientists, including neuroscientists and philosophers. I also provide a specific neuroscientific example that demonstrates how adopting the animal's perspective can simplify the characterization of the representation relation. Finally, I suggest that taking this perspective supports a specific thesis regarding content determination: the statistical dependence hypothesis.

1 Introduction

As a cognitive scientist, there are at least two possible ways to characterize the contents of internal representations. One is to adopt the typical scientific, objective, perspective on the representational states and contents of an organism, I call this the 'observer's perspective'. Adopting this approach, we would present a stimulus, perhaps a target moving at 1 m/s, to an animal and then determine which mental states were activated by that presentation. Those activated states would

then be considered candidates for representations of that stimulus. Ignoring, for the moment, concerns about having a naïve causal theory of representation and about reproducing the behavior across trials, this approach essentially tells us how likely an observed neural state is, given a stimulus.

Another, largely overlooked, means of characterizing the contents of internal representations is to adopt what I call the 'animal's perspective'. That is, we can look at the mental states of the animal and try to guess how likely it is that a certain stimulus (e.g., something moving at 1 m/s) is in the environment. In order to take this perspective, very different constraints must be placed on defining stimuli and analyzing behavioral responses. In other words, adopting one perspective over another has serious consequences – it is a difference that makes a difference.

In this paper I argue that the observer's approach is by far the most common amongst cognitive scientists, including philosophers such as Dretske, Dennett, Fodor, and Quine, and neuroscientists such as Desimone, Georgopolous, van Essen, and Hubel and Weisel. More importantly, I argue that this approach is very seriously flawed in virtue of being incomplete. I show that rectifying this flaw suggests an alternative characterization of representational content. Although I discuss the close relation between the observer's and animal's perspectives, I begin by distinguishing them in order to highlight the limitations and strengths of adopting either one exclusively. In particular, I present an example that demonstrates how adopting the perspective of the animal can result in a simpler characterization of the representation relation. I conclude by suggesting that the importance of the animal's perspective is properly captured by

a recent proposal for characterizing mental content called the statistical dependence hypothesis (Eliasmith, 2000; Usher, 2001; Eliasmith 2006).

2 Two perspectives, one problem

When faced with scientific problems, such as the problem of characterizing representations, we have had great success in dealing with them from a third person perspective, so a methodological bias in favor of the observer's perspective is only natural. In general, this is an important perspective to adopt in order to construct *objective* solutions – solutions that we can easily share with others. However, when it comes to representational problems, it is not so clear that this is an appropriate viewpoint to take.

Consider the specific problem of 'neurosemantics', that is the problem of how neurobiological systems have contentful states.[1] In addressing this problem, it is the information-processing neurobiological system that is the locus of concern. This scientific question, unlike questions about quarks, molecules, or tectonic plates, concerns something that may have a perspective of its own. If the system of interest (the animal) does have a perspective, and that perspective is relevant to answering the questions we are interested in asking, then we may be able to adopt either the usual observer's perspective *or*, in these special cases, that of the animal under study.

---

[1] This is in obvious analogy to the problem of 'psychosemantics,' more familiar in cognitive science (Fodor, 1981).

A 'perspective', as I shall use the term, is a relation between an information processor and a transmitter of information. Perspective is determined by *what* information is available to an information processor from a transmitter. Notably, we don't have to know what the information is *about* in order to distinguish one set of informational states from another. This is because information-theoretic descriptions can be taken strictly to be descriptions of energy transfer, and we do have a way of tracking energy flow without reference to 'aboutness' (Fair 1979, p. 228). So, by distinguishing 'perspectives' I mean to distinguish information-theoretic descriptions of energy flows. This means that perspectives are commonplace and can be attributed to individual neurons and brain areas as well as to entire brains.

To claim that there is a difference between the observer's and the animal's perspective, then, is to claim that in a given situation, an animal's (first person) perspective and an observer's (third person) perspective provide access to different information. Specifically, an animal (and any of its sub-components) can only access information available through sensory receptors. However, properly situated observers can access *that same information, as well as information available through their own sensory receptors* about the same situation. So, the observer has two sources of information; the animal's receptors, and their own.[2]

---

[2] It is irrelevant to the point being made here that the observer must access the information available from the animal through the observer's sensory apparatus. The fact remains that the observer's perspective includes two distinct sources of information, only one of which the animal's perspective includes.

Most cognitive scientists concerned with representation have adopted the observer's perspective. However, there have been notable exceptions. For example, Fitzhugh (1958) describes a means of determining the nature of the environment given the response of nerve fibers. Just as a brain (or its parts) infer the state of the world from sensory signals, Fitzhugh attempts to determine what is in the world, once he knows a nerve fiber's response to an unknown stimulus. He purposefully limits the information he works with to that available *to the animal*. The 'extra' information available via the observer's perspective is only used after the fact to 'check his answers'; it is not used to determine what the animal is representing. Fitzhugh's is one of the first in a significant line of experimental approaches that has recently been extended in the book *Spikes: Exploring the neural code* (Rieke, Warland et al. 1997). One of the main themes of this book is echoed in this chapter: our representational characterizations can change when we adopt the perspective of the animal.

In his book *Content and Consciousness*, Daniel Dennett (1969) also realized that the animal's perspective is an important one:

> Whereas we, as whole human observers, can sometimes *see* what stimulus conditions cause a particular input or afferent neuron to fire, and hence can determine, if we are clever, its 'significance' to the brain, the brain is 'blind' to the external conditions producing its input and must have some other way of discriminating by significance (p. 48).

However, Dennett does not appear to have realized that adopting the animal's perspective may have important consequences for a theory of content, because he assumes the standard perspective elsewhere in the same book: "[T]he investigators working with fibres in the optic nerves of frogs and cats are able to

report that particular neurons serve to report convexity, moving edges, or small, dark, moving objects because *these neurons fire normally only if there is such a pattern on the retina*" (p.76, my italics; see also pp. 42, 126). In this second quote, and elsewhere, Dennett has assumed that the pattern, *as determined from the observer's perspective*, is what is being represented. However, as he noted in the previous quote, bits of brains don't necessarily represent what whole human observers do.

In contrast to Dennett's ambiguous commitment to the animal's perspective, work in artificial intelligence has generally embraced this perspective. Researchers in this field realize that the problems that agents solve must be solved given only one source of information – sensory input. For example, this kind of 'first-person' strategy is adopted by the influential tradition in machine vision of constructing three-dimensional scenes from basic features (Marr 1982). Nevertheless, theories of representational content in organisms have decidedly *not* taken a cue from such traditions in artificial intelligence. This is, perhaps, not surprising given that researchers in artificial intelligence often distinguish their pragmatic concern for understanding how to solve a given problem from concerns of how the brain *actually* solves such problems. This, of course, doesn't stop such research from suggesting hypotheses about how the brain *might* solve such problems (but, for a neurobiologically motivated critique of some such hypotheses based on Marr's program see Churchland, Ramachandran et al. 1994).

Artificial intelligence researchers, then, tend to share the conviction that trading the third person perspective for a first person perspective not only makes

sense given the kinds of problem at hand, but is also necessary for avoiding unwarranted assumptions about the nature of the environment. In characterizing neurobiological systems, however, most neuroscientists and philosophers adopt a third person perspective. In particular, neuroscientists tend to assume a set space of possible distal stimuli and try to determine how the system reacts to those distal stimuli (and philosophers tend to assume that neuroscientists have a good methodology). This, however, isn't the problem that an animal must solve in the real world. Rather, the set of possible stimuli is unknown, and an animal must infer what is being presented given various sensory cues. In the next three sections, I contrast these two ways of answering questions about the representation relation.

3 One way to find some answers

The standard methodology for approaching representational problems is the intuitive one. If you were asked to determine what states or processes played a representational role in a given system (i.e., to solve the Problem of Representations (Cummins 1989)) a natural approach would be to present the system with various things it would have to represent and to look for the processes and states that are activated by the presentation of those stimuli. This is precisely the current methodology in neuroscience, and one endorsed by many philosophers. Let me consider this approach in these two disciplines in turn.

3.1 The observer's perspective in neuroscience

For instance, the large corpus of experiments performed to characterize shape-related responses in neurons in early parts of visual cortex such as V1, V2 and V4 adopt this methodology (Knierim and Van Essen 1992; Gallant, Braun et al. 1993; DeYoe, Carman et al. 1996; Callaway 1998). First, a neuron is found with a recording electrode and its receptive field is determined. The receptive field of a neuron is the part of the visual field that, when occupied by a stimulus, causes the neuron to respond (i.e., to fire above its base firing rate). The neuron's preference for color and other non-shape related features is also determined. All the stimuli presented to the neuron have the non-shape related features it prefers. Now, a set of predetermined stimuli, such as crosses, oriented bars, spirals, and sinusoidal gratings, are presented to the neuron and its responses are recorded. The experimenter then proceeds to characterize the responses of the neuron over a series of trials in order to account for the variability of responses to the same stimuli. What the experimenter is constructing, then, is the conditional probability function that a certain neural response, r, occurs given a stimulus, s: $p(r|s)$. So if we are told, for example, that a spiral is in some neuron's receptive field, we can use the probability function we have constructed to predict how that neuron is likely to behave. Presumably, if the experimenter picks enough different stimuli to present to a neuron, he or she will be able to get some sense of what the neuron is representing, that is, to what dimensions (e.g., curvature, length, etc.) it responds.

This kind of experiment has been performed since Hubel and Wiesel's (1962) classic experiments in which they identified cortical cells selective to the orientation and size of a bar in a cat's visual field (such neurons are often

problematically called ‘edge detectors’). The ‘bug detector’ experiments of Lettvin et al. (1988/1959), perhaps better known to philosophers, take a similar approach. In the ‘bug detector’ experiments, retinal ganglion cells (i.e., ‘bug detectors’) were found that respond to small, black, fly-sized dots in a frog’s visual field. More recently, this method has been used to find ‘face-selective cells’ (i.e., cells that respond strongly to faces in particular orientations) in monkey visual cortex (Desimone 1991). In fact, because the Hubel and Wiesel studys were so influential, nearly all single electrode experiments done in cortex follow this basic methdology, whether in parietal cortex (Andersen, et al. 1985), occipital cortex (Newsome and Pare, 1988), temporal cortex (Desimone 1991), motor cortex (Georgopolous, et al. 1986), or frontal cortex (Boch and Goldberg, 1989). In all of these cases, what is deemed important is recording how a neuron responds to known stimuli. In other words, the observer’s perspective is adopted, since both the neuron’s response and the nature of the stimulus (e.g., edges, flies, and faces) are used to characterize the neuron’s behavior.

This method clearly dominates neurophysiological research (Gross, Rocha-Miranada et al. 1972; Zeki 1980; Felleman and Van Essen 1991; Roelfsema, Lamme et al. 1998). It is also the method used by neuroscientists to determine the relata of the representation relation (i.e., to solve the Problem of Representation (Cummins 1989)). In the case of face-selective cells, the representation relation cam be completed as follows: {the neuron that is being recorded from} represents {that face *x* degrees from *y* degrees (where *y* degrees is

the preferred orientation of the cell)} with respect to {the monkey's brain}.[3] These are presumed to be the right relata because, in order, the neuron responds to the stimulus, the observer knows that the stimulus is a face at *x* degrees from *y* degrees, and the neuron doesn't respond that way outside of the monkey's brain. Notice the central role of the observer's perspective in determining the relata in the representation relation. The precise content of a given neural firing is determined by the observer's *independent* knowledge of the stimulus. It is, in general, dangerous to have such *a priori* (with respect to the animal) commitments determine the results of an investigation. After the next section I discuss how we can, at least partially, avoid this result by adopting the animal's perspective.

3.2 The observer's perspective in philosophy

First, however, it is important to show that philosophers have adopted related tactics in trying to characterize the representation relation,[4] as they are the ones directly concerned with the theoretical foundations of cognitive science. Consider, for example, Fred Dretske's (1988) approach.

---

[3] I take it that characterizing the representation relation requires filling out the schema: {representation} represents {content} with respect to {system}. Only the first two elements of the relation are of interest here. And, furthermore, nothing central to this paper hangs on the choice of this particular schema.

[4] To be clear, I take it that both philosophers and cognitive scientists are interested in the same problem when it comes to the representation relation. This is why the philosophers I discuss here take their project to be naturalistic. Similarly, this is why the cognitive scientists I discuss are not interested in the evolution of representation, or the learning of a particular represention per se, but in the underlying (metaphysical) relation between internal and external states.

Dretske argues that the problem of representation only arises for systems that use intrinsic indicators as representations (e.g., the 'bug detector' cells representing bugs to a frog). To understand this kind of representational relationship he calls neuroscience to his aid. He accepts 'bug detectors' as representations of edible bugs because neuroscience has shown that particular cells fire when given bug-like stimuli (ibid., pp. 68-9). So the representational relation is the causal one between bugs and neural firings; the causal relation that is described by the conditional probability of the neural firings given the presence of bugs. Dretske is not alone in this kind of appeal to neuroscience. Philosophers have often thought that the details of cognitive function could be left to neuroscientists (see e.g., Dennett 1969; Millikan 1984; Churchland 1986; Churchland 1989; Dennett 1991).

But, Drestke is a particularly interesting case because he *seems* to be interested in the conditional probability that there is a stimulus in the environment given a response (i.e., $P(s|r)$), not the related, but converse probability function which neuroscientists are constructing (i.e., $p(r|s)$).[5] This is important because, as I discuss in more detail in the next two sections, I think $p(s|r)$ has been mistakenly ignored. But, if Dretske explicitly discusses $P(s|r)$, how can I claim that it has been ignored? The reason is that Dretske (1981) claims that $P(s|r)$ has to be equal to one, i.e., he claims that there *has to be* the stimulus in the environment given a

---

[5] For notational clarity, I should note the difference between $P(x,y)$ and $p(x,y)$. The former is a particular, real-valued probability (i.e., the probability that specific events x and y occur together), whereas the latter is a function which describes the likelihoods for all combinations of the random variables X and Y (i.e., the probability *function* that maps the events $X=x$ and $Y=y$ for all x and y to their probabilities). Of course, the two are closely related since

particular neural response[6] in order for that response to carry information about the stimulus. This is to say that *if* there is a given neural response *then* there is a given stimulus. In effect, then, Dretske has turned the probability statement into a logical one by forcing the unity criteria on the probability.

There are two problems with this result. First, from an experimental point of view, this condition on neural meaning prevents Dretske's analysis from having any methodological import. It is never the case, after all, that probabilities of this kind, as measured experimentally, are one. Therefore, on Dretske's analysis it is never the case that a measured neural response can be said to carry information about a stimulus.[7]

Second, and more importantly for my purposes, Dretske's criterion can only be satisfied by adopting a rather extreme form of the *observer's* perspective; the observer must be ideal. In particular, the observer must have complete knowledge of channel conditions, the animal's background knowledge, and the state of the stimulus in order to verify that a given response carries information about a stimulus. For these reasons, Dretske's theory does not adopt what I have been calling the animal's perspective. That is, Dretske's theory eliminates the perspectival nature of $P(s|r)$ by forcing a criterion of a unitary conditional

---

$p(x_k,y_k)=P(X=x_k,Y=y_k)=p_k$. Using $P(s|r)$ here makes little difference to my central point, but more accurately reflects Dretske's discussion.

[6] More precisely, Dretske claims that $P(s|r)=1$ given background knowledge and certain channel conditions. These two extra conditions make no difference here.

[7] Dretske may claim that his is a metaphysical reduction of the notion of representation, but he then must explain why all empirically characterized representation relations, none of which meet his criterion, are still considered representation relations. And, even if he succeeds in offering such an explanation, he must tell us why the original criterion in conjunction with this explanation should be preferred over an account that doesn't necessitate further explaining.

probability; all relevant information must be available in order to determine that this conditional probability is one. Since the animal's perspective is defined by a limit on information available from a transmitter, and there can be no limits on the information available under Dretske's characterization, Drestke's theory clearly does not adopt the animal's perspective in the relevant sense.

Even those philosophers who, unlike Dretske, reject neuroscience as the arbiter of cognitive theories have generally accepted the standard methodology – normally by placing psychology in neuroscience's stead. Quine (1960), for example, motivated by his behavioristic tendencies, warns that we should steer clear of looking "deep into the subject's head" or at the subject's "idiosyncratic neural routings" (p. 31). In contrast, Quine describes in great detail experiments in which we are asked to evaluate the response of a subject given some stimuli (e.g., a rabbit). In effect, Quine argues that even if the conditional probability of some response (e.g., the word 'gavagai') given some stimulus (e.g., a rabbit) is equal to one, we still can't make claims about what the stimulus is being seen *as* (e.g., a rabbit, or undetached rabbit parts). What is important for my purposes is that the conditional probability that behaviorists like Quine are interested in is still that of the response given the stimuli; it is this conditional probability that is constructed under the standard methodology.

The same is true of philosophers motivated by *cognitive* psychology, such as Fodor (1975, p. 34-7). For example, in Fodor's discussion of concept learning, he takes it that a subject's response profile is what is modeled by psychological theories. What psychologists are doing, then, is recording the subjects' responses to a known set of stimuli. This allows them to achieve their goal of predicting

subjects' responses knowing the presented stimuli. In order to do this, they have effectively constructed the same conditional probability function as the behaviorists and neuroscientists: the probability of a response given a stimulus.

These examples from the various disciplinces of cognitive science, though only a small sample, show a convergence on a particular *methodology* for characterizing the representational properties of cognitive systems. They depend on the assumption that constructing the conditional probability function of the likelihood of a response given a stimulus is the best way to characterize the relation between representations and sensory stimuli.

4 The strangeness of taking the familiar route

Neuroscientific experiments such as those discussed above are intended to address *both* of Cummins' representational problems because they help to characterize a physical process that is correlated with external stimuli, and they then use that correlation to determine the relata of the representation relation. This experimental paradigm is geared towards characterizing the neural response objectively, that is, for a third party observer. Because there are so many sources of uncertainty when applying this kind of approach to a complex system, the measurements of the output vary, even with well-controlled inputs (see section 5 for a simple example). Not surprisingly then, we construct histograms that tell us the probability of getting a particular output given the input. From this third person perspective, the inputs are well defined and the outputs are probabilistically related to the inputs. In other words, it just makes sense to

construct the conditional probability of the indeterminate output given the determinate input. That probability function, what I have been calling $p(r|s)$, is a means of describing the physical processes inside the system we are probing.

If we take a step back for a moment and think carefully about the problem neuroscientists and philosophers are both trying to address, this approach begins to seem a little odd. In the end, we are interested in understanding the problem of neurosemantics. That is, we want to know how, and in what way, *animals* (or their information processing parts) rely on internal states to stand for things in the outside world. And, we want to know what the relation is between those internal states and the things in the outside world. We don't want to know (just) how to cause certain internal states in an animal. But, constructing conditional probabilities of the response given the stimulus tells us how to control the animal with known stimuli, not how the stimuli could be inferred from the responses, or, more importantly, what the relation is between the two.

This response-given-stimulus conditional probability may make sense from our perspective, but, and this cannot be overemphasized, *that conditional probability makes no sense from the perspective of the animal*. In the real world, an animal (or its information processing parts) must try to coordinate behaviors based on the neural firings from its sensory apparatus. There is no sense in which the animal could know what stimulus is being presented prior to having some set of neurons activated; this far, Dennett (1969) is right. This is important for characterizing the representations in neurobiological systems because, in the frog for example, that neural activity is used by subsequent neurons to detect and react to bugs; bugs aren't somehow *used* to cause neural firings.

Another way of thinking of this difference is to realize that constructing the response-given-stimulus conditional, $p(r|s)$, captures the process that *generates* neural responses. If we present a certain stimulus to a neuron, we can (approximately) determine the response we expect the neuron to generate. This is a different problem from *inferring* the stimuli in the world from the neural response. In this second case, we would try to (approximately) determine what stimuli had caused the response we see.[8] If we want to understand how an animal can use its neural representations, we want to understand how it can make such inferences, not just how neural action potentials are generated.

Perhaps the reason neuroscientists and philosophers haven't tried to understand neural function in terms of the conditional probability I am arguing for (i.e., $p(s|r)$) is a methodological one. Perhaps, in other words, it is just easier to find $p(r|s)$ than $p(s|r)$ and *that* explains why we have to adopt the perspective supported by former instead of that supported by the latter. But this doesn't seem to be the case.

First, we must realize that the statistical relation that we are *most* interested in capturing is the combined (or joint) probability function that describes the likelihood of a stimulus *and* a response, $p(s,r)$. This function describes the probability that the stimulus, s, and the response, r, occur together (or with some suitable delay). The reason we are most interested in this joint probability function is because it captures *all there is to know* about the probabilistic relation between a stimulus and a response. From the joint

---

[8] This can be undertaken by an observer, and nevertheless not adopt an observer's perspective

probability function we can determine the marginal probability functions ($p(s)$ and $p(r)$) as well as either conditional probability function ($p(s|r)$ and $p(r|s)$). In other words, there is nothing more to know about the relation between the two variables r and s than what there is to be found in the joint probability function.

There are three ways of determining (or, more realistically, *approximating*) a joint probability function. The first is to determine it experimentally. That is, we can randomly present a set of stimuli that drive a cell, record the firings and construct the joint histogram. Notably, this is not the same as showing stimuli and constructing a histogram of the response probabilities for each stimulus (i.e., the standard methodology). In the next section I discuss a specific example of this difference. The second and third ways of determining the joint probability function are either: 1) to find it from the response-given-stimulus probability, $p(r|s)$, if we know the probability of the stimulus, $p(s)$ as in equation (1); or 2) to find it from the stimulus-given-response probability, $p(s|r)$, if we know the probability of the response, $p(r)$ as in equation (2).

$$p(s,r) = p(r \mid s) \cdot p(s) \quad (1)$$

$$p(s,r) = p(s \mid r) \cdot p(r) \quad (2)$$

Given these three ways of determining the joint probability function, we can learn something quite interesting about the methodological assumptions of traditional neuroscience and philosophy. Namely, that efforts have been focused on characterizing only *part* of the relationship between stimuli and responses. In particular, $p(r|s)$ has been characterized, but this isn't all there is to know about

---

about what is being represented. I discuss this more fully in section 5.

the relation between a stimulus and a response. In order to completely characterize the relationship, we also need to know $p(s)$ as in (1).

The importance of the probability of a stimulus occurring, $p(s)$, is often overlooked by the standard methodology. If we aren't careful about $p(s)$, then our choice of stimuli to present to a neuron can greatly skew our estimate of the joint probability function and we will mischaracterize the relationship between stimulus and response. For example, if I present only one stimulus over and over, the probability of that stimulus will be one, and the joint probability will be equal to the conditional probability, $p(r|s)$. This, of course, isn't because that's what the joint probability *really is*, but rather because my choice of $p(s)$ is a particularly bad one, one that is unlikely to represent the probability of naturally occurring stimuli. In order to get a good estimate of the joint probability, we need to have a guess as to what $p(s)$ is. As important and difficult as generating that guess may be, it is not relevant for my purpose of showing that the standard methodology isn't simpler. What *is* important is that we *must* put a lot of work into determining $p(s)$, or we will poorly characterize the relationship we are after.

In the case of determining $p(s|r)$, we seem, at first glance, to be at a methodological disadvantage. We can't, after all, force the neuron to have a response and then see what the stimulus that caused it was. However, from (2), it is plain that we can characterize this conditional probability if we characterize the joint probability function first. Furthermore, we don't need to worry about $p(r)$ here (as we needed to worry about $p(s)$ under the traditional methodology) because it can be calculated directly from our estimate of $p(s,r)$ (by marginalizing the joint probability function). But, estimating the joint probability function isn't

easy. We need to present the neuron with a good selection of stimuli, and to record the responses of the neuron. What do I mean by a 'good selection'? Well, the naturally occurring $p(s)$ would be a good selection. That, of course, is just what we needed to know in order to properly characterize the relationship between stimulus and response in the traditional methodology. In other words, we need to know just as much about the probabilistic relationships (i.e., we have to make the same tough guesses) in determining $p(s|r)$ from (1) via the joint probability function, as we need to know in order to properly characterize the stimulus/response relationship under the standard methodology.

In sum, characterizing the complex relationship between the environment and an animal's internal representations is no more difficult from one perspective than from the other. Furthermore, there are a number of considerations in favor of adopting the animal's perspective. In particular, it's what the animal must do, and *that* is what we are interested in understanding. So, taking the third person perspective, that is, adopting the traditional methodologies of neuroscience and philosophy, may not be the best bet in solving the interesting representational problems. The alternative is, of course, to adopt the perspective of the animal.

5 The other way to find some answers

Though constructing the response-given-stimulus conditional probability, $p(r|s)$, is by far the most prevalent means of trying to understand representation in neurobiological systems, it is not the only one. The alternative, as just discussed, is to construct the stimulus-given-response conditional probability, $p(s|r)$.

Fitzhugh (1958) suggests embracing this latter approach, though his suggestion does not seem to have attracted much interest until recently (Bialek, Rieke et al. 1991; Theunissen and Miller 1991; Abbott 1994; Mainen and Sejnowski 1995; Rieke, Warland et al. 1997). In this section I discuss a specific example that shows the significant difference adopting one perspective over the other can make.

I have already suggested a few reasons why the animal's perspective may be important for characterizing representation. But are there reasons to think the animal itself could or does use the stimulus-given-response conditional? For the animal to do so, according to equation (1), it would need to take advantage of the joint probability function (or an estimate of the joint probability function) and the probability of a response occurring. In other words, before anything else, the animal needs an internal statistical model of the environment's relation to its neural responses. The simple fact is, we have to start with a model of the stimulus before we can construct the probability of a stimulus given a response. Fortunately, there is evidence that young animals, including children, do have a sense of the statistical structure of their world (Soja, Carey et al. 1991; Spelke and Van de Walle 1993). For example, there is evidence that children, at the tender age of three months, perceive object unity (Spelke and Van de Walle 1993, p. 134). These sorts of results suggest that animals come into the world with innate mechanisms that help them guess at what stimulus in the environment causes

some particular neural firings.[9] Of course, these initial models can be updated on the basis of experience.

Having to begin life with a statistical model of the world may seem unduly nativist to many. However, such models don't need to be very detailed (or even very good) to be useful (Friston, 2003). Researchers in machine vision have taken advantage of this fact and applied it to object recognition. They have turned from traditional 'descriptive' models that are learned from scratch to 'generative' models that *assume* an initial model and then *build up* better representations on the basis of that assumed model and experience (Frey and Jojic 1999). Using these new approaches, researchers have been able to solve some traditionally difficult problems with computationally simple algorithms and very general models of the statistical structure of the world. So, not only is it possible to construct stimulus-given-response conditional probabilities (as outlined in the last section), but doing so is both biologically reasonable and has lead to advances in fields solving related problems. These are two good reasons to think this may be a fruitful approach.

But, what about an actual neurobiological system solving an actual neurobiological problem? Since 1988, Robert de Ruyter van Steveninck and William Bialek have worked to characterize the motion processing system in the blowfly (de Ruyter van Steveninck and Bialek 1988; Rieke, Warland et al. 1997). The neurons they are particularly interested in are called H1 neurons and are

---

[9] This innateness claim is actually quite weak and is generally admitted by both 'nativists' and 'non-nativists' alike (Chomsky and Katz 1975, p. 70; Fodor 1981, p. 275).

about 4 synapses away from the fly's photoreceptors. These neurons show a high sensitivity to the velocity of stimuli in the fly's environment.

By tethering a fly, and recording from an H1 neuron for an extended period, these researchers were able to build up a good estimate of the joint probability of velocity and firing rate. With this data, they directly compared the difference between using the stimulus-given-response conditional probability and the more traditional response-given-stimulus conditional probability (see Figure 1).

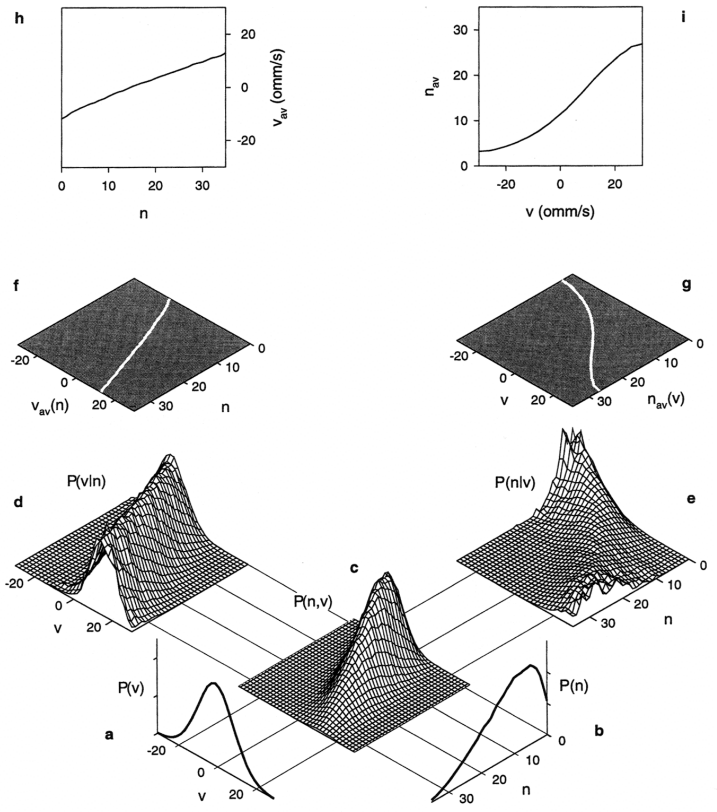

Figure 1: Joint, marginal, and conditional probability functions (a, b, c, d, e), and the differing characterizations of the stimulus/response relationship (f, g, h, i) depending on the conditional used (from Rieke, et al. 1997).

Figure 1 demonstrates the important differences that can arise from taking the animal's perspective instead of the observer's perspective. Beginning at the bottom of this figure, (a) and (b) show the probabilities of a stimulus (velocity) and of a response (number of neural spikes in a time window) respectively, for some H1 neuron. These are the marginal probability functions of the joint probability of the variables, which is shown in (c). From (c) we can discern that there is a statistical dependence between the two probabilities in (a) and (b) since $p(n, v) \neq p(v) \cdot p(n)$. This is as we would expect if the neural response is related to the velocity. The next two graphs, (d) and (e) are generated using equations (1) and (2) of the previous section, and show the conditionals $p(v|n)$ (i.e., $p(s|r)$) and $p(n|v)$ (i.e., $p(r|s)$) respectively. A graph of the best estimate of the velocity given some response is shown in (f) and (h). As is standard practice, this best estimate is presumed to be the average. These two graphs, then, characterize the problem from the perspective of the fly. The best estimate of the response given some velocity is shown in (g) and (i). These two graphs characterize the problem from the observer's perspective.

As can be seen by comparing graphs (h) and (i), adopting the fly's point of view results in a much more linear relation between the stimulus and response (i.e., the function from one to the other is nearly a straight line) than does adopting the third person perspective. In fact, (i) looks much like the standard sigmoid function used in many artificial neural networks, and determined by many neurobiological experiments. This relation between stimulus and response, found by adopting the observer's perspective, is extremely nonlinear. In general, if we can characterize a system as linear, it will be much easier to analyze than if

we have to deal with the inherent complexities of nonlinear responses. In this sense, our description of the problem is much simpler if we adopt the animal's perspective over that of the observer. As well, this result is encouraging because it suggests that particular instances of the representation relation in neurobiological systems may not be unduly complex (i.e., nonlinear instead of linear) if we adopt the appropriate perspective.

6 The baby and the bath water

If the animal's perspective is advantageous, as this result suggests, should we abandon cognitive science as traditionally done? The answer is no. I have been intentionally overstating the case for the differences between these two methodologies to show the strengths of the alternative. In fact, the two approaches are deeply connected. If we look again at equations (1) and (2), we can see precisely what that connection is. In particular, equating the right hand sides of both equations leads to:

$$p(r \mid s) \cdot p(s) = p(s \mid r) \cdot p(r) \qquad (3)$$

This equation is known as Bayes' rule. What it tells us is that if we can completely characterize one of the conditional probability functions, along with $p(s)$ and $p(r)$, then we can completely characterize the other. However, complete characterization of unknown probability functions through sampling is extremely difficult. So, rather than discarding one methodology in favor of another, we should try to characterize these probability functions in as many ways as possible. This gives us multiple means of discovering the same underlying probability

function, $p(s,r)$. And this kind of cross-validation is an invaluable tool for any scientific enterprise.

So far, however, researchers have approached the problem from mainly one standpoint – that of the observer (and only partially so, as $p(s)$ is often ignored). Not only would it be more ecumenical, but it would also be better science to use all of the tools we have available. If our estimates of the joint probability function converge, then our confidence in the accuracy of the estimate would be significantly greater than an estimate from only one source. Convergence is never a bad thing.

The tight relation between $p(s|r)$ and $p(r|s)$ also helps show what the real difference is between the two approaches. As I argued in the last section, the amount of work involved in getting at either conditional is about the same. So, this methodological switch wouldn't be about saving time. Rather, it is about constructing the right conditional probability in the right way, or more importantly, under the right assumptions. Dretske argued for constructing the right probability, but his assumptions about the nature of that probability lead to difficulties. We must not only construct this probability, but also do so under the assumption that the animal has no *a priori* access to the nature of the stimulus. The animal may have some innate statistical model, but it doesn't have to be one that exactly mirrors the statistical structure of stimuli in the environment as Dretske's criterion mandates.

Another way of stating this 'no *a priori* access' assumption is: we should not adopt the observer's perspective about *what* is being represented. So far, I have been suggesting this by claiming that we must take the animal's perspective

and not the observer's perspective. But, strictly speaking, we can't *literally* adopt the perspective of the animal, because we aren't literally the animal. Rather, we *must* take an observer's perspective because we *are* observers. What I mean to say, then, is that we should *direct* our third person perspective *through* the animal. This is the real difference between the two perspectives. The observer's perspective is a third person perspective, *simpliciter*. What I have been calling the animal's perspective is still technically a third person perspective, but it is 'filtered' through the animal; we limit our access to the animal's information channel when representing the world (even though we can use *our* channel to help verify the inferences we make on the basis of the animal's perspective).[10] And, this is an important difference, as the blowfly example shows.

In section 3, I mentioned that we could avoid having *a priori* commitments determine detailed content ascriptions. In the case of the monkey face-selective cells, taking the standard perspective leads to a characterization of the representation relation as: {the neuron that is being recorded from} represents {that face *x* degrees from *y* degrees} with respect to {the monkey's brain}. So, if the experimenter presents a stimulus at 45 degrees from center, and there is an increased probability of response from a neuron, the experimenter may claim that the neuron represents the stimulus 45 degrees from center. Notice, of course, that this content is *completely determined* by the choice of stimulus presented by the observer. In other words, the content is {that face *x* degrees from *y* degrees},

---

[10] I should note that how the animal gets to this particular state (i.e., gets to have this particular information channel) is not one with which I am concerned here (presumably, this is the role of learning).

because the observer *knows* that the stimulus is *x* degrees from *y* degrees, having presented that as the stimulus.

If, instead, we attempted to determine the representation relation from the animal's point of view, we would first construct the joint probability function of, say, the firing rate and the orientation of the stimulus. We would then find $p(s|r)$ and, given a firing rate, we would determine the best guess as to *s*. So, the representation relation may look much the same: {the neuron that is being recorded from} represents {that *there is a* face *x* degrees from *y* degrees} with respect to {the monkey's brain}. This minor terminological change (i.e. the introduction of 'there is a'), denotes a very important difference in possible content. Consider, again, the experimenter presenting a face at 45 degrees from center. Adopting the animal's perspective, the increased firing resulting from that stimulus is used to estimate the orientation of the face. It could well be that, given various lighting effects, occlusion, etc. that the particular increased firing is more likely to indicate a face 50 degrees from center. This, then, is the content of that representation for the animal. And, this is clearly *not* the same content as determined by adopting the observer's perspective. This difference can be expressed by noting that in the first case, the content is identical to the stimulus, but in the second case, the content is a property ascription in the form of an hypothesis about the world. So, the stimulus is the same in both cases, but the content is different. Under the standard methodology, content is determined by *a priori* knowledge about what is being presented to the cell. Under the alternate methodology, the content is determined by statistical inference from a firing rate to a likely stimulus. Thus, the displacement determined by this second method

could be different from that of the actual stimulus. This is not so under the standard methodology. These, then, are definitely *not* the same characterization of the representation relation.

7 The statistical dependence hypothesis

My discussion so far has focussed on the methodological side of typical representational characterization. But I think there is also a more theoretical lesson that can be drawn from these considerations. In other words, taking the alternate methodology seriously provides important insights into the nature of representational content. Recall two things that we have learned: 1) the joint probability distribution completely characterizes the relation between stimulus and response variables; and 2) neurons are said to represent what they have statistical dependencies with (under both methodologies). I think we can put these claims to work for a theory of content.

First, given that joint probabilities fully characterize the relation between stimuli and responses, if we had the set of all joint probabilities between any stimulus and the responses of some neuron, we would have a complete characterization of how that neuron relates to any particular stimulus. Second, responses are said to represent what they have dependencies with. Presumably then, it makes sense to say that the things (objects, events, properties) a neuron best represents are what it has its highest statistical dependency with. Furthermore, a neuron can be a better 'stand-in' for what it has the highest statistical dependence with than for anything else. Since representation is

'standing-in', and content is partly what is 'stood-in' for, we would say that a neuron's content is (at least partly) what it has this highest statistical dependence with.

Putting these two claims together results in a hypothesis about the nature of meaning in neurobiological systems. I call this the statistical dependence hypothesis (Eliasmith, 2000; Usher, 2001; Eliasmith 2006):

> *The set of causes relevant to determining the content of neural responses is that set that has the highest statistical dependence with the neural responses under all stimulus conditions.*[11]

Notice that the hypothesis suggests that content is determined by response*s*, not a single response. Response *profiles* statistically depend on sets of causes, not momentary responses. It is well known that neurons have graded responses to stimuli. In this sense it is misleading to call them 'detectors' of any kind. Neurons don't 'detect' things (i.e., they don't determine that there *is* an edge or there *isn't* one), they respond selectively to input; the more similar the input, the more similar the response. The statistical dependence hypothesis highlights this ubiquitous, often ignored, property of neurons.

The statistical dependence hypothesis says that given a complete characterization of how a neuron (or a group of neurons) responds via the set of all joint probabilities (i.e., the set of joint probabilities under all stimulus

---

[11] Hyvarinen (1999) notes: "The mutual information is a natural measure of the dependence between random variables." (p. 107). Average mutual information between random variables is defined as $I(x;y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \rho(x,y)\log\frac{\rho(x,y)}{\rho(x)\rho(y)} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \rho(x,y)\log\frac{\rho(x|y)}{\rho(x)}$, so the mutual information of two events is $I(a;b) = \log\frac{\rho(a|b)}{\rho(a)}$. Usher (2001) adopts mutual information as a means of understanding representation, as I have elsewhere (Eliasmith, 2000).

conditions), the causes relevant to content of that neuron's (or group's) response are those that its (their) response profile corresponds to the best. We would expect content to be (at least partly) determined by the *best* corresponding neural responses because those responses carry the most information about the relevant causes. Notably, this doesn't assume that representations are 'normally right' – representations have all kinds of statistical dependencies, not just the best one. But, neural responses are, in a sense, about what they are the best at being about.

The statistical dependence hypothesis is about what we should take neurons to mean; i.e., how we should determine their content *in general*. But, what about active, real-world representing? How do we know what this particular representation that is active right now is about? How do we know what it has as a referent? I think a more limited version of the same hypothesis helps answer these questions. I'll call this corollary the occurent representation hypothesis:

> *The referent of an occurent representation is the cause that has the highest statistical dependency with the representation under the particular stimulus conditions in which it is occurent.*

This hypothesis, then, serves to tell us that, right now, *this* representation is about *that* thing in the world.

A simple example should help to clarify the application of both the statistical dependence hypothesis and its corollary. Consider, again, an H1 neuron in the blowfly. According to the statistical dependence hypothesis, the meaning carried by this neuron is determined by its highest statistical dependence under all stimulus conditions. Given past experiments, the response profile of this neuron is most highly dependent on horizontal velocity in the visual field under all stimulus conditions. Now, what do we say when a particular stimulus is moving

in the visual field? We say that the referent of the representation is that stimulus, since, under these conditions it has the highest statistical dependence with the neural response. And, we say that the neural response *means* that there is such-and-such a velocity in the visual field. If, however, we flashed a number of stimuli in quick succession, providing the illusion that there was movement[12] and resulting in a response from this H1 neuron, things would be different. We would then say that the referent of the response was the set of stimuli events (since they have the highest statistical dependence with neural firings under these conditions). However, we would still say that the neuron means that there is such-and-such a *velocity* in the visual field (even though there isn't) because under *all* stimulus conditions it is velocity that this neuron picks out.[13] This is simply a case of misrepresentation.

There are many more things that need to be said about the statistical dependence hypothesis. While I have discussed in greater depth elsewhere (e.g. Eliasmith 2006), the point in this instance is merely that an at least *prima facie* new way of understanding representational content falls directly out of the previous methodological considerations. In fact, one way to understand the flaw in adopting the observer's perspective is that it results in a blurring of referent and content. Notice that the perspective of the observer incorporates two sources of information when determining content; i.e., both what the observer takes the

---

[12] This effect is called the phi phenomenon by psycholgists and is well exemplified by a marqee (see Sarris 1989).

[13] This raises the philosophical worry about how we can justify distinguishing 'velocities' from 'nearby flashes' as distinct sets of causes. I consider these worries elsewhere (Eliasmith, 2000).

stimulus to be *and* how the animal's perceptual system responds to the stimulus are included. Adopting the animal's perspective makes it quite clear why and how we should keep these two sources separate. Similarly, the statistical dependence hypothesis and its corollary provide a way to understand meaning that makes this distinction explicit.

8 Conclusion

There are significant shortcomings of the traditional characterization of representational content in cognitive science as a result of the nearly univocal adoption of the observer's perspective. I have argued that there is an important alternative, the animal's perspective, that, when investigated in greater detail, results in new theoretical insights into the nature of representation. Furthermore, this distinction between perspectives highlights precisely what information is needed to properly characterize the representation relation. Undoubtedly the best means of gathering this information is to adopt *both* perspectives, keeping in mind their complimentary strengths and weaknesses.

9 References

Abbott, L. F. (1994). "Decoding neuronal firing and modeling neural networks." *Quarterly Review of Biophysics*. **27**(3): 291-331.

Andersen, R. A., G. K. Essick, et al. (1985). "The encoding of spatial location by posterior parietal neurons." *Science* **230**: 456-458.

Bialek, W., F. Rieke, et al. (1991). "Reading a neural code." *Science*. **252**: 1854-57.

Boch, R. A. and M. E. Goldberg (1989). "Participation of prefrontal neurons in the preparation of visually guided eye movements in the rhesus monkey." *Journal of Neurophysiology* **61**: 1064-1084.

Callaway, E. M. (1998). "Local circuits in primary visual cortex of the macaque monkey." *Annual Review of Neuroscience*. **21**: 47-74.

Chomsky, N. and J. Katz (1975). "On innateness: A reply to Cooper." *The Philosophical Review*. **84**(1): 70-87.

Churchland, P. (1989). *A neurocomputational perspective*. Cambridge, MA, MIT Press.

Churchland, P. S. (1986). *Neurophilosophy*. Cambridge, MA, MIT Press.

Churchland, P. S., V. S. Ramachandran, et al. (1994). A critique of pure vision. In C. Koch and J. Davis (Eds.), *Large-scale neuronal theories of the brain*. Cambridge, MA, MIT Press.

Cummins, R. (1989). *Meaning and mental representation*. Cambridge, MA, MIT Press.

de Ruyter van Steveninck, R. and W. Bialek (1988). "Real-time performance of a movement sensitive neuron in the blowfly visual system: Coding and information transfer in short spike sequences." *Proceedings of the Royal Society of London Ser. B*. **234**: 379-414.

Dennett, D. C. (1969). *Content and consciousness*. Cambridge, MA, MIT Press.

Dennett, D. C. (1991). *Consciousness explained*. New York, Little, Brown and Company.

Desimone, R. (1991). "Face-selective cells in the temporal cortex of monkeys." *Journal of Cognitive Neuroscience*. **3**: 1-8.

DeYoe, E. A., G. J. Carman, et al. (1996). "Mapping striate and extrastriate visual areas in human cerebral cortex." *Neurobiology*. **93**: 2382-2386.

Dretske, F. (1981). *Knowledge and the flow of information*. Cambridge, MA, MIT Press.

Dretske, F. (1988). *Explaining behavior*. Cambridge, MA, MIT Press.

Eliasmith, C. (2000). How neurons mean: A neurocomputational theory of representational content. Ph.d. Thesis in Philosophy. St. Louis, Washington University.

Eliasmith, C. (2006). *Neurosemantics and categories*. In C. Lefebvre and H. Cohen (eds.). Categorisation in Cognitive Science. Amsterdam: Elsevier.

Fair, D. (1979). "Causation and the flow of energy." *Erkenntnis*. **14**: 219-50.

Felleman, D. J. and D. C. Van Essen (1991). "Distributed hierarchical processing in primate visual cortex." *Cerebral Cortex*. **1**: 1-47.

Fitzhugh, R. (1958). "A statistical analyzer for optic nerve messages." *Journal of General Physiology*. **41**: 675-92.

Fodor, J. (1975). *The language of thought*. New York, Crowell.

Fodor, J. (1981). *Representations*. Cambridge, MA, MIT Press.

Frey, B. and N. Jojic (1999). *Estimating mixture models of images and inferring spatial transformations using the EM algorithm*. IEEE Conference on Computer Vision and Pattern Recognition, Boulder, CO, IEEE Computer Society Press.

Friston, K. (2003) "Learning and inference in the brain." *Neural Networks.* 16(9): 1325 - 1352.

Gallant, J., J. Braun, et al. (1993). "Selectivity for polar, hyperbolic, and Cartesian gratings in macaque visual cortex." *Science*. **259**: 100-3.

Georgopoulos, A. P., A. B. Schwartz, R. E. Kettner. (1986). "Neuronal population coding of movement direction." *Science* **243**(1416-19).

Gross, C. G., C. E. Rocha-Miranada, et al. (1972). "Visual properties of neurons in inferotemporal cortex of the macaque." *Journal of Neurophysiology*. **35**: 96-111.

Hubel, D. and T. Wiesel (1962). "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex." *Journal of Physiology (London)*. **160**: 106-154.

Knierim, J. J. and D. C. Van Essen (1992). "Neuronal responses to static texture patterns in area V1 of the alert macaque monkey." *Journal of Neurophysiology*. **67**: 961-980.

Lettvin, J., H. Maturana, et al. (1988/1959). What the frog's eye tells the frog's brain. In W. McCulloch (Ed.), *Embodiments of mind*. Cambridge, MA, MIT Press.

Mainen, Z. F. and T. J. Sejnowski (1995). "Reliability of spike timing in neocortical neurons." *Science*. **268**: 1503-1506.

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

Millikan, R. G. (1984). *Language, thought and other biological categories*. Cambridge, MA, MIT Press.

Nagel, T. (1974). "What is it like to be a bat?" *Philosophical Review*. **83**(4): 435-50.

Newsome, W. T. and E. B. Pare (1988). "A selective impairment of motion perception following lesions of the middle temporal visual area (MT)." *Journal of Neuroscience* **8**(6): 2201-11.

Quine, W. V. O. (1960). *Word and object*. Cambridge, MA, MIT Press.

Rieke, F., D. Warland, et al. (1997). *Spikes: Exploring the neural code*. Cambridge, MA, MIT Press.

Roelfsema, P. R., V. A. F. Lamme, et al. (1998). "Object-based attention in the primary visual cortex of the macaque monkey." *Nature*. **395**: 376-81.

Soja, N., S. Carey, et al. (1991). "Ontological categories guide young children's inductions of word meaning: object terms and substance terms." *Cognition*. **38**(2): 179-211.

Spelke, E. and G. Van de Walle (1993). Perceiving and reasoning about objects: Insights from infants. In N. Eilan, B. Brewer and R. McCarthy (Eds.), *Spatial representation*. London, Blackwell**:** 132-161.

Theunissen, F. E. and J. P. Miller (1991). "Representation of sensory information in the cricket cercal sensory system. II. Information theoretic calculation of system accuracy and optimal tuning-curve widths of four primary interneurons." *Journal of Neurophysiology*. **66**(5): 1690-1703.

Usher, M. (2001). "A statistical referential theory of content: Using information theory to account for misrepresentation." Mind and Language 16: 311-334.

Zeki, S. (1980). "The representation of colours in the cerebral cortex." *Nature*. **284**(5755): 412-8.