How Neurons Mean
A Neurocomputational Theory of Representational Content

Chris Eliasmith
Assistant Professor
University of Waterloo
chris@twinearth.wustl.edu

Ph.D. Dissertation
Department of Philosophy
Philosophy-Neuroscience-Psychology Program
Washington University in St. Louis

May 2000

ACKNOWLEDGEMENTS

This dissertation is the product of a series of significant evolutions of my initial ideas. There are many people who deserve credit for ensuring that these changes were in the right direction. They include Charles H. Anderson, William Bechtel, Andy Brook, Andy Clark, Pete Mandik, Dominic Murphy, Steve Peterson, Jesse Prinz, Pepa Toribio, David Van Essen, William Wimsatt, and Chase Wrenn, all of whom read and provided insightful comments on all or part of this dissertation and its ancestors. Of these, Charles H. Anderson, William Bechtel, Andy Clark, Steve Peterson, Jesse Prinz, Pepa Toribio, David Van Essen, and William Wimsatt, were kind enough to be members of my defense committee. I would like to express special thanks to William Bechtel, Andy Clark, and Jesse Prinz who formed the core of that committee.

I would also like to single out both Charles H. Anderson, my mentor in computational neuroscience, and William Bechtel who, as my dissertation advisor, has patiently and expertly guided my progress. As well, Pete Mandik and Chase Wrenn deserve special mention for constructive discussions that focused my efforts early on. Finally, I'd like to thank those who are part of the evolution of this thesis in a different sense, and without whom I would not have started, let alone finished: Jen and Alana Eliasmith, Jerry and Janet Elias, and Steve Elias, Kim Pendrith, and Michael Elias – thanks.

Despite all of this help, I have undoubtedly introduced or missed mistakes. Those, of course, are my responsibility alone.

ABSTRACT

Questions concerning the nature of representation and what representations are about have been a staple of Western philosophy since Aristotle. Recently, these same questions have begun to concern neuroscientists, who have developed new techniques and theories for understanding how the locus of neurobiological representation, the brain, operates. My dissertation draws on philosophy and neuroscience to develop a novel theory of representational content.

I begin by identifying what I call the problem of "neurosemantics" (i.e., how neurobiological representations have meaning). This, I argue, is simply an updated version of a problem historically addressed by philosophers. I outline three kinds of contemporary theory of representational content (i.e., causal, conceptual role, and two-factor theories) and discuss difficulties with each. I suggest that discovering a single factor that provides a unified explanation of the traditionally independent aspects of meaning will provide a means of avoiding the difficulties faced by current theories. My central purpose is to articulate and defend such a factor.

Before describing the factor itself, I summarize the necessary background for evaluating a solution to the problem of neurosemantics. The resulting analysis results in thirteen questions about representation. I provide a methodological critique of the traditional approach to answering these questions and argue for an alternative approach. I discuss evidence that suggests that this alternative provides a better means of characterizing representation.

After having established the nature of the problem and a preferred methodology, I briefly describe my theory of content. I then outline a neurobiologically motivated theory of neural computation that I and others have helped Charles H. Anderson develop. I use the computational theory to define the relations relevant to understanding representational content at various levels of analysis. I then answer each of the thirteen questions about representation.

In conclusion, I defend this theory from potential philosophical criticisms. This defense includes an explication of how concepts are to be accounted for on this theory, and a consideration of the problem of misrepresentation. I also show how this theory is immune to the standard critiques facing each of causal, conceptual role, and two-factor theories of content.

TABLE OF CONTENTS

CHAPTER 1

Setting the Stage

God is in the details. – Mies van der Rohe (1886-1969)

1 In the beginning

God, some say, is in the details. Others think that is where the Devil resides. In the case of cognition, the details are details about the brain. Surprisingly, perhaps, both views are right; these details are beautiful *and* difficult. There are few, if any, philosophers and neuroscientists who would disagree on this point. Nevertheless, there are many philosophers and neuroscientists who think that these same details, no matter how beautiful or difficult, have nothing to do with a proper characterization of our mental lives. The debate between reductionists and anti-reductionists regarding mental function focuses on just this issue: mental function can be reduced to neural function, or it cannot be.

This is not, as some may suspect, a conflict between neuroscientists and philosophers. There are those philosophers who think neuroscience is the only way to properly explain mental function (see e.g. Churchland 1981), and there are those neuroscientists who think neuroscience won't ever succeed in explaining aspects of mental function (see e.g. Eccles 1974). In the middle of these two extremes lies the position I would like to assume as a working hypothesis; namely that neuroscience is at least *relevant* to helping us solve *some* problems about mental function. This, I take it, is a weak enough position so as to be as untendentious as any.

Even those, like Jerry Fodor (1975), who famously argue against there being any priority to neuroscience when it comes to understanding cognition, admit that knowing neuroscience *may* help us know things about mental function: brains, after all, "do their mental stuff" (1998, p. 89). As Fodor (1975) is at pains to point out, however, admitting that much is not admitting very much at all: while *physical* brains may do *mental* stuff, it doesn't follow that *mental* objects and relations can be reduced to *physical* objects and relations (p. 17). While this may be logically true, it is methodologically naïve to conclude from this that we should rhetorically wonder "Why, why, does everyone go on so about the brain?" (Fodor 1999). It seems rather obvious why "everyone" interested in mental function goes on so about brains: brains are the only agreed upon instances of physical systems exhibiting mental function. Methodologically speaking, if we get a good theory about how brains perform the mental functions they do, we have *at the very least* a partial theory of how physical things give rise to mental things (or realize mental relations). Such a partial theory would be a great improvement over what is currently on offer, even if it is only partial. And, of course, there is always the prospect that such a theory can be generalized to cover more than brains: we *can't* rule out this possibility without having seen such a theory to start with. These, I take it, are good reasons for thinking that knowing neuroscience will help unravel some of the mysteries of our mental lives.

In particular, I believe that one of the problems neuroscience can help solve is that of characterizing the relationship between mental representations and the world they represent. This is not a unanimous belief. Drestke (1995), for example, claims: "A working premise behind the Representational Thesis is that a better understanding of *the mind* is not to be obtained by knowledge – no matter how detailed and precise – of the biological machinery by means of which the mind does its job" (p. xiv). In other words, thinking of mentality as being centrally representational is logically independent of what we know about brains (Fodor would agree). Notice that Dretske doesn't deny the *utility* of neuroscience, he denies that neuroscientific data alone can suffice to explain *all possible* mental representation. Put this way, neuroscientists should *not* have problem with Drestke's claim for two reasons. First, neuroscientists don't *just* generate data, they also generate theories which explain the data. Second, neuroscientific theories may be theories about mental representation without theories of mental representation being neuroscientific theories. Thus, if a theory generated by a neuroscientist speaks to the problem of mental representation (as it can), it can't be dismissed just because a neuroscientist generates it – that would be pure *ad hominem.* We can conclude from this that philosophers and neuroscientists can *both* generate theories about representation and, whatever the theory, it must respect neuroscientific data. In this respect at least, philosophers and neuroscientists might share some research territory.

In fact, philosophers and neuroscientists *do* both generate theories about representation in neurobiological systems (Millikan 1984; Dretske 1988; Miller, Jacobs et al. 1991; Abbott, Rolls et al. 1996; Rieke, Warland et al.

1997; Fodor 1998). However, there have been few attempts to provide theories spanning these disciplines. Indeed, neuroscientists and philosophers of neuroscience have both commented on the surprising lack of attempts to provide such a theory in any detail (see, e.g., Churchland 1993; Crick and Koch 1998, p. 103). Neuroscientists tend to be preoccupied with anatomical and physiological connections between stimuli and their neural effects (see, for example, any recent issue of *The Journal of Neuroscience*), while philosophers concentrate on questions concerning content or meaning of mental states (see, for example Cummins (1989), or any recent issue of *The Journal of Philosophy*). Conversely, neuroscientists seem to be uninterested in the metaphysical status of representations and the representation relation, while philosophers tend to assume that they can "leave the details to the neurophysiologist" (Dennett 1969, p. 42; Dretske 1988, esp. ch. 3; see also Fodor 1998, pp. 7, 73).

Given these divergent interests, we might think that neuroscientists and philosophers are both interested in representation, but they are interested in quite different questions concerning representation. Perhaps the central questions of these two disciplines are independent. Philosophers have, indeed, argued that neuroscience isn't relevant to theories of representational content (Fodor 1975; Dretske 1995).[1] This seems false in two ways. First, if what is meant by 'neuroscience' is just 'neuroscientific data, *simpliciter*', then neuroscience may not determine theories but this data is certainly *relevant*; if your theory predicts lots of things that conflicts with lots of accepted data, so much the worse for your theory. Second, it seems even less reasonable if what is (more reasonably) meant by 'neuroscience' is 'theoretical and empirical work in neuroscience'. Arguing that theoretical work in neuroscience isn't relevant to theories of representation seems a bad idea simply because neuroscientists clearly have theories which quantify over representations. Neuroscientists characterize lots of representation relations (Gross, Rocha-Miranada et al. 1972; Bialek, Rieke et al. 1991; Warland, Landolfa et al. 1992; Rieke, Warland et al. 1997). They are more than comfortable claiming things of the form '*X* represents *Y*'. This, of course, doesn't mean that neuroscientists have the *right* theory of representation, but they do seem to be assuming some account of what the representation relation is and what representations are. If philosophers are out give a theory of what the representation relation is and what representations are, work neuroscience is relevant because those philosophers will have to either say how neuroscientists are misusing the term, or how they are correctly using the term. Either way, philosophers should not assume that neuroscience is irrelevant to their theories of representation.

The converse is also true: neuroscientists should not assume that philosophy is irrelevant to their theories of representation. As I have noted above, quantifying over representations and engaging in representational talk doesn't entail a correct theory of representation. Theories of representation, be they philosophical or neuroscientific, should address a number of concerns – most commonly voiced by philosophers. For one, any theory of representation must be able to explain misrepresentation: How is it that some neural state, for example, can be *about* some state of affairs (e.g., a dog) when it is caused by a different state of affairs (e.g., a cat)? For another, a simple causal theory of representation (assumed by much of neuroscience) won't do. We can't, in other words, say '*X* represents *Y*' just in case '*X* causes *Y*'. The problem is simply that causal relations are far more common than representational relations. Not only, for example, does the dog cause my 'dog' representation, but so do intervening states such as the photons hitting my retina, and so do preceding states such as the dog's ancestors. Where and how, philosophers ask, should we draw the line?

What do these considerations show? It seems to me that they show that the time has come for philosophers to take neuroscientific details seriously and for neuroscientists to address the kinds of questions philosophers pose about representation. Of course, this conclusion is sound only if neuroscience and philosophy *really do* have a problem to share. The best way to show unequivocally that this is so, is to figure out what that problem is. In the next section I consider the history of inquiry into the nature of mental representation, with an eye to discovering what might concern both neuroscientists and philosophers. I then argue that these disciplines share a common interest in what I call the 'problem of neurosemantics' – the problem of how neurons mean. I then argue that, despite philosophy's traditional reliance on language as the route to understanding mental meaning, it makes more sense to approach the problem of meaning from a neuroscientific perspective. In the final section of this introductory chapter, I outline how the remainder of this dissertation is structured. As will become apparent, my goal is to provide a solution to the problem of neurosemantics that is of interest to both neuroscientists and philosophers.

---

[1] Conversely, it's hard to find neuroscientists dismissing the importance of philosophy. However, neuroscientists also don't dismiss the importance of economics to neuroscience, so it is largely unclear if they think philosophy is *at all* relevant to their work (thanks to Brian Keeley for pointing this out).

2 A brief history of mind

For thousands of years we have been trying to understand how our perceptual experiences relate to the world that causes them. In this section, I examine a small subset of these attempts in order to show that contemporary neuroscientific and philosophical inquiries into mental representation are *both* concerned with similar problems. I will show, in other words, that neuroscience and philosophy share a common ancestry when it comes to representational problems. If this is indeed the case, perhaps it will be less surprising that theories of mental content, like the one I propose in chapters 5-8, can adopt insights from both disciplines. The exemplar theories I have chosen span the approaches taken to understanding mentality in the Western tradition and include theories committed to dualism, materialism, empiricism and rationalism.

Over a thousand years ago Stoicism, a philosophical school founded by Zeno (334-262 B.C.E.), developed a unique, materialistic theory of content. The Stoics held that mental representations – what they called 'impressions' – were of at least two kinds, sensory and non-sensory:

> Sensory impressions are ones obtained through one or more sense-organs, non-sensory are ones obtained through thought such as those of the incorporeals and of the other things acquired by reason (Diogenes Laertius 7.49-51).[2]

The roots of sensory impressions are in objects in the world that the Stoics label "impressors" (Aetius 4.12.1-5). Cicero, in his *Academia*, discusses how these sensory impressions inform non-sensory impressions that are then employed by the mind to build up complex representations, and eventually concepts, or "conceptions" (2.21). Impressors, as a class, are distinguished from "figments," which cause "imaginations" and occur in "people who are melancholic and mad" (Aetius, 4.12.5). A difficulty arises in distinguishing imaginations from conceptions because both have no impressor; e.g., there is no generic dog. The Stoics solve this problem by claiming that conceptions, presumably unlike imaginations, are either "naturally" and "undesignedly" or "through instruction and attention" (Aetius 4.11.1-4) constructed from sensory impressions that are "arranged by their likenesses" (Cicero, 2.30-1). This link to sense perceptions allows conceptions to be properly classified as non-sensory impressions.

However, this solution raises a further question: How are those sense impressions related to sensory impressors? The Stoics considered this question explicitly. Diogenes Laertius, for example, suggests that "confrontation" is the link between impressors and sensory impressions (7.53). Cicero speaks of impressions being "activated" by impressors (2.30-1). Both solutions seem plainly causal: we have impressions of impressors because they cause those impressions in us.

Having identified this relation, Laertius goes on to claim that there are many other kinds of links between impressions (sensory and non-sensory) themselves, including "similarity", "analogy", "magnification", "diminution", "transposition", "combination", "opposition", "transition", and "privation" (7.53). So, for example, similarity of impressions can result in our tokening one when we token the other "like Socrates on the basis of a picture" (7.53). For each kind of link, there is a different sort of *rule* relating impressions. I will call such relations between representations *transformations*. Transformations, then, are manipulations of representations in accordance with some rule.[3] The Stoics took such transformations to be an important part of the explanation of our cognitive abilities.

Whatever we may think of the Stoics' classificatory framework or their characterization of possible transformations, it *is* of interest what they take their main problems to be. There are three main concerns for the Stoics. First, they are concerned with getting the right classification. That is, they are attempting to identify different kinds of *mental objects*. Second, they felt a need to posit the *relation* between those objects and the world. For the Stoics, this link was a *causal* one. These two concerns come together when the Stoics claim, for instance, that sensory impressions are directly caused by objects, whereas conceptions are more distantly related to the sensory impressions that give rise to them. The Stoics, then, are interested in understanding the objects of thought and their relation to the world; i.e., mental representations and the representation relation. Third, the Stoics are concerned about characterizing the *relations between mental objects*. They want to account for how some impressions can give rise to others. How, they wonder, do we get from mere "confrontation" to conceptions? In

---

[2] All quotes are taken from (Long and Sedley 1987, pp. 236-253).

[3] It is unimportant, for my purposes, whether we consider transformations as relations between two representations, or as processes of manipulating a single representation.

other words, they are interested in understanding the kinds of *transformations* that impressions can undergo. In summary, then, the Stoics wondered 1) what the mind works on, 2) how that 'mental material' is given to us, and 3) how the mind does its work on that material.

A thousand years later, empiricists and rationalists also wondered about human cognition. Descartes, a rationalist, wanted to show that reason is less fallible than the senses. The framework he relies on in arguing for this conclusion divides our mental life in to three separate, but related, "grades of perception":

> In order rightly to see what amount of certainty belongs to sense we must distinguish three grades as falling within it. To the first belongs the immediate affection of the bodily organ by external objects… The second comprises the immediate mental result, due to the mind's union with the corporeal organ affected… Finally the third contains all those judgments which, on the occasion of motions occurring in the corporeal organ, we have from our earliest years been accustomed to pass about things external to us (Descartes 1641/1955, p. 251).

The first grade, which Descartes calls "cerebral motion", is the "passive" physiological transduction of sensory stimuli. The second grade of perception arises in the mind because it is "intimately conjoined with the brain" (ibid., p. 252). This grade of perception results from the mixture of the physical and mental. It is here, in the second grade, that *mental* representations or "ideas" arise for Descartes (ibid., p. 52). The third and last grade, called "judgment" by Descartes, serves to interpret the possibly misleading picture of the world presented via the two previous grades. In cases of perceptual illusion (e.g., a straight stick that looks bent when placed into water), judgment can sometimes rectify the misleading representation presented by the first two grades. Judgment, for Descartes, serves to map our perceptions onto true or false propositions. It is these propositions, present in our "understanding", that Descartes is most interested in. Nevertheless, as someone trying to understand the mind, he feels compelled to give a story of how judgments are related to the senses.

Notably, Descartes' three-part distinction was adopted by many subsequent perceptual theorists including Malbranche, Berkeley, and Reid (Atherton in press). More importantly, despite providing a somewhat different picture of cognition than that adopted by the Stoics, Descartes has similar concerns. Descartes wants to say how we get to our final, true/false judgments. To repeat, his story is that we are physiologically "affected" by objects resulting in "sensations", these then cause an "immediate mental result" ("perception"), via the pineal gland (Descartes 1641/1955, pp. 345-6), and finally we use such perceptions to form judgments about the world. Specifically, Descartes posits internal physiological representations, or "images," in the first grade of perception (Descartes 1641/1955, p. 52), and properly so-called *mental* representations, or "ideas," in the second. Descartes also discusses how the mental representations are transformed into true or false judgments. A judgment, it seems, is some kind of complex transformation that maps representations of perceived properties onto representations of actual properties. Descartes discusses the example of seeing the sun as a small yellow disk about the size of our thumbnail, yet judging the sun to *be* a large sphere many times bigger than earth (Descartes 1641/1955, p. 161). Therefore, though Descartes' story is significantly different than the Stoics, like them he posits *mental representations*, a *relation* between those representations and the world, and *transformations* of those representations to explain our mental life.

Though on the other side of the rationalist/empiricist debate, John Locke (1700/1975) similarly characterizes the problems he is interested in. He distinguishes between what he calls "simple" and "complex" ideas, and claims that the simple ideas are joined, by various means, to form the complex ones:

> For having by *Sensation* and *Reflection* stored our Minds with simple ideas…all our complex *Ideas* are ultimately resolvable into simple *Ideas*, of which they are compounded, and originally made up, though perhaps their immediate Ingredients, as I may so say, are also complex *Ideas* (II, 22, 9).

Though Locke is often criticized for his overly liberal use of the term 'ideas', which results in him conflating representations with their contents (see, e.g., Yolton 1993, p. 91-2), he clearly has some notion of mental entities that are causally derived from sensory receptors and that help explain mentation. Locke, himself, occasionally refers to ideas as *representations of* things (II, 30, 5; II, 31, 6). In particular, he thinks of simple ideas as "sensible representations" of the external world (IV, 3, 19) that are compounded to form all other ideas.

Locke is greatly interested in the various means by which the ideas may be compounded. He suggests a number different transformations (or "first Faculties and Operations of the Mind") which ideas might undergo, including (much like the Stoics) "Composition," "Enlarging," "Abstraction," and "Comparing" (II, 11, 4-14). In

particular, Locke, like Descartes, places much emphasis on the role of judgment, which, he suggests, "alters the appearances into their causes" (II, 9, 8). As an example of the role of judgment, Locke discusses "perceiving" a small golden globe, even though only (the idea of) a flat, shadowed circle is "imprinted" (II, 9, 8). Judgment performs the important task of determining the actual properties of things from the properties we "receive". So, the basic features of Locke's theory of mind are much like those of Descartes, and thus much like those of the Stoics: *mental representations*, causally *related* to external objects, are *transformed* by various means.

Having survived over a thousand years, it is not surprising that this picture has survived three hundred more, to the present day. Consider, for instance, the theory of mental representation espoused by Fodor (1975; 1987; 1994; 1998). Though Fodor's theory of content has changed, the problem he is addressing remains essentially the same:

> [This], I suppose, *is* the problem of perception … For though the information provided by causal interactions between the environment and the organism is information about physical properties in the *first* instance, in the *last* instance it may (of course) be information about any property the organism can perceive the environment to have (Fodor 1975, p. 47).

Fodor has unequivocally and consistently held a "representational theory of mind" (1998, p. 1). He is quite explicit that this theory posits mental representations and that our mental life stems from computations over those representations (ibid., pp. 7-9). Computations, of course, are a computer-age versions of transformations; mental computations are mental processes which modify (e.g., compound, associate, etc. (ibid., pp. 9-12)) mental representations.

One important difference between Fodor's inquiry and the historical inquiries I've considered so far is that Fodor, like most of his contemporaries, is more concerned about the representation *relation* than about representations or transformations. The Stoics, Descartes, and Locke all assumed that the causal relation just *automatically* determined what mental representations were about; mental representations are about the objects that cause them. Fodor and his contemporaries have realized that simple causation won't properly explain what representations are about (see e.g. Dretske 1988, p. 74; Fodor 1998, p. 73). If I am given a picture of a dog, for example, it is the picture that causes my mental representation, but it is the dog that my representation is about; representational content and cause can come apart.

Fodor thinks that content is determined by "nomic relations" (ibid., p. 73). So, for example, he claims that "'dog' [the word] and DOG [the concept] mean *dog* because 'dog' expresses DOG, and DOG tokens fall under a law according to which they reliably are (or would be) among the effects of instantiated *doghood*" (ibid., p. 75). Fodor, then, posits some other kind of *metaphysical* regularity to underwrite the meaning of mental representations. So, the picture of a dog may cause me to token my 'dog' representation, but that representation has a nomic relation with *dogs*, not *pictures of dogs*. Therefore, that representation is about dogs, and not dog pictures.

I have been tracing the history of these problems in order to show that neuroscience and philosophy share their genealogy. But so far I have said little about neuroscience. Notice, however, that the explanation provided by each of these schools of thought is progressively more 'mechanism conscious'; i.e., the way sensation happens is becoming more important. The Stoics compared perception to the imprinting of ring seals into wax (Laertius, 7.49-51), and provided no explanation of how transformations occur. Descartes (and even more so Malbranche (Atherton in press)) had a reasonably sophisticated physiological explanation (discussed in his *Dioptrics*), of how external motions were transduced into internal motions that were images of sensory stimuli. Despite the very mechanistic view of the first two grades of perception, Descartes did not think it possible to give a mechanistic account of the third grade. Locke, writing only a few years after Descartes, has a similar kind of story to tell. Fodor, though no friend of neuroscience (see e.g., Fodor 1995), has a very mechanistic view of mentation throughout. For Fodor, representations are transduced by some purely physical process. Furthermore, he supposes thinking to *be* computation (where by 'computation' Fodor means Turing-like discrete symbol manipulation (1998, pp. 10-11)). Fodor has chosen, then, to adopt the mechanisms important for cognition from the field of computational theory (coupled with psychology). Others choose to find the relevant mechanisms in neuroscience (Churchland 1989; Churchland and Sejnowski 1992; Akins 1996; Rieke, Warland et al. 1997).

There are good reasons to look to neuroscience rather than to traditional computational theory. Most importantly, computational theory places no constraints on what functions can be computed, and what the computations are defined over. This leaves the problem of what and how things can be represented completely undetermined. Neuroscience, on the other hand, is in the business of figuring out what constraints there are on

neural processing: How many, and which, neurons represent what parts of the environment? What intensity of external stimuli is needed to drive neurons? What can neurons do (and what can't they do)? Neuroscientists, then, take single neurons and groups of neurons to *be* representations (Gross, Rocha-Miranada et al. 1972; Felleman and Van Essen 1991; Abbott, Rolls et al. 1996). Neuroscientists characterize the *relation* between these representations and the outside world (Rieke, Warland et al. 1997). And, neuroscientists are interested in knowing what kinds of *transformations* the brain can perform on these representations (Andersen and Zipser 1988; Zipser and Andersen 1988). In each case, neuroscientists are asking the same sorts of questions as Fodor, Locke, Descartes, and the Stoics and, in each case, they are turning to the brain for constraints on the kinds of answers they can provide. Unlike traditional computational theory, then, the solutions proposed by neuroscientists *are* constrained. More importantly, these solutions are constrained by a system that is *known* to have the property we are trying to explain. This doesn't mean that computational systems can't have meaning. It does mean, however, that rather than assume something about the brain based on our computational theories, it makes more sense to try and understand computation *as it relates to the brain* (Churchland and Sejnowski 1992; Bower 1998).

Psychology can help Fodor here. In particular, psychologists discover constraints on mental processing by looking at real, thinking systems. These are constraints that can be incorporated into theories of mental processing without turning directly to the brain. But, we should notice something important about the current state of psychology: one of the most rapidly expanding sub-fields is cognitive *neuroscience*. Even psychologists, Fodor's one-time allies, are turning to the brain to get a better handle on the right mechanisms. The mental vocabulary of these psychologists, much to Fodor's chagrin, is quickly becoming permeated with terms from neuroscience: mental objects and processes are becoming understood in terms of (and type-identified with) physical objects and processes. These psychologists, then, have accepted the conclusions of the methodological argument I presented in the first section: we *should* look at the brain to understand mental processes. To summarize: even though we may agree with Fodor that computation, *in some sense*, is important for understanding mental processes, we need to turn to the brain to understand what that sense is.

Historically speaking, it is perfectly natural to combine our best details about mechanisms in the brain, with our theories about mental representation. Neuroscience, currently, provides those details. I'm suggesting that we do the scientifically respectable thing and bring the evidence to bear on our theories, even on our metaphysical theories (see, e.g., (Quine 1960) for arguments to the effect that this is unavoidable). Fodor is, again, right that all the details in the world about the brain won't *determine* what content is, but whatever content is, it better be *consistent* with those details. Perhaps, then, one good way to generate a theory of representational content that is consistent with these details is to keep the details in mind while generating the theory. This is precisely what I propose to do.

3 The problem of neurosemantics

I propose, then, to tackle the same problem that the Stoics, Descartes, Locke, and Fodor are interested in, *while* paying heed to what we know about the brain. Given recent developments in neuroscience, including brain scanning techniques such as PET and fMRI, long term inter-cellular recordings, and new theories on neural coding, there is good reason to believe that we are currently in a unique position to address this age-old problem. But what, precisely, *is* this problem?

Recall that each approach I examined is concerned with three things: mental representations, the representation relation, and transformations of these representations. Thus, they share a related family of questions. Here are formulations of what I take to be central questions for each approach, formulated in their own terminology:

> *Stoics*: What is the nature of our various kinds of impressions? How are they related to the world? And, how are the impressions related to each other?
>
> *Descartes*: What are sensory images and what are mental perceptions and how, precisely, are they formed? What are the relations between sensed objects, sensory images, perception and judgment? And, how does our judgment act on perceptions?
>
> *Locke*: How are simple ideas different from complex ideas? What is the relation between our ideas and what they are ideas of? And, how do our simple ideas combine to form complex ideas?

Fodor: What properties do mental representations have? What, if anything, underwrites the nomological relations between these representations and their contents? And, what class of computations operate on these representations?

Neuroscientists: How do single neurons or groups of neurons represent? What is the relation between the neural representation and the external environment? And, what kinds of computation can biological systems perform on these representations?

Some of these questions are never explicitly posed by a given school, but all either assume, assert or argue for an answer, and all are appropriate questions to ask. As well, there is a common thread linking these problems together. The common thread is the subject of these questions – *us*. Human beings are, as far as each of these positions is concerned, an undeniable subject to which their concepts apply. Furthermore, we, assuming materialism is correct, are neurobiological systems.[4] Therefore, one problem these positions definitely share is the problem of how we, qua neurobiological system, have representational content. As Dennett (1969) has put it: "What, if anything, permits us to endow neural states with content?" (p. 44). This, I take it, is the 'problem of neurosemantics.'

A few comments are in order. First, I presume that a number of other questions will *have* to be addressed in order for this problem to be satisfactorily solved. For example, what are neurobiological representations? How are the representations related to their contents? How are contents related to one another (i.e., what transformations can be realized by neural computations)? In other words, I take it that the problem of neurosemantics can only be solved if each of the three *kinds* of historical questions is answered. I dedicate part of chapter 3 to deriving and posing a more complete set of questions.

Second, I don't intend to specify the particular transformations that obtain between our representations. A completed neuroscience is needed to do that. What I do intend to provide is a framework that will help us identify, describe, and explain the *kinds* of transformations neurobiological representations can enter into. Recent theoretical advances in the neurosciences help provide (as I show in chapter 6) just this kind of framework. These advances suggest a theory of content ascription that I outline and defend in chapters 5, 7 and 8.

Third, I would like to repeat a point I made in the first section of this chapter: although the problem of *neuro*semantics is *possibly* more limited than a problem about *psycho*semantics (to use Fodor's term), this isn't necessarily the case. There are three well-known possibilities of the relation between these two problems: 1) the former may reduce the latter (Place 1959); 2) the former may eliminate the latter (Churchland 1981); or 3) the former can only ever provide a solution to a very limited and uninteresting subset of the latter (Fodor 1975). Only the last possibility entails that neurosemantics is a more limited problem. Furthermore, it is possible that a theory of neurosemantics could be generalized to satisfy those who hold the third position. I want to remain officially agnostic as to which of these relations holds between neural and psychological theories. However, it should be clear by now that my hunch is that the third position can't be right.

4 Mental content

In the last twenty or so years, there have been a plethora of philosophical theories trying to answer questions about mental content (see chapter 2). They have run the gambit from covariance theories (Dretske 1981; Fodor 1981; Dretske 1988; Fodor 1998), to functional role (Harman 1982; Block 1986; Harman 1987), to adaptational role (Millikan 1984; Dretske 1988; Dretske 1995). These theories aren't particularly concerned with neurons, but rather with mental states, mental representations, and concepts. But, because these theories assume materialism to be true, there is a very clear sense in which they *are* assigning content to neurons. In the remainder of this section I provide a brief characterization of what content *is*, such that it is assignable to neurons (as well as to concepts).

Generally speaking, contemporary theories of *mental* content are part of a tradition concerned with *linguistic* content. In this tradition, the content or meaning of a sentence is the abstract proposition that the sentences expresses. Thus, the sentence 'The star is bright' expresses the same content as 'Der Stern ist hell' and 'L'étoile est lumineuse' even though the sentence types are different. For mental states, it would be the

---

[4] Notably, Locke and Descartes did not hold materialism to be true. However, this is essentially irrelevant to what they take the important *problems* to be; the metaphysics changes, but the questions don't. Of course, what counts as a good solution *will* be quite different.

thought 'The star is bright' that has this same content. Content, then, is what a representation tells you about what it represents. An equivalent way of saying this is that content is the set of properties ascribed to something by its representation. Given this definition, *two* aspects of content become evident: there is the set of properties and the thing they are ascribed to. These two aspects have gone by the names of 'sense' and 'reference', 'intension' and 'extension', and 'meaning' and 'denotation'. Whatever the choice of terms, there seem to be two different problems that need to be solved. The first is the problem of *fixing* content, i.e., figuring out what the representation refers to. The second is the problem of *determining* content, i.e., figuring out what properties are assigned to the object of the representation.

To illustrate the difference between these two aspects of content, consider a variation of Frege's (1892/1980) now famous 'evening star' example. If I tell you that 'The evening star is Venus', then there is a possibility that I am telling you something new. You may, in fact, quickly deduce that, since the morning star is Venus, the morning star is the same as the evening star. The two aspects of content come apart in this example as follows: even though 'the morning star' and 'the evening star' are *about* the same thing, namely Venus, it may not be the case that the *properties ascribed* by someone's representations 'the morning star' and 'the evening star' are the same. So, even though the content may be *fixed* to the same thing, the content may be *determined* to be different. My seeing the morning star and my seeing the evening star may both cause my content to be fixed to Venus, but I may ascribe different properties in each case (e.g., that one appears in the morning and the other in the evening).

One way my talk of property ascription differs from standard accounts is that it is often thought that properties alone aren't enough to understand content determination. So, for example, even if all the properties I ascribed to the morning star and the evening star are the same, the claim would be that they *still* have different '*senses*'. This complaint can't remain, however, once we realize that one of the properties ascribed by a representation to its object *in virtue of its representing that object* is that *that particular* representation ascribes those properties. As tautologous as that may sound, it shows a minimal sense in which no two syntactically different terms *could* have identical senses; this is precisely the result that is desired by those lodging the complaint to begin with. What they have failed to realize, it seems, is that claiming that *all* properties ascribed by two representations are the same *entails* that the property of being represented by that particular representation is the same (i.e., the syntax is the same). Thus, it is impossible for the all of the properties to be the same and the senses to be different. Therefore, it is perfectly acceptable to identify senses with the properties ascribed by a representation.

So, how can *neurons* ascribe properties? Consider a system of peripheral sensory neurons like those found in the eye. These cells transduce light intensities and transmit a series of discrete, rapid, nearly identical voltage discharges (called 'neural spikes') down the optic nerve. The pattern of spikes transmitted varies as a result of changes in light intensity. If content is the set of properties ascribed to something by its representation, then the retinal ganglia neurons have content. In particular, they ascribe the property of there being a certain temporal and spatial density of photons at a certain retinal location. But, the real question is, can neurons have *interesting* content, i.e., content rich enough to underwrite our mental abilities? Can we figure out how neurons might support not only a humdrum content at the sensory periphery but also the 'full-blooded' content of morning stars? I think the answer to both questions is 'yes' (see chapters 5-8). Notably, the term 'content' is often reserved solely for language, or language-like mental structures. However, I will use the term 'content' more broadly, and take it as part of my project to show how a notion of 'content' can apply both to what is found in single neurons and to what is found in language-like mental structures; I am interested in understanding what *unifies* content, so understood.

5 Language and meaning

Frege's distinction between 'sense' and 'reference' stems from his work on language. For many philosophers, it is natural to extend insights about language to the mental realm. The reasons are various. For example, it is often argued that we think in a 'language of thought' that has all the structural properties of natural language (see e.g. Fodor 1975). If this is true, any insights we gain about natural language apply equally to our mental language. As well, some have argued that the purpose of language is to express our thoughts (see e.g. Chisholm 1955). In this case, studying the product of thought may give us insight into the processes that produce it. Nevertheless, I think there are better reasons *not* to rely heavily on insights about language for understanding thought. In this section I show, contrary to the traditional approach in philosophy, why language is *secondary* to understanding mental

content. Although linguistic abilities must be accounted for by a theory of mental content, there are reasons to think we should avoid taking language as a starting point.

Many philosophers who have proposed semantic theories have focused on the *propositional* content of beliefs and language (see e.g. Loar 1981; Evans 1982; Harman 1982; Lycan 1984; Block 1986; Fodor 1998). This project has been less than obviously successful. As Lycan (1984), a proponent of the approach, has put the point:

> Linguistics is so hard. Even after thirty years of exhausting work by scores of brilliant theorists, virtually no actual syntactic or semantic result has been established by the professional community as *known* (p. 259).

But Lycan, like most, is determined to continue with the project using the same methods, and shunning others: "And there must be some description of this processing that yields the right predictions without descending all the way to the neuron-by-neuron level" (ibid., p. 259). After thirty (forty-five by now) years of difficulty, it seems rather likely that those neuron-by-neuron details actually *do* matter to a good characterization of the syntax and semantics of mental representations (and perhaps, through them, language).

There are reasons other than a simple lack of success to think that language may not be a good starting point for such theories. For one, many theorists agree that mental content should be naturalized. That is, content deserves a scientific explanation that refers to objects found in nature. Linguistic objects, like words, are presumably one kind of object found in nature. But, it is a mistake to give an explanation of content *in terms of* words since this is to explain one poorly understood natural concept in terms of another. In fact, such an explanation would be perfectly circular if we were giving an explanation of content that relied on the content-carrying capacity of words.

Worse yet, language is only one small domain of the application of natural content. Language, as most linguists understand it, is a human specialization. Thus it is unique to one species in millions. This is a good reason to think that starting with language, or focusing on language, when constructing a theory of content is a dangerous tactic. This is true unless we have *prima facie* evidence that most non-human animals don't have internal representations; but we don't have such evidence.[5] Furthermore, there *is* clear evidence that language is *not necessary* for content.[6] People who have had the misfortune of growing up without natural language, but later learn language, are able to recall events that preceded their linguistic competence (Nova 1997). So, we need a theory that can account for content in the *absence* of natural language. Furthermore, the use of symbols for communication in the animal kingdom is rampant. Bee dances, monkey calls, whale songs, bird songs, etc. are all instances of communicating properties of the environment via symbols that refer to the things having those properties. So, it seems likely that it is much *more* common for there to be content without language than content with language. Content, it seems, is prior to language.

In addition, if we think that linguistic capacities are the result of a somewhat continuous evolutionary process, then the fact that language is a human specialization suggests that it is a far more complex phenomenon than "merely" having neural states with content. Even those, like Chomsky (1986), who think that language is a specifically human ability that *doesn't* have evolutionary precursors, argue that language is particularly complex. Being able to deal with linguistic complexity suggests uniquely powerful computational abilities. Thus humans, by all indications, have the most computationally powerful brain of any animal. To begin explorations of content by examining a phenomenon found solely in the most complex exemplar systems with content just seems a bad tactic (Bechtel and Richardson 1993). This, in fact, might serve to *explain* the lack of progress noted by Lycan. If language were taken to be an endpoint in a continuum of content complexity, then the fact that our theories of language are not compelling, as Lycan suggests, would be expected rather than surprising.

These are all reasons to think that constructing a theory of content beginning with language should be exceedingly difficult, as has proven to be the case. And, even if a successful language-based theory of content is constructed, these are then reasons to be unsure about how such a theory will apply to non-linguistic cases – which are the majority. Starting at the "neuron-by-neuron" level avoids such difficulties. We have quite good

---

[5] Notably, positing an 'internal language' for animals raises the problem of why there is no behavioral evidence that animals have a representational system approaching the complexity of human *language*, proper.

[6] The debate as to whether language is sufficient for content is one that focuses on the abilities of machines. Some, most famously Searle (1992), claim that a computer could master a natural language yet not have content. Many others disagree (Turing 1950; Hofstadter and Dennett 1981; Thagard 1986; Churchland and Churchland 1990).

naturalistic descriptions of neurons (especially compared to words). We don't yet have any clear examples of 'interesting' content in non-neural systems. And, the complexity of neuron (not neural) function is far less than that of the brain areas involved in language production. So, starting at the neuron-by-neuron level should not be so quickly shunned. Rather, it may be more naturalistic, more widely applicable, and more likely to succeed.

A terminological consequence of my rejecting linguo-centrism about mental representations is that I will use the terms "meaning" and "content" interchangeably, as others have done (Dretske 1988, p. 52; Cummins 1989, p. 12). I note this because there are those who distinguish content from meaning where the former is mental and the latter linguistic (Loar 1981, p. 1; Peacocke 1986, p. 3).

6 The plan of attack

My primary goal is to work out a tenable solution to the problem of neurosemantics. A more modest secondary goal is to show how I think such a solution *should* be constructed. My aims are two-fold then; partly theoretical and partly methodological. Before attempting to realize either goal directly, in the next chapter I critically survey current contemporary theories of content. As a result of this analysis, I propose a strategy for avoiding the difficulties contemporary theories have. In addition, this survey presents the philosophical background for the ensuing discussion.

In the subsequent chapter (chapter 3), I begin to address the methodological project. In particular, I analyze the problem of neurosemantics in terms of the representation relation. This analysis highlights thirteen questions that must be answered in order to successfully solve the problem. I again argue there, as I have here, that these are questions of interest to both philosophers and neuroscientists. Continuing with the methodological project in chapter 4, I discuss a non-traditional approach to understanding the representation relation. I show that the traditional approach, assumed by philosophers and neuroscientists alike, can result in unduly complex characterizations of the representation relation. Drawing from recent theoretical and experimental work in computational neuroscience, I discuss an alternative approach – what I call 'taking the animal's perspective' – that improves such characterizations. Given the successes of this new approach, I adopt it for the remainder of the thesis, and show how it can inform a theory of content (via what I call the 'statistical dependence hypothesis').

In the fifth chapter, I begin the theoretical part of my project. The first half of this chapter is dedicated to explicating the basic assumptions of my theory of content, with particular emphasis on defending an appropriate theory of cause. In the second half of chapter 5, I briefly outline the theory of content that I articulate and defend in chapters 6 through 8. This outline is intended to motivate the ensuing discussion of a neurocomputational theory that, I believe, is an integral part of a solution to the problem of neurosemantics.

In chapter 6 I provide enough detail of the neurocomputational theory to show how it underwrites a solution to the problem of neurosemantics. In particular, I describe a means of characterizing the relation between representational levels that defines the transformations supported at those levels *and* the causal relation to the external world. The computational theory I present is based on the work of a number of computational neuroscientists including Rieke and Bialek (1997), Miller (1991), Georgopoulos (1986), and more directly that of Anderson (1994; 1998; Eliasmith and Anderson 1999; Eliasmith and Anderson forthcoming; Eliasmith and Anderson in press).

With these details in hand, I revisit the theory outlined at the end of chapter five in order to provide a fuller account. This, then, is where the neuroscientific details and philosophical considerations meet. I discuss how, on the basis of the neurocomputational theory I adopt, the representation relation can be understood, and I answer each of the questions about representation posed in chapter 3. In addition, I provide a detailed example of the application of the theory.

In the final chapter, I defend this account against the philosophical concerns that have posed difficulties for previous theories of representational content as I discuss in chapter 2. Most importantly, this defense includes an explication of how concepts are to be accounted for on this theory, and a consideration of the problem of misrepresentation. I suggest that the plausibility of this account is due to its reliance on both philosophical and neuroscientific results concerning the nature of representation.

Churchland and Sejnowski (1992), a philosopher and a neuroscientist respectively, note that "'Data rich, but theory poor' is a description frequently applied to neuroscience" (p. 16). I think it may be fair to say the opposite of philosophy. Perhaps, then, neuroscience and philosophy aren't nearly as strange bedfellows as many would think. Perhaps, too, neuroscience and philosophy should combine forces to understand cognition; especially since they share a common problem. I am not merely interested in claiming *that* this should happen, I'm interested in *showing* how it can.

CHAPTER 2

Contemporary Theories of Content

Do not believe in anything merely on the authority of your teachers and elders. – Buddha

1 Introduction

In this chapter I survey a variety of theories of content proposed by contemporary philosophers. I also discuss the difficulties that each of these theories faces. This brief survey divides the theories currently on offer into three categories: causal theories, conceptual role theories, and two-factor theories. I discuss each of these kinds of theories by choosing two influential proponents from each. Though by no means exhaustive, this survey covers by far the majority of positions available regarding the nature of content. To conclude this chapter, I suggest a strategy for avoiding the problems faced by contemporary theories. Although the strategy is admittedly vague, I show how adopting it can result in a precise theory of content in chapters 5-8.

2 Causal theories

Causal theories of content have as their main thesis that mental representations are about what causes them. My 'dog' thoughts mean dog because dogs cause me to token them. The theories of content proposed by Jerry Fodor (1990; 1998), and Fred Dretske (1981; 1995) are influential examples of the two most common kinds of causal theories; synchronic and diachronic. Synchronic theories, like Fodor's, do not depend on the history of the system to determine representational content. Diachronic theories, like Dretske's, do. In both cases, however, the meaning of a mental representation is determined by its causal relations to the external environment. For this reason, causal theories are also called 'externalist' theories of meaning.

The motivations for holding an externalist theory are varied. Historically speaking, both Descartes (with his 'immediate affection' relation) and the Stoics (with their 'confrontation' relation) implicitly assumed that causation was important for determining meaning. The intuition that motivated their outright *assumption* that cause determined meaning is enshrined in contemporary causal theories. However, philosophers have more recently realized that a naïve causal theory is problematic and have thus proposed various extensions to a simple causal theory (see section 2.2).

A more recent motivating factor for causal theories lies in a series of thought experiments invented by Hilary Putnam (1975) and extended by Tyler Burge (Burge 1979). These so-called 'Twin Earth' thought experiments have served to make externalism compelling to many philosophers of mind and language. A simple example is as follows: Suppose there is a molecular duplicate of earth somewhere far away, call it Twin Earth. On Twin Earth, the entire population of earth is reproduced down to every last neural connection. In fact, the only difference between Twin Earth and earth is that the substance we call water has a microstructure of XYZ rather than H_2O. Notably, all of the phenomenal properties of XYZ and H_2O are the same, only the chemical makeup is different. Now, let us consider a pair of earth/Twin Earth twins, Hilary and Twin Hilary. Notice that on earth, Hilary's 'water' thoughts refer to H_2O but on Twin Earth, Twin Hilary's 'water' thoughts refer to XYZ. In fact, if we brought a sample of XYZ to earth and Hilary called it water, we would want to say that Hilary was wrong. It isn't water, it's Twin water because it's XYZ and not H_2O. What this means, then, is that Hilary's 'water' thoughts mean something different than Twin Hilary's 'water' thoughts (namely, H_2O instead of XYZ). Since the only difference, *ex hypothesi*, between Hilary and Twin Hilary are their causal relations (Hilary is causally related to H_2O and Twin Hilary is related to XYZ), we know that cause has to determine meaning.

A third motivation for causal theories is their success at explaining communication and shared meanings. Putnam has also discussed the consequences of such intuitions in a social setting. Analogous to the Twin Earth thought experiment, Putnam constructs a thought experiment in which a person (say Hilary, again) is unable to perceptually distinguish between elm trees and beech trees. Putnam claims that Hilary would be considered to be wrong if he called elms 'beeches' or vice versa. The only way this intuition can be explained is by appeal to an externalist theory of meaning; or, more precisely, to an externalist theory that relies on a group of experts to determine the meaning of certain terms (like 'elm' and 'beech'). Under this sort of theory, communication and shared meanings can be explained because members of the same communities will have the same experts (and environments) determining the meanings of their terms. We can then explain communication by appeal to these

socially determined meanings. In particular, we successfully communicate when our usages align with those of experts (i.e., when our terms mean the same thing).

3 Problems with causal theories

The biggest problem for causal theories is explaining misrepresentation. Consider, for instance, my looking at a cat that I represent as a dog. Intuitively, we want to consider this a typical case of misrepresentation. However, a naïve causal theory wouldn't clearly show why this is misrepresentation as opposed to the correct representation of the disjunction of the set of cats with that of dogs (i.e., cats *or* dogs). Since, in other words, a cat is *causing* me to token this representation according to a causal theory, the representation is about the cat. However, a dog also causes me to token the same representation. So now this representation is causally related to the set 'cat or dog'. How can we explain representational *mistakes* under such a theory?

Clearly we can't. This is why the main focus of contemporary theories has been to better understand the nature of the representation relation. In particular, if this relation isn't just causal, what other ingredients *do* we need? In the remainder of this section, I consider two solutions to this 'problem of misrepresentation', one from Fodor and one from Dretske. I also show why each is unsatisfactory.

Fodor posits what he calls 'nomic' relations to explain representation. These are lawful causal relations that obtain between a representational state and what it is about. So, there is a nomic relation between my 'dog' representation and dogs. In order to explain misrepresentation, Fodor further posits a particular kind of relation between these relations; i.e., a second-order relation between first-order relations. Specifically, he suggests that there is an *asymmetric dependence* between misrepresenting nomic relations and correctly representing nomic relations. In the above example, there is a nomic relation between cats and my 'dog' representation. There is also a nomic relation between dogs and my 'dog' representation. Fodor holds that the cat nomic relation is dependent on the dog nomic relation. He also holds that this dependence is asymmetric because the dog nomic relation doesn't depend on the cat nomic relation. Presumably we can take 'dependence' as meaning something like 'wouldn't exist without'. The claim, then, is that misrepresentation occurs whenever we have this kind of asymmetric dependence.

The biggest difficulty with this 'theory' of misrepresentation is that it is too vague. This is true in two senses. First, as Cummins (1989) and Hutto (1999) have pointed out independently, Fodor's solution seems more like a redescription of the problem that is supposed to be solved than an actual solution. Fodor (1987) admits as much "The treatment of error I've proposed is, in a certain sense, purely formal … it looks like any theory of error will have to provide for the asymmetric dependence of false tokenings on true ones" (p. 110). The point of a *solution* is to say what *determines* those dependencies. Fodor has left it open as to whether the asymmetric dependence is determined by evolutionary facts about a representer, or the representer's learning history, or naming ceremonies, or some kind of dispositions, or, for that matter, statistical dependence relations. Hutto (1999) complains that "in absence of this vital detail [the asymmetric dependency thesis] is of no use to the naturalist" (p. 47) and that Fodor "fails to give a scientifically respectable explanation of the dependency relationship" (p. 48). Fodor does add the extra constraint that the dependence must be synchronic, but there is nothing about asymmetric dependencies in particular that supports the additional constraint. Second, Fodor provides no principled means of determining what nomic relations depend on which others. He provides examples, but not a way of knowing when such relations hold. Why, for example, would there *not* be an asymmetric dependence between stereotypical dog nomic relations and atypical dog nomic relations? Or, better yet, between atypical doorknob nomic relations and typical doorknob nomic relations (see Fodor 1998).

Dretske (1988) has a very different solution to the problem of misrepresentation. He claims that mental representations have evolutionarily determined functions.[1] The function of a 'dog' representation is to represent dogs because, over the course of evolutionary history, that kind of representational state has been used to represent dogs. Given this account, we can pick out cases of misrepresentation because only in such cases do representations *not* perform their function. Cats might cause my 'dog' representation to be tokened, but my 'dog' representation doesn't have the function of representing cats, therefore it is a case of misrepresentation.

---

[1] Dretske often relies on learning history rather than evolutionary history to fix functions, but in either case his theory is diachronic.

Again, two difficulties arise for this theory. First, much like Fodor, Dretske's theory is vague in that it doesn't give any detail as to how we can determine what the function of a particular state is. Obviously, evolutionary (and learning) histories have *some* cases of misrepresentation during the function-fixing phase. How many correct cases are needed to determine *the* function of a neural state? How are the correct and incorrect cases to be distinguished during *this* process?

A second and independent concern is that diachronic theories conflict with central intuitions about meaning. This conflict can best be highlighted by considering Donald Davidson's (1987) swampman thought experiment. In this thought experiment we are asked to suppose that someone, say Don, is standing beside a swamp. Suppose also that a large bolt of energy strikes Don, eliminating him, and independently strikes the swamp, causing a molecular doppelganger of Don to appear. We would suppose that Swamp Don, when he goes out into the world, behaves in all the same ways that Don would have behaved. We would thus also suppose that Swamp Don represents things (and misrepresents them) in just the ways that Don does. However, according to Dretske's theory, Swamp Don doesn't have representations or meanings *anything like* those had by Don. The reason Dretske thinks this is acceptable is because "such [internalist] premises are suspect when applied to fantastic situations" (Dretske 1995, p. 148). He thinks, in other words, that intuitions that meanings should be the same for both Don and Swamp Don are more fallible than his theory. However, despite Dretske's being appalled by 'fantastic' thought experiments, a similar story can be recreated in the 'scientific' language of artificial neural networks. Suppose networks are randomly generated until one is found that has all the same weights as some trained network. The first network will have the same behavior, but a very different history than the second. Our intuitions about meaning are just as strong in this second, far more realistic case; we wouldn't think that one network has meanings if the other doesn't. It seems, then, that we should be more concerned about the viability of diachronic theories than with the applicability of intuitions to fantastic situations.

4 Conceptual role theories

Conceptual role theories hold that the meaning of a term is determined by its overall role in a conceptual scheme. As I use the term, 'conceptual role theory' can denote any theory that ascribes the meaning on the basis of a causal, computational, functional, inferential, or conceptual role. Theories of this sort have been proposed by Gilbert Harman (1982) and Brian Loar (1981), and are often called 'internalist' theories of content because they depend on factors internal to an agent to determine meaning. Under such theories, the meaning of a term is determined by the inferences it causes, the inferences it is the result of, or both. So, for example, 'dog' means dog because we use it to infer properties like 'has four legs', 'is furry', 'is an animal', 'is friendly', etc. all of which are properties of dogs.

A motivation for this kind of theory may be that, historically speaking, the deepest insight of the Stoics and Descartes into the nature of meaning was that transformations matter. The Stoics and Locke were explicitly concerned with transformations such as magnification, analogy, and combination (Laertius, 7.53). Descartes, too, discusses the mappings of perceptions onto judgments (1641/1955, p. 161). What these philosophers are doing is trying to understand the relations *between* mental representations. They think, in other words, that transformations help determine meaning. Conceptual role theories take this one step further and insist that such relations are all there is to meaning.

A second, more recent, motivation driving such theories can be seen by reconsidering Frege cases (see section 4 of chapter 1). Recall that Frege (1892/1980) considers the possibility that when we are told that 'Hesperus' (the evening star) refers to the same thing as 'Phosphorus' (the morning star), we learn something new. Or, more generally, when we are told of the coreference of two terms, their meanings might change. A causal theory can't account for this intuition because we know that Hesperus and Phosphorus have the same referent (Venus) and reference is all there is to meaning under such a theory. The idea is that even when causes are the same, meanings can be different.

A conceptual role theorist can explain our intuition quite easily by noting that the inferences warranted by each term can be quite different. 'Hesperus' warrants the inference 'will be seen in the evening' whereas 'Phosphorus' warrants the inference 'will be seen in the morning'. Thus the terms differ in meaning and when we are told of their coreference we *learn* something. In particular we learn that the inferences for one are warranted for the other, so the meanings of both terms change appropriately.

A final motivation for conceptual role theories is their explanatory success. Consider Twin Earth again. In the Twin Earth case, we would expect Hilary and Twin Hilary to behave in exactly the same way. Of course,

ex hypothesi, causes are different in the two cases. It seems unlikely, then, that we can explain the *sameness* of behavior in terms of causes given this *difference* in cause. However, we can explain *sameness* of behavior given a *sameness* of conceptual role. In particular, the internal states of the twins are identical, so Hilary's 'water' representation will have all of the same inferential (and therefore behavioral) consequences as Twin Hilary's 'water' representation. If we expect what we mean to determine how we behave, we should consider conceptual role theories a success.

5 Problems with conceptual role theories

The two main problems with conceptual role theories are their inability to account for truth conditions, and their vulnerability to charges of relativism (for a review of other problems see Fodor and Lepore 1992). Truth conditions are a problem for conceptual role theories precisely because such theories deny any importance to causes in determining meaning. If we think that truth determines meaning in some way, then conceptual role theories are in trouble. For example, if we think that my pointing to H_2O and saying 'water' and my pointing to XYZ and saying 'water' are instances of my being right and wrong respectively, then we think that truth determines meaning. This follows because, in the second case, my being wrong depends on the meaning of my term referring to something else, namely H_2O. Conceptual role theories say nothing of a connection between my concepts and their referents in the world. These theories can't explain, then, why I'm right in one case and wrong in the other.[2] In this sense, conceptual role theories entail that meaning is cut off from the environment. Thus, these theories can't explain how we refer to *actual objects* rather than to sets of inferences.

The second difficulty for conceptual role theories is the problem of relativism (see Fodor and Lepore 1992). Given that the meaning of a term depends on its overall role in a conceptual scheme, it's not clear that any two individuals ever have the same meanings. Presumably individual differences in conceptual schemes are quite common. You might know a lot more about poodles than I do. In other words, you might draw many more inferences based on your 'poodle' representations that I ever would. If this is the case, we clearly don't share the meaning of the term 'poodle'. But, things are worse. Not sharing the meaning of 'poodle' means we don't, given a conceptual role theory, even share the meaning of the term 'two'. Since the meaning of a term like 'two' depends on its relation to all other terms in a conceptual scheme (including 'poodle'), and you and I have different inference-individuated terms (i.e., versions of 'poodle') in our conceptual schemes, the meanings of *all* of our terms are different. Furthermore, the meanings of my terms right now will be quite different from the meanings of my terms a few days from now (assuming I learn at least one new inference). Given this sort of criticism, Lepore (1994) concludes that conceptual role theorists must endorse the claims that "no one can ever change his mind; and no two statements or beliefs can ever be contradicted (to say nothing of refuted)" (p. 197). This extreme form of relativism would make explaining such important cognitive phenomena as communication and conceptual change impossible.

6 Two-factor theories

A common theoretical move to make in philosophy of mind has been to avoid the problems of causal theories and the problems of conceptual role theories by combining these two kinds of theories into one 'two-factor' theory. Exemplars of this sort of theory are those proposed by Ned Block (1986) and Hartry Field (1977). On these theories, causal relations and conceptual role are "two distinct components" or two *independent* aspects of the meaning of a term (Field 1977, p. 380). However, these *are* taken to be two parts of one thing: "the two-factor approach can be regarded as making a conjunctive claim for each sentence" (Block 1986, p. 627) or "referential meaning is *part of* meaning" (Field 1977, p. 399, italics added). In other words, both aspects of meaning, be they reference and sense, extension and intension, denotation and connotation, or what ever we would like to call them, are part of *meaning* generally.

If we look again at the problems of meaning that were historically of greatest concern, we notice that there are, in fact, *two* problems. There is the problem of understanding the world/mind relation *and* the problem of determining the nature of internal, mental transformations. Perhaps, then, it makes the most sense to consider both problems when constructing a theory of meaning. This, of course, is precisely the route taken by two-factor theorists. That, then, is one possible motivation for holding such a theory.

---

[2] Harman (1987) avoids such criticisms by allowing his 'causal' roles to extend out into the world. However, Block (1986) shows that this makes his theory a two-factor theory, which means it has other difficulties to deal with (see section 2.5-6).

A second possible motivation is one that I have already hinted at. If we have a two-factor theory, we *should* be able to solve all of the problems of causal theories and conceptual role theories. Notice that the problems faced by these theories are mutually exclusive. In other words, problems for causal theories are solved by conceptual role theories and vice versa. This means that if we can successfully combine these two kinds of theories we will have the best of both worlds and thus a theory that solves all the problems. However, things aren't so easy.

7 Problems with two-factor theories

It is central to two-factor theories that the factors are independent. However, this raises a grave difficulty for such theories. In criticizing Block's theory, Fodor and Lepore (1992) remark "We now have to face the nasty question: *What keeps the two factors stuck together?* For example, what prevents there being an expression that has the inferential role appropriate to the content *4 is a prime number* but the truth conditions appropriate to the content *water is wet*?" (p. 170). If, in other words, there is no relation between the two factors it is quite possible that massive misalignments between causal relations and conceptual role occur; I will call this the 'alignment problem'.

It is clear that two-factor theorists take themselves to be explaining *one thing* (i.e., meaning), but given the alignment problem it is not clear what could possibly be *the* meaning of a given neural state. In what sense could *a* meaning be defined by the conjunction of '4 is a prime number' and 'water is wet'. The only sense in which the referential aspect is *part of* meaning is the same sense in which Venus is *part of* the set of 'me and Venus' – by stipulation. If we really think meaning is unified, as even two-factor theorists seem to think, the alignment problem is a serious problem indeed.

There is a second difficulty with two-factor theories. Lepore (1994) has pointed out that if meaning is to be a conjunction of a causal and a conceptual role factor, then the relativistic problems that confronted conceptual role theories will be problems again. If we think meaning is determined, even partly, by conceptual role then any change in conceptual role is a change in meaning. As we saw in section 5, this sensitivity to changes in conceptual role makes shared meanings, conceptual change, and communication difficult to explain – at least *more* difficult to explain than on a straight causal theory.

8 A strategy for constructing a theory of content

Given the difficulties with each of causal, conceptual role, and two-factor theories what options are left for a new theory of content? Before I answer this question, I would like to consider the problems themselves in a little more detail. Note, in particular, that two-factor theories seem to *almost* have all of the resources needed to solve standard problems with content. There are two difficulties that remain, however, the alignment problem and problems with relativism.

Consider the difficulties with relativism first. What, exactly, is the problem that is faced by conceptual role and two-factor theories when it comes to relativism? Recall that the meanings of all terms in a conceptual scheme are relative to that scheme according to these theories. The problem, according to critics, is that this entails that for any two people (or the same person at two times) the meaning of a given term is never *the same* for both people. However, that isn't obviously a problem in itself. It's only a problem if we think meanings *should be* the same in the first place. Thinking that meanings should be *exactly the same* has a number of untoward consequences for causal theorists.

Proponents of causal theories note the ease with which causal theories can explain communication and shared meanings. If meanings are causes, the story goes, meanings are shared because causes are shared. Even ignoring problems with individuation of causes, there is still something odd about this claim. In particular, meanings don't seem to be things of the you-have-it-or-you-don't variety. The vast literature on vagueness hints at this fact (Fine 1975; Dummett 1978; Fuhrmann 1988; Williamson 1994). If the meaning of a term is its causes, and causes *aren't* vague, meanings shouldn't be vague because causes aren't. If, in other words, piles cause me to token my 'pile' representation, then the meaning of the term 'pile' is quite determinate; it's all those things that make me token that representation. Of course, the meaning of the term isn't determinate, it's vague. This means that if causal theorists think causes aren't vague, causal theories of meaning can't explain a ubiquitous property of semantics.

Perhaps, then, causes are vague. But, what would this claim mean for a causal theorist? It would amount to the claim that nomic relations between classes of objects and our concepts would have to be imperfectly fixed. In other words, some dogs would not token our 'dog' concept (if all dogs did token our dog concept, then the

previous problem would arise). What does it mean for our concept to be imperfectly fixed? It would mean that some people token their 'dog' concepts for some dogs that other people don't token their 'dog' concept for. This would account for the type-level 'dog' concept being imperfectly fixed on the type-level set of dogs. However, and here is the problem, this also means that two people seldom actually share the concept 'dog'. Under such a picture, two people on share the 'dog' concept if and only if they call all and only the same things dogs. This, it seems, is probably an unlikely state of affairs. Of course, relativism of one concept, like 'dog', doesn't entail relativism for all concepts as in the case of conceptual role theories. But, given the problem with the 'dog' concept, there's little reason to think that *any* two people will share *any* concepts. Therefore relativism is just as much a problem for such causal theories.

What all this means is that the causal theorist faces a dilemma. Either causes are determinate and vagueness can't be accounted for, or causes are vague (i.e., imperfectly fixed) so vagueness can be accounted for, but people seldom share the meaning of their concepts. So, causal theorists can either explain vagueness and be subject to relativist concerns, or not explain vagueness at all.

It seems clear to me which horn of the dilemma should be embraced: charges of relativism should fall on deaf ears. Rather, meanings should be taken to be more like body parts. Two body parts can be the same (e.g., your body part 'nose' and my body part 'nose' are the same body part), even if non-identical (e.g., our noses are differently shaped). Similarly, meanings can be the 'same' even if non-identical. Notably, in both cases, 'sameness' is a matter of degree. Given the characterization of meanings (or senses) in section 4 of the last chapter, there is a natural hypothesis about how to determine the sameness of meanings. In particular, meanings will be similar to the extent that they ascribe the same properties to their referents. Of course, under such a characterization of sameness of meanings we will need a different theory of communication, shared meanings, etc. Such theories will need to show how, for example, communication is possible despite non-identical meanings. Such a theory would presumably have a much easier time explaining different degrees of miscommunication since it would be based on a theory with degrees of sameness of meaning. Explorations of theories of communication are clearly beyond my scope, but I take it that it isn't *prima facie* unlikely that a theory of communication (and concept change, etc.) can succeed even if meanings are never *strictly speaking* identical. If this is at least a live option, I think we can safely ignore charges of relativism like those voiced by Lepore (1994).

These considerations make a good case for disregarding the problems of relativism. This is a benefit for both conceptual role and two-factor theories. However, I think it is more of a benefit to two-factor theories because there are no resources available to conceptual role theories that can explain the relation between meanings and truth conditions. In other words, I don't think conceptual role theories *can* solve their remaining problem. However, I do think that the alignment problem of two-factor theories can be solved, but it won't be solved by a traditional two-factor theory.

In particular, alignment won't be a problem if we can describe each of the two factors in terms of a third *underlying* factor. If we can provide a unified description of conceptual role and causal relations, then we know precisely how these two factors are aligned. They are aligned because they are simply two different consequences of the same underlying description. If such a factor exists, it will provide for a unified explanation of meaning. The strategy I will adopt in constructing a theory of content, then, will be one of attempting to find something that can fill this role. However, I won't return to an explicit consideration of what this factor might be until chapter 5. In the meantime, I will show, more specifically, what kinds of questions such a factor must be able to help us address (chapter 3) and discuss how I think we should go about finding such a factor (chapter 4).

9 Summary

I have surveyed contemporary theories of content and outlined the difficulties faced by each. I have shown that current causal theories don't provide a satisfactory solution to the problem of misrepresentation. Conceptual role theories, in contrast, suffer from the inability to satisfy intuitions that meaning and truth are closely related. Two-factor theories, while solving these problems independently, cannot account for the unified character of content. In particular, two-factor theories suffer from the alignment problem; i.e., the problem of showing how the factors relate. I have suggested that two-factor theories hold the most promise for solving the difficulties faced by other theories. In order to solve the problems of two-factor theories, I have proposed that we should seek an explanation of each factor in terms of some other underlying factor. If we can find such a factor, the alignment problem will be solved because the two factors will simply be descriptions of some *one* underlying process – a process that underwrites meaning generally.

CHAPTER 3

Family Ties

And 't is a shameful sight / When children of one family / Fall out, and chide, and fight. –
Isaac Watts (1674-1748)

1 Introduction

It is often as important to ask good questions, as it is to give good answers. In both philosophy and science, conceptual revolutions have occurred because of good questions. Descartes ushered in much of modern epistemology and metaphysics by asking: "[H]ow do I know that He has not made it so that there would be not earth at all, no heavens, no extended thing, no figure, no magnitude, no place and yet that all these things would seem to me to exist not otherwise than they seem to now?" (Descartes 1641/1990, p. 93). Einstein reconceived physics by asking: "Are two events (e.g., the two strokes of lightning A and B) which are simultaneous *with reference to the railway embankment* also simultaneous *relatively to the train?*" (Einstein 1961, emphasis original). Both questions are good because they suggest new ways of thinking about problems.

In all likelihood, the questions I am interested in answering are not nearly so monumental. It is nevertheless important to say what, precisely, the questions are. We have already seen, in the last chapter, some of the answers philosophers have provided for the question "What is content?" But, I take it, if we better define the questions that need to be answered by a theory of content, then we might be able to find specific ways to avoid the problems had by these theories. I have outlined a general strategy to avoid these problems, now I would like to be specific enough about the questions that need to be answered to be able to construct a theory that reaps the benefits of adopting that strategy.

I have a second major goal in this chapter as well. I am interested in showing why the questions I pose should interest *both* philosophers and neuroscientists. So far I have focused on the philosophical accounts of content. But, there are reasons to think that 'purely' philosophical accounts are unlikely to provide a satisfactory solution to the problem of neurosemantics. This is especially true if the problem itself spans philosophy and neuroscience. If, in other words, I can show that neuroscientists and philosophers are interested in the same questions, then neither philosophy nor neuroscience is more likely to solve such problems. In fact, it is likely that a satisfactory solution will have to depend on insights from both fields.

In this chapter I set about the two tasks of defining the questions and showing that they should interest both disciplines, by first characterizing representation in general. I then argue that, for a complete understanding of representation, we need to address representational problems in both neuroscience *and* in philosophy that are often wrongly thought to be distinct. Finally, I pose the set of questions that need to be answered in ordered to solve the problem of neurosemantics. In sum, the purpose of this chapter is to carefully define the problem I am trying to solve in the remainder of the dissertation.

2 Representational in-laws

Representations are everywhere. In particular, they are everywhere in neuroscience and philosophy of mind. But, are they everywhere the same? On the face of it, it seems not. Neural spikes, firing rates, neural populations, images, syntactically structured symbols, and abstract concepts are all called 'representations'. But, why are they all called representations? That is, what characteristics, if any, do they share? In this section, I argue that they *prima facie* share (along with many other things), the ability to enter into a particular kind of relation; the representation relation. And, more than this, they are commonly taken to do so.

The representation relation is at least a three-place relation: $\{X\}$ represents $\{Y\}$ with respect to $\{Z\}$. I say 'at least' here because it is possible to characterize representation as a higher-order relation (e.g., $\{X\}$ represents $\{Y\}$ with respect to $\{Z\}$ with $\{W\}$ for the purpose of $\{V\}$). However, to solve the problem of neurosemantics we only need a theory that accounts for the first three relata; moreover, any fewer is insufficient. It is true that in many cases, 'represents' is used as if it were a two-place relation: e.g., "This pen represents a dog." However, in cases of 'representation designation' (i.e., "I hereby declare this thing to represent that thing") there needs to be a

designator. In the pen/dog case, the designator either assumes the role of the third relata, i.e., "This pen represents a dog *to me*," or has left the third relata elliptical, i.e., "Let this pen represent a dog *to you*."

But who would play the role of such a designator in the case of mental representations? In other words, why can't we say that a mental representation *just represents* something (by virtue of its functional role, or causal relations, etc.)? Wouldn't it be careless to posit a 'someone' that the mental representation is 'to'? Clearly, that would lead to an infinite regress. In fact, I think it would be careless to posit a *'someone'*. But notice that 'something represents something to *someone*' is not the characterization I have suggested. Rather, I take it that whenever a mental representation represents something, it always does this *with respect to* some *system*. 'Systems' are much different that 'someones' and the representing isn't '*to*' a system, it is '*with respect to*'. The latter denotes a context, not an agent.

Consider two common theories of representation: causal role theories and conceptual role theories. Each clearly identifies a three place relation in describing representation: for causal role theories there is the representation, the thing it represents, and the context under which it is a representation and not just an effect[1]; for conceptual role theories, there is the representation, the thing it represents, and the role it plays (i.e., its context as defined by the system of concepts). So, claiming that representation is a three-place relation is nothing new.

One of the reasons that the three place relation is so ubiquitous stems, I believe, from the nearly universal commitment amongst philosophers and neuroscientists to understanding neurobiological systems as information processing systems (see e.g. Dretske 1981; Bialek and Rieke 1992; Van Essen and Anderson 1995; Rieke, Warland et al. 1997; Fodor 1998; Koch 1998; Eliasmith in press). Formally, the information relation is a three place relation: {channel} carries {information} with respect to a {receiver} (Reza 1994, p. 2). These three places are necessary and sufficient for an adequate and general definition of Shannon and Weaver-style information (Shannon 1948/1949). Also notice that they loosely align with the three places of the representation relation as defined above (i.e., channels and vehicles carry information and content with respect to receivers and systems). So, construing neurobiological systems as information-processing systems that represent naturally leads to a commitment to a particular kind of representation relation. In particular, it's not surprising that both the representation relation and the information relation are three place relations since the nature of the latter informs intuitions about the nature of the former given such a (rather common) commitment.

In addition, there is evidence that neurons represent extrinsically, as opposed to intrinsically. In other words, they represent because of their place in the system as a whole, not because of some fundamental property. There are at least three good sources of evidence for this claim. First, the general plasticity of the brain has been well catalogued (Karni, Meyer et al. 1995; Rauschecker 1999). One example comes from the ability of motor cortex to 'rewire' when an appendage such as a finger is lost; the neurons representing the finger are recruited to represent other, neighboring body parts (Wall, Kaas et al. 1986). Second, there are only a few classes of cells compared to the many representational roles they play. For example, pyramidal cells that represent, say, edges in early visual cortex are physically indistinguishable from pyramidal cells that represent motion later on in visual cortex. Third, the amount of cortex devoted to a representation can vary with changes in experience, but there is no evidence that the cells themselves change (Rauschecker 1999). Given this evidence, and given that neurons participate in representational relations, we should think that their place in the system is important for defining that relation.

If my argument from authority (i.e., all the other representational theories assume three places), my argument from analogy (i.e., neurobiological systems are information processors, information theory demands a three place relation, so neurobiological information processing (via representations) unsurprisingly depends on a three place relation), and the empirical evidence are not convincing, then perhaps the best support for the claim that representation is usually assumed to be a three place relation will come in the next chapter, where I describe how researchers actually characterize the relation. It is clear that in practice the representational relation is three places (see section 2).

Given, then, that there are three places, I'd like to suggest the following schema for the representation relation (the 'representation schema'):

A {vehicle} represents a {content} with respect to a {system}.

---

[1] For Fodor, such a context is defined by the nomic relations that obtain. In particular, those that have asymmetric dependencies and those that do not. For Dretske, the context is determined through an evolutionary and learning history.

Terminologically I have adopted the traditional names to the first two relata, 'vehicle' and 'content.' I call the third relata the 'system' simply in an attempt to be general. However, researchers in neuroscience and philosophy of mind tend to narrow the scope of this schema to include only natural biological systems (particularly human beings). In other words, they are not interested in representation *writ large*, but only in neurobiological representation. I adopt this narrower perspective as well. So, the third element of the schema will be, for my purposes, 'the (complete) nervous system'. This is in contrast to understanding 'system' as either a more general physical system or as an abstract representational system (like a language) (see e.g. Goodman 1968).

The overlapping research programs of neuroscience and philosophy can be understood in terms of a mutual interest in explaining the elements of this schema and the relation it defines. Neuroscience, for example, tries to determine the properties of neurons and the complex systems they form, ever with an eye to understanding how both the elements and their amalgamation work. Talk of representation in neuroscience is undeniably rampant (see e.g. Felleman and Van Essen 1991). For example, if a neuron fires relatively rapidly when an animal is presented with a certain set of stimuli, the neuron is said to "represent" the property that the set of stimuli share (see e.g. Desimone 1991). Of course, a neuron that normally responds to, for example, faces while it is in visual cortex won't respond to faces when the neuron isn't *in* the visual cortex. Therefore, the neuron's systemic relations are important. So, neuroscientists are interested in the elements of the representation schema: neurons (vehicles) represent the stimulus (content) with respect to (in the context of) the brain (system).

Philosophers, too, are deeply interested in representation. In fact, they have been asking questions about representations since the time of Aristotle.[2] Perhaps it is not surprising, then, that no aspect of the representation schema has gone unexamined by philosophers. Contemporary philosophers have spent time wondering about the various properties of different kinds of vehicles (Haugeland 1991), the nature of representational systems (Dretske 1988), but probably the most effort has been spent trying to understand how to determine what a representation is of (i.e., the content) (Dennett 1969; Millikan 1984; Fodor 1987; Dretske 1988). Philosophers, then, abstractly probe the elements of the representation schema: various kinds of representations (vehicles) represent things (content) with respect to a representational system (system).

Being able to characterize, even this roughly, both neuroscience and philosophy in terms of the representation schema shows something of their family ties. Each is determined to better understand the relata and relation defined by the schema. However, for those convinced (by Fodor for one; see chp. 1) that this is a mere similarity that is hiding deeper differences, more argument is necessary. That is the purpose of the next section.

3 Sharing the problem of neurosemantics

Robert Cummins (1989) has identified what he feels are the traditional scientific and philosophical problems of representation. They are the "Problem of Representations" and the "Problem of Representation" respectively:[3]

1. *Problem of Representations (Scientific)* – When confronting the Problem of Representations, we try to determine "which states and processes are involved in which [representational] activities and how" (ibid., p. 1). Cummins argues that this is a problem for empirical science (ibid., p. 1).

2. *Problem of Representation (Philosophical)* – When confronting the Problem of Representation, we attempt to account for the "nature of the (mental) representation relation" (ibid., p. 2). Cummins thinks this is a philosophical problem (ibid., p. 1).

Although this distinction might capture the relative emphases of science and philosophy, I think that the identification of one problem as philosophical and the other as scientific goes against the demonstrably common goal of neuroscience and philosophy to fill in the representation schema.

As I showed in section 1, both neuroscientists and philosophers are engaged with both kinds of problems. Neuroscientists explore solutions to the 'philosophical' Problem of Representation by making representational claims about neural firing rates. In particular, they try to discover what kind of causal relation the representation

---

[2] See Cummins (1989) for a brief history of theories of representation in philosophy.

[3] Dretske also has something like this in mind with his distinction between "representational facts" versus "facts about representation" (Dretske 1995, p. 3).

relation is. Similarly, philosophers construct representational theories that depend on the particular processes and states they posit. In fact, one might characterize the philosophical debate between connectionists, symbolicists, and dynamic systems theorists as a debate over the importance of various kinds of processes and states to cognition (Newell 1990; Churchland and Sejnowski 1992; van Gelder and Port 1995; Eliasmith 1996). Thus, no one field has anything like a monopoly on either of these representational problems. Rather, solutions to either of the representational problems are informed by solutions (perhaps partial) to the other.[4]

What Cummins has, in effect, proposed, is that understanding the representation relation and understanding the relata of the representation relation (representations) are two independent problems. If the problems *are* independent, then there may be grounds for claiming that one is philosophical and the other scientific. However, I would like to show that there are good reasons for thinking that there is no such independence in the first place.

There are at least four arguments against this independence. The first is a historical argument. Historically speaking, all of those interested in either problem have posed solutions to both. The Stoics, for example, identified impressors, impressions, *and* a causal connection. Descartes identified sense objects, perceptions, *and* a causal connection. Fodor has identified physical objects, mental representations, *and* a nomic connection. Neuroscientists have identified stimuli, neural firings, *and* a causal connection. So, given that prior solutions have not been provided independently we shouldn't expect to generate solutions independently. However, technically speaking, this argument contains a logical fallacy (i.e., *argumentum ad verecundiam*).

So, for those inclined to disregard the preceding argument as fallacious, perhaps a more logical argument will suffice. Technically speaking, a relation in first-order logic is defined by a set of ordered *n*-tuples of objects in the universe. This is a very straightforward way in which the relata (elements of the ordered *n*-tuples) and relation (the set of those ordered elements) are not independent. You simply can't have the relation without relata.

For those inclined to think that this argument from first-order logic doesn't have much to do with relations in natural language, perhaps epistemological considerations will help. Epistemological theories generally fall into one of three categories: coherentist, foundationalist, or a combination of the two (sometimes called 'foundherentist' (Haack 1993)). Anyone who holds a standard form of coherentism will grant that relations are closely tied to their relata because knowledge about either is conceptually tied to the other (Lehrer 1974; BonJour 1985). Anyone who holds a version of foundherentism will outright grant that empirical (i.e., scientific) data can make an important difference to your metaphysics (Quine and Ullian 1970; Thagard 1992).

Lastly, for those who find appeals to historical precedence arbitrary, think first-order logic is a poor model of natural language, and hold a foundationalist theory of epistemology, consider the following (I will use Fodor's theory to give examples of the following premises):[5]

a) The representation relation must be defined by necessary and sufficient conditions (i.e., intensionally as opposed to extensionally in the case of logic). Example: *x* represents *y* iff *x* is nomologically related to *y*.

b) The foundations that justify the definition of the relation must be independent of the foundations that justify identification criteria of the relata. Example: Reasons for accepting the nomological definition of representation can't be reasons for accepting that certain objects can fill the roles of *x* and *y*.

c) Therefore, if philosophy is just in the business of offering definitions of relations, identification criteria for relata can't help determine what definitions are right. Example: What *x* and *y are*, or *can be*, can't determine the correctness of relational definition.

d) For a relational definition to be right, it must allow only certain kinds of entities to enter into the relation. Example: One of the reasons nomological relations might be right is because only

---

[4] Cummins (1989) also has a sense that the problems are closely related, claiming of the Problem of Representation that "this question ... can be answered only by examining the scientific theories or frameworks that invoke mental representation" (p. 26). However, it is not clear that Cummins sees the relation going both ways or that their interdependence is as strong as I suggest.

[5] An analogous argument can be offered to show that science is interested in the relation as well as the relata.

things with physical properties can be representations and only things with physical properties enter into such relations.

e) Therefore identification criteria (e.g., has physical properties) help determine which definitions are right.

f) Therefore philosophy is not just in the business of offering the correct relational definitions.

In a sense, this is a weaker version of argument number 2 (the first-order logic argument). In this case, the claim isn't that relations *just are* sets of ordered pairs; i.e., that every property of the relata help define the relation (since every property determines the identity of the objects in the relation). Rather, the claim is that at least *some* properties of the relata inform how we define the relation.

I have presented four reasons for thinking that Cummins is mistaken when he distinguishes scientific problems about the representation relata from philosophical ones about the representation relation. So, in order to completely characterize either the relata or the relation, it is necessary to address both representational problems. But, in order to solve both problems, we have to ask good questions. Not surprisingly, these questions will cross the traditional boundaries between philosophy and neuroscience.

4 Basic vehicles, higher-order vehicles and thirteen questions about representation

4.1 Two kinds of vehicles

The main purpose of this section is to be explicit about the questions we must answer in order to solve the problem of neurosemantics; i.e., to understand how neurobiological representations have the many kinds content they do. Before posing these questions in section 4.2, however, it will be useful to draw the distinction between *basic* vehicles and *higher-order* vehicles. This distinction gives a sense of what everyone interested in neurosemantics agrees on. Furthermore, it is a distinction that will help to shape the questions in the next section.

Notice that in the examples I provide in section 2 vehicles are neuronal firings for neuroscientists, while for philosophers of mind, vehicles are anything that supports a representation (including words, sentences, and images). Vehicles in these two cases seem to be quite different. However, both disciplines are committed to the materialist position that the brain underlies all vehicles.[6] In this simple sense, then, both disciplines are committed to neurons being *the* vehicles on which all other vehicles depend in neurobiological systems: if there were no neurons, there would be no words, sentences, or images. Of course, the same claim holds for certain chemical compounds, or electrons and other elements of matter. Why not draw the 'basic' line somewhere else? The reason, I think, is that neurons have a unique spatial and temporal discontinuity within neurobiological systems. They are the largest physically distinct objects that don't themselves contain strong spatial discontinuities. And, their temporal behavior is salient at the same scale because of the strong temporal discontinuities between neural spikes. In addition, neuroscientists and philosophers take neurons to be basic because neuroscientists have had success *already* under this assumption. Neurons, as functional units, are something of a highest common denominator among a vast array of neurobiological systems.

The privileged place of neurons in current neuroscience (and accepted by philosophers (see e.g. Dennett 1969; Dretske 1988)) suggests a distinction between 'basic vehicles' and 'higher-order' vehicles. With respect to our mental lives, the basic vehicles are neurons and higher-order vehicles are the likes of mental images, words, and mathematical variables. Notably, I don't distinguish the physical neurons themselves from their behavior (e.g., output signals or spikes). So neurons are basic as a functional object, not a physical one. This avoids questions of whether the *neurons* are vehicles or whether *neural responses* are vehicles. I take it that both are vehicles, and thus group them by considering neurons as *functional units*.

Basic vehicles, then, are those vehicles that ultimately support all neurobiological representation. A simple consequence of this is that if someone were to posit a vehicle that could not possibly be supported by neurons, it would be ruled out as a representational vehicle in neural systems. In addition, it is presumably the case that

---

[6] Of course, many researchers are not materialists (e.g. Eccles 1974; Nagel 1974; Jackson 1986), but I am assuming materialism for the course of the thesis as it seems to be the most promising and widespread hypothesis (c.f. Churchland 1993).

anything that does not change neurons' behavior does not change the representations in a system. For example, if a blow to the head doesn't affect neuronal function (and doesn't change sensory input, of course), then it's irrelevant to the representations in the head. So, we can consider neurons as 'basic' because in every unequivocal case of mental representation, neurons are there. In other words, neurons are *necessary* for the existence of higher-order vehicles and mental representations in neurobiological systems.

In contrast, what counts as a higher-order vehicle is definitely more 'up for grabs.' It is these higher-order vehicles that are at issue in many debates: e.g., the feature detectors/filters debate in neuroscience (Van Essen and Gallant 1994) and the symbolicism/connectionism/dynamicism debate in philosophy (see e.g., Newell 1990; van Gelder and Port 1995; Eliasmith 1996; Eliasmith 1997). However, we do know that higher-order vehicles must be constructed, in some sense, out of basic vehicles. In other words, relations *between* basic vehicles will determine how they constitute the relevant higher-order vehicle.[7] The utility of this distinction will become clearer in chapters 5-8. However, this brief characterization should suffice to make it clear what the distinction is and why it might be important.

4.2 Thirteen questions

Given the preceding considerations concerning the nature of the representation schema, its interest to both philosophers and neuroscientists, and the distinction between basic and higher-order vehicles, answering the following five questions should result in a theory of how to complete the representation schema and thus solve the problem of neurosemantics (to the satisfaction of both philosophers and neuroscientists):

1. What are the basic vehicles?
2. What are the higher-order vehicles?
3. What determines the content?
4. What is the system?
5. What relations hold between vehicles, contents, and the systems they are in?

This last question can be clarified by identifying the six addition questions it subsumes:

6. What is the relation between basic and higher-order vehicles?
7. What gives a basic vehicle its content?
8. What is the relation between the basic vehicles and the system it is in?
9. What gives a higher-order vehicle its content?
10. What is the relation between a higher-order vehicle and the system it is in?
11. What is the relation between a vehicle's content and the system it is in?

These eleven questions can be pared down to nine. Most obviously, question 5 is now redundant. However, question 3 is also redundant for a more subtle reason. If taken to ask "How, in general, do we identify the content of a representation?", then question 3 is asked in a more detailed way by the combination of questions 7, 9 and 11. If each of 7, 9 and 11 are answered satisfactorily, there would be nothing left to know about question 3; i.e., we'd know that the content of the vehicle just is what is given to the vehicle in such-and-such a way (questions 7 and 9) as determined by such-and-such a relation to the rest of the system (question 11).

So, re-arranging the questions into a more logical order, and taking questions 3 and 5 to be redundant, we are left with the following nine questions about representation in the brain:

1. What are the basic vehicles?

---

[7] The 'order' of a relation is simply the number of levels of relations entering into a given relation plus one. So, a first order relation is of the form '*X* {relation} *Y*' where *X* and *Y* are not relations. A second order relation is of the form '*X* {relation} *Y*' where *X* and *Y* are either or both relations, and so on.

2. What are the higher-order vehicles?
3. What is the relation between basic and higher-order vehicles?
4. What is the system?
5. What is the relation between the basic vehicles and the system they are in?
6. What is the relation between the higher-order vehicles and the system they are in?
7. What gives a basic vehicle its content?
8. What gives a higher-order vehicle its content?
9. What is the relation between a vehicle's content and the system it is in?

These questions capture much of what we wish to know about the brain itself, but ignore the important role of the external environment (see chapter 2). Any discussion of content is going to be incomplete without reference to the relation between the system, its representations, and the world. Representations are about things in the world, after all. For these reasons, content has traditionally been considered, at least partly, a world-brain relation (see chapters 1 and 2). These considerations give rise to four more questions, bringing the total to thirteen:

10. What is the relation between the basic vehicle and the external environment?
11. What is the relation between the higher-order vehicles and the external environment?
12. What is the relation between content and the external environment?
13. What is the relation between the system and the external environment?

Notably, these last four questions may also be redundant. In fact, certain theories currently on offer assume that they are. For example, given a causal theory, if we know what gives a vehicle its content (questions 7 and 8) then we know what the relation is between the vehicles and their environment (questions 10 and 11) – the answer in both cases is cause. However, it isn't obvious that these questions are necessarily redundant because, for instance, a two-factor theory distinguishes the content-determining conceptual role relations from the representation-environment causal relation. I take it that whether these questions are redundant or not will only become clear given a satisfactory theory of content. So, it is best to retain them until it becomes obvious one way or the other.

Note that formulating the questions in this way makes the interdependence of Cummins' two problems about representation more evident. Questions 5 and 6, in particular, are about neither representations nor content exclusively, they concern both. This is clear when we consider two contemporary theories of content. Proponents of conceptual role theories hold that once we know the relations between a vehicle and the system, we know what the content of the vehicle is (Block 1986; Cummins 1989). In this case, questions 5 and 6 would be about content. However, we may think instead that content is causally determined. And, we may still be functionalists about vehicle individuation. In this second case, questions 5 and 6 would be about characterizing vehicles. So, our answers these more specific questions may be answers to either of the two representational problems (depending on our answers to other questions). This, then, is another example of how representational problems are not independent as Cummins suggests.

The only one of these thirteen questions which neuroscientists and philosophers agree on an answer to is, somewhat by definition, the first: neurons are the basic vehicles. Both neuroscientists and philosophers have suggested or assumed answers to the remaining questions, though perhaps focusing on a subset. Nevertheless, the disciplines share the common underlying goal of answering all of these questions because they share the common problem of neurosemantics. Another way of stating the goal of this thesis is that it is an attempt to provide a framework for answering all of these questions (and perhaps providing satisfactory answers to at least a few).

5 Summary

I began by arguing that representation is a three-place relation. In particular, the relation can be captured by the representation schema: A {vehicle} represents a {content} with respect to a {system}. This representation schema, I suggested, can be used to describe the problem of neurosemantics in both neuroscience and philosophy.

To show this, I provided examples of the application of this schema by both neuroscientists and philosophers. I then provided additional arguments to show that problems about what the representations are (a 'scientific' problem) and problems about what the representation relation is (a 'philosophical' problem) are not independent as has often been presumed. This means that the problem of neurosemantics is best addressed by considering insights from *both* neuroscience and philosophy.

In the last section, I suggested that philosophers and neuroscientists have already reached some measure of agreement about representation in neurobiological systems. In particular, both presume neurons (*qua* functional units) are basic vehicles; i.e., are the parts out of which all other, higher-order, vehicles are somehow built. Given this distinction, and the previous considerations about representation, I posed thirteen questions that, if answered, would solve the problem of neurosemantics. The remainder of this thesis is dedicated to describing a methodological and theoretical framework for answering these questions.

CHAPTER 4

A New Perspective on Representational Problems

The many truths we cling to depend greatly on our point of view.
– Obi-Wan Kenobi

1 Introduction

In this chapter, I am concerned with determining the right *way* to answer the thirteen questions about representation. There are reasons to think that neuroscientists and philosophers share not only an interest in the representational problems I have outlined, but that they also share an approach to solving these problems. In this chapter I argue that their shared approach is flawed, or at least incomplete. More importantly, exposing this methodological flaw provides insights into constructing a theory of content. Although the considerations I present in this chapter may occasionally seem far a field of a *theory* of representational content, I show, in the end, how these very same methodological considerations provide a deeper theoretical understanding of what is important to such a theory. More specifically, I propose the first major piece of the theory I am constructing; the *statistical dependence hypothesis*. This hypothesis tells us how we can determine what a given representation is *about.*

But first it is important to see why traditional approaches to characterizing representation by both neuroscientists and philosophers are problematic. The difficulty stems from the shared assumption that the best way to characterize the representation relation is from what I call the 'observer's perspective,' instead of what I call the 'animal's perspective'. I eventually discuss the close relation between the observer's and animal's perspectives in section 5, but I begin by distinguishing them in order to highlight the limitations and strengths of adopting either perspective exclusively. In section 3, I present examples from both neuroscience and philosophy that capture this methodological prejudice in favor the observer's perspective. In section 4, I show why assuming this perspective can hinder our understanding of the representation relation. By way of contrast, I provided a detailed example of how adopting the perspective of the animal can result in a simpler characterization of the representation relation. Then, in section 6, I show how these considerations lead to a characterization of the relation between vehicles and what they are said to be about.

2 Two perspectives, one problem

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

Consider, again, the problem of neurosemantics described in chapter 1. It is about the nature of representations inside a neurobiological system. The information-processing neurobiological system 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 it does have a perspective, and that perspective is relevant to answering the questions we need to ask, then we *may* be able to adopt either perspective – that of the observer, or that of the neurobiological system – when addressing representational problems about *that* system.

My use of the term 'perspective' may bring to mind concerns with subjective experiences or consciousness (e.g., along the lines of Nagel 1974), but I mean to avoid such discussions. I have in mind something much weaker than a conscious perspective. 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 are 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 (presumably, more so than consciousness) 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 animals and observers have access to different information in a given situation. An animal (and each of its information processing sub-components) can only access information available through sensory receptors. Properly situated observers can access that information, as well as information available through their own sensory receptors about the same situation. In other words, the observer has two sources of information; the animal's receptors, and their own.[1] Given the current state of neuroscientific inquiry, only a small number of neurons can be recorded from simultaneously. Thus, the 'animal's perspective' as I am using the term, generally refers to a tiny part of the total information available to an animal. Nevertheless, I think there is an important lesson to learn even from this limited access to the animal's 'total' perspective.

Most neuroscientists and philosophers 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 theories can change when we adopt the perspective of the animal.

In his book *Content and Consciousness*, philosopher Daniel Dennett (1969) also realized that the animal's perspective is the more natural 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 that 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). However, biologically reasonable theories of representational content, i.e., theories of the kind I am interested in constructing, 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

---

[1] 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.

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.

Experiments adopting this methodology are performed to characterize shape-related responses in neurons in early parts of visual cortex such as V1, V2 and V4 (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 occurs given a stimulus. 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 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 has dominated, and still 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 schema introduced in chapter 2 would 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}. 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. In section 4, I discuss how I think we can, at least partially, avoid this result by adopting the animal's perspective.

Before I do, however, it is important to show that philosophers, too, have adopted related tactics in trying to characterize the representation relation. Consider, for example, Fred Dretske's (1988) approach. He argues for a distinction between three types of representational systems:

Type I – Systems with no intrinsic power of representation at all; e.g., a pen used to stand for a unicorn.

Type II – Systems that use natural signs as conventional representations; e.g., falling sand particles used to represent time.

Type III – Systems that use their intrinsic indicator functions as representations; e.g., the 'bug detector' cells representing bugs to a frog.

In fact, the Problem of Representation arises only in the third case because in type I and type II systems the representational relationship is stipulated by a user (what I called 'representation designation' in section 1 of the last chapter). So how does Dretske come to understand the representational relationship in type III representational systems? 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., $p(r|s)$).[2] This is important because, as I discuss in more detail in the next two sections, I think $p(s|r)$ has been wrongly ignored. But, if Dretske talks about $P(s|r)$, how can I claim that the related probability function has been ignored? The reason is that Dretske (1981) claims that $P(s|r)$ has to be unity, i.e., that there *has to be* the stimulus in the environment given a particular neural response (and given background knowledge and certain channel conditions) 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. 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 (like the account offered in chapter 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 perspective of the animal'. That is, Dretske's theory eliminates the perspectival nature of $P(s|r)$ by forcing a criterion of a unitary conditional 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 are no limits on the information available to the ideal observer, 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

---

[2] It is important to 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 $p(x_k,y_k)=P(X=x_k,Y=y_k)=p_k$.

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: the probability of a response given a stimulus.

These examples from neuroscience and philosophy, 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. In the next section, I discuss some of the problems with adopting this assumption. In the section following that, I describe an alternative methodology for approaching representational problems that avoids these difficulties.

4 The strangeness of taking the familiar route

Neuroscientific experiments such as those discussed above are intended to address *both* representational problems because they help to characterize a physical process that is correlated with external stimuli, and they 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. This is the natural perspective for us, as scientists, to take. We have a system we are interested in understanding, we know what we are presenting to the system, and we have a means of measuring the system's output. 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 will 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.[3] 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 spikes 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 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).

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

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

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 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, on 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

---

[3] This can be undertaken by an observer, and nevertheless not adopt an observer's perspective *about what is being represented.* I discuss this more fully in section 5.

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 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 (or its information processing parts) 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.[4] 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. 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 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

---

[4] 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).

the difference between using the stimulus-given-response conditional probability and the more traditional response-given-stimulus conditional probability (see Figure 1).

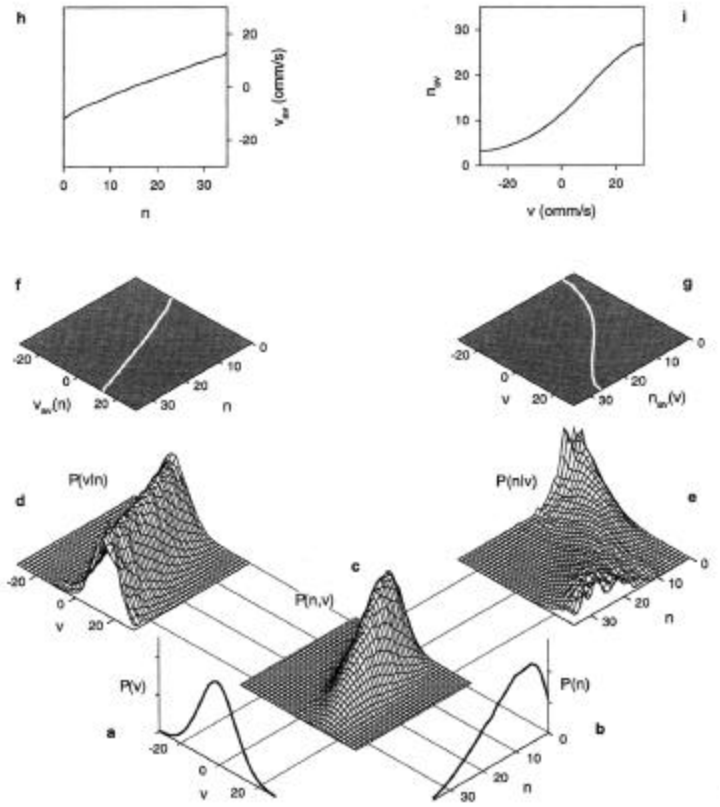


Figure 1: Joint, marginal, and conditional probability functions (a, b, c, d, e), and the differing characterizations of the stimulus/response relationship (f, g, h, i) depending on the conditional used (from Rieke, Warland et al. 1997).

Figure 1 demonstrates the important differences that can arise from taking the observer's perspective versus the animal'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). This is the problem as solved 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.

If the animal's perspective is advantageous, as this result suggests, should we abandon neuroscience, psychology and philosophy 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. 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). And, this is a difference that can make a difference, as the blowfly example shows.

In section 2, I promised to discuss how 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}. Notice, of course, that this content is *completely determined* by the stimulus presented by the observer. In other words, the content is {that face *x* degrees from *y* degrees}, 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 would 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}. However, notice the slight difference in the content in these two cases. Under the standard methodology the content is {*that* face *x* degrees from *y* degrees}. Under the alternate methodology the content is {*that there is a* face *x* degrees from *y* degrees}. This difference can be expressed by noting that in the first case, the content is identical to the referent, but in the second case, the content is a property ascription in the form of an hypothesis about the world. So, the referent is the same in both cases, but the content is different. Another way of understanding this difference is to notice that what the displacement, *x*, is, is determined in a *much different way* in each case. Under the standard methodology, it is determined by *a priori* knowledge about what is being presented to the cell. Under the alternate methodology, the displacement 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.

6 The statistical dependence hypothesis

Taking the alternate methodology seriously provides important insights into the nature of representational content. Recall two things that we have learned along the way: 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 is 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 a neurobiological system. I will call this the statistical dependence hypothesis:

> *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.*[5]

Notice that the hypothesis suggests that content is determined by responses, 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 relies on this ubiquitous 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 conditions[6]), 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:

---

5 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 $I(x;y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} r(x,y)\log\frac{r(x,y)}{r(x)\,r(y)} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} r(x,y)\log\frac{r(x|y)}{r(x)}$, so the mutual information of two events is $I(a;b) = \log\frac{r(a|b)}{r(a)}$. Usher (unpublished) adopts mutual information as a means of understanding representation. Similarly, I adopt mutual information as a measure of statistical dependence in chapter 8.

6 Note that the term 'stimulus conditions' isn't intended to be especially perception oriented or especially response oriented. I discuss examples of *both* perceptual representation and motor representation in chapters 6 and 7.

> *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.

Perhaps a simple example will 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 neuron 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[7] 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.[8] This is simply a case of misrepresentation.

There are many more things that need to be said about the statistical dependence hypothesis, but I will leave further discussion until chapters 5-8 where it can be put in the context of a more complete theory of content, and thus better defended. The reason I have introduced the hypothesis here is that its initial formulation relies on what we have learned from considering the problematic methodology assumed by most neuroscientists and philosophers. 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 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.

7 Summary

There are significant shortcomings of the traditional methodology employed in neuroscience and philosophy, and there is an important alternative worth investigating in greater detail. I have characterized the difference between methodologies as one in perspective; the observer's perspective versus the animal's perspective. There is evidence that adopting the animal's perspective can simplify the project of characterizing the representation relation. As well, the distinction between perspectives puts important constraints on our theory of meaning. The statistical dependence hypothesis is one aspect of a theory of meaning that is consistent with these constraints. The remainder of this thesis is engaged in the project of further exploring exactly what kinds of insight we can gain into the representational problems faced by philosophy and neuroscience when we adopt the perspective of the animal.

---

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

[8] Of course, this raises the deep philosophical worry about how we can justify distinguishing 'velocities' from 'nearby flashes' as distinct sets of causes. I consider these worries in more detail in chapter 6.

CHAPTER 5

A Theory of Content

A theory is something nobody believes, except the person who made it. – Attributed to Albert Einstein (1879-1955)

1 Introduction

So far, I have proposed a series of somewhat disconnected considerations regarding a theory of content. I have suggested that a good strategy for building such a theory is to look for a single factor that underlies both causal and conceptual role factors (chapter 2). I have argued that a theory of neurosemantics will have to consider both philosophical and neuroscientific results regarding the nature of content and have proposed a precise set of questions both disciplines are interested in answering (chapter 3). And, in the last chapter, I suggested that we should adopt the animal's perspective and the statistical dependence hypothesis when constructing a theory of content. Together, then, these chapters show that we need to consider philosophical and neuroscientific results: chapter 2 being an example of the former and chapter 4 being an example of the latter. In this chapter, I begin the process of bringing these considerations together to develop a theory of content.

2 Some assumptions

To begin, I will be explicit about what I assume in constructing this theory. I take the assumptions to be non-controversial for the majority of contemporary philosophers of mind and neuroscientists. They are explicitly shared by at least Fodor (1998), Dretske (1988), Harman (1982), Millikan (Millikan 1984), Block (1986), Cummins (1989), Churchland (1989), and Dennett (1987). They are so common in neuroscience as to be unarticulated; they are background hypotheses paradigmatic of science.

First, I assume materialism. I take materialism to be the thesis that all there is, is matter. Matter is whatever has physical properties and enters into physical relations. Physical properties and relations are the properties and relations that can be characterized by the physical sciences. As regards mental phenomena, this means that every mental token is a physical token. But, as mentioned earlier (see chapter 1), I remain agnostic as to whether the relation between mental and physical *types* is one of reduction, elimination, or independence.

Second, I assume that I am offering a naturalistic theory of meaning. That is, a theory that is expressible without either 'that'-clauses or semantic terms. A theory, in other words, that doesn't use terms like 'understands that', 'believes that', 'means that', 'denotes', or 'refers to' except insofar as to explain those terms. Naturalistic theories are intended to show how meaning or representation relations are part of the natural order. It would, of course, be circular to employ meaning-related terminology to explain how meanings are, in some sense, reducible to non-meanings.

From the first two assumptions, it follows that causes are probably going to be important for explaining content. If we want to explain meanings in terms of non-meanings and non-meanings are things that have physical properties and enter into physical relations, where cause a central physical relation, then we likely want to be explaining meanings in terms of causes. Now, this says nothing of what the important sets of causes are: they may be internal or external, personal or social, or possible or actual. The point of a theory of content is to say precisely which causes are relevant to content determination. Of course, such a theory may also say something about other kinds of physical relations that are important for meaning, but I presume cause is one that *must* be addressed; especially since the natural order is usually taken to be, at bottom, a causal one.

3 Causes and conceptual roles

To say that I will assume cause plays an important role in content determination is not to say anything very specific. The purpose of this section is to be more specific. It is surprisingly common for philosophers offering a theory of content that relies on cause *not* to offer a theory *of* cause. It is surprising because if the theory of cause is a bad one, any theory of content dependent on it should be expected to have serious difficulties. It is even surprising for conceptual role theorists not to offer a theory of cause. Since most such theorists are offering *naturalistic* theories, they are still in the business of explaining meaning in terms of causes, it just happens to be

that internal causes are the important ones. For them not to worry about what causes are, is for them not to worry about the limitations on the possible (inferential, causal, computational) relations between concepts. So, despite the title of this section, I will be most concerned with defending a theory of cause. I take it that naturalistic descriptions of other kinds of relations will be essentially causal ones as well.

What we need from a theory of cause is two things: 1) a token-level causal account; and 2) a type-level account of lawful dependencies. We need both aspects because a workable naturalistic theory of content will depend on causal regularities. Causal regularities are both *causal* and *regular*. Explaining token-level causes will account for the 'causal' part, and explaining lawful dependencies will account for the 'regular' part. The account I offer here tightly interweaves these two aspects of a causal theory for representational content.

Cause is sometimes called "the cement of the universe," and has been analyzed in numerous conflicting ways since the ancient Greeks (Mackie 1974). There are a number of theories currently on offer, including Humean-inspired 'constant conjunction' theories, the identification of causes with necessary and sufficient conditions (or the more sophisticated INUS (insufficient but necessary, unnecessary but sufficient) conditions), and probability-based theories of casual relations (see Sosa and Tooley 1993). It is not clear that any of these analyses have succeed in capturing *the* notion of causation, leading some to the conclusion that "[t]he attempt to 'analyze' causation seems to have reached an impasse; the proposals on hand seem so widely divergent that one wonders whether they are all analyses of one and the same concept" (Kim 1995). Notably, I don't need to analyze *the* notion of causation; I'm not concerned with explaining every possible notion of cause. Rather, I need only a scientifically and philosophically respectable notion that will underwrite a theory of content.

Let me begin, then, by examining the notion of lawful dependency. Hume was right to say that all the evidence we can ever get for causation comes from what he calls "constant conjunction." In particular, I agree that, at least *prima facie*, "[t]he constant conjunction of our resembling perceptions [i.e., impressions and ideas], is a convincing proof, that the one are the causes of the other" ((Hume 1739/1886, I, 314). I agree, in other words, that representations are closely related to lawful dependencies, and that all the *evidence* for lawful dependencies comes in the form of constant conjunctions. Because things in the world result in our seeing them, we say that they cause us to represent them: this dog is constantly conjoined with my representing this dog.

But, how constant is 'constant'? One of Hume's fundamental insights is that we are in an epistemological predicament. Whatever the metaphysical state of the universe, we are limited in such a way as to only have finite data about causal relations. If we notice 'nearly whenever this happens, that happens' we are entitled to infer lawful dependencies between this and that. Of course, this doesn't mean that that is all there is to cause. Hume, for one, thought otherwise.[1]

Nevertheless, under this empiricist characterization of our relation to causal regularities, we can at least begin to *identify* lawful dependencies by there being a somewhat constant conjunction between causes and effects. One way to write this more precisely is as follows:

Event A is causally related to event B if and only if $P(A,B) \neq P(A) \times P(B)$.

Or, in other words:

Event A is causally related to event B if and only if $\neg(P(B|A) = P(B)\ \&\ P(A|B) = P(A))$.

These definitions are equivalent. They mean that two events are in a lawful dependency relation if and only if they are not probabilistically independent. This kind of definition is Humean in the sense that it presumes that the fact that two events happen together more than by chance (i.e., they are conjoined) is evidence that they are lawfully related. The conjunction in this case is probabilistic and may not be 'constant' or 'necessary' in Hume's sense, but the definition is very much in the Humean spirit.[2]

---

[1] Hume says: "An object may be contiguous and prior to another, without being considered as its cause. There is a NECESSARY CONNEXION to be taken into consideration; and that relation is of much greater importance, than any of the other two above-mention'd" (ibid., I:378). However, Hume also realized that the notion of a 'necessary connection' is not one directly accessible to us: "From the mere repetition of any past impression, even to infinity, there never will arise any new original idea, such as that of a necessary connexion" (ibid., I:389).

[2] In fact, a probabilistic definition is preferable to Hume's, as he has conflated causation with a necessary connection; presumably cause is just as relevant to nondeterministic processes (see e.g. Dretske and Snyder 1972; Anscombe 1993).

This account is clearly unsatisfactory because it is perfectly symmetrical. If you notice a probabilistic dependence between two events, you don't know which caused which or if they were just caused by some third event *C*: this won't do for a general understanding of causation. What we need, then, is a means of *directing* causal relations. This is where a theory of token-level causes comes in.

David Fair (1979) has suggested that the transfer of energy-momentum (just 'energy' from now on) can provide the ingredient we need.[3] He argues for a physicalist reduction of the notion of cause to the concept of energy-momentum transfer as applied in physics. So, for example, gas tanks cause gas gauges to register gas levels because there is a transfer of energy-momentum between gas tank levels and gas gauges. Perhaps the buoyancy of the gas transfers energy-momentum to a bob in the gas tank, which transfers energy-momentum to a lever, which transfers energy-momentum to the gauge needle. Notably, this is a theory of cause independent of any probabilistic relations – and that is why it will do for token-level causes. Energy-momentum, then, isn't only a necessary ingredient for directing causes in a probabilistic theory; it's the substrate of cause itself. Still, probabilistic dependencies *are* a means of identifying that *some* energy-momentum transfer has occurred, and thus they are a means of discovering lawful dependencies. In fact, some scientists argue that these correlations *stem* from energy-momentum transference (Prigogine 1996, p. 78-81).

Fair's kind of physicalistic reduction of cause raises a number of concerns, many of which (including the individuation of energy-momentum transference, an acceptance of a purely physical object ontology,[4] and identification of energy-momentum across time) he has handled deftly. However, Sosa and Tooley (1993) raise three criticisms which Fair does not adequately address, but which I would like to deflect. The first is a concern about the analysis of the concept of cause. One *may* claim that causation must be an intrinsic relation and thus that it must be the same in all possible worlds. If this is the case, it is easy to dream up possible worlds in which there is cause and not, for example, transfer of energy. If the identification of cause with energy-momentum transfer is contingent, then it provides no hope for analyzing the concept itself. Whatever we may think of possible worlds in general,[5] the possible worlds necessary to support this argument are particularly problematic. They are, for one, more subject to Dennettian 'inconceivability' criticisms than most (Dennett 1991; Dennett 1995): whatever our intuitions tell us about a world that differs from the actual one with respect to one of the most fundamental quantities, they probably don't tell us much. But, there is a deeper difficulty: such worlds, I think, are *im*possible worlds.

In order to have a world in which anything happens at all (i.e., in which things are caused), we need at least two things: 1) a world and 2) change. It seems unlikely that we can have either of these without energy-momentum (just 'energy' from now on) transfer. If there is any 'stuff' in our world, then there is energy in our world because energy is just that stuff: energy just *is* mass, heat, charge, etc. If there is change in our world, then some stuff turns into other stuff and we get energy transfer: adding change to energy gives power, motion, current, momentum, etc. We can call energy whatever we like, but the fact remains that it is a fundamental characterization of any changing world. Energy transfer, understood in this way, is a functional role of sorts. But, it is so basic that it is *the* functional role. Nothing functions without energy transfer. To claim that you can conceive otherwise (see e.g. Sosa and Tooley 1993) is like claiming you can conceive round squares. Or, another way of putting it: To try and conceive of a changing world without energy transfer is like trying to conceive of a square without a geometry (not a particular geometry, any geometry). Not only are such worlds hard to conceive, they are impossible.

A second concern about Fair's reduction is that while someone may claim "it is true of any enduring object in this world that its temporal parts are causally interrelated" this "does not involve any transference of energy or momentum from one object to the other" (Sosa and Tooley 1993, p.4). In other words, because *temporal* parts of an object are causally related, but there isn't any energy transfer between temporal parts, energy transfer can't explain causal relations. Such concerns are best answered by noting that the assumed distinction between objects and energy is invalid. This is to say, as Strawson (1987 p. 260) does, that objects and energy aren't different things, they concurrently constitute the world. This is what Einstein's famous equation, $E=mc^2$, tells us. So we

---

3 Similar theories are suggested by Strawson (1987), Castaneda (1984), and Aronson (1971).

4 These considerations also show how my talk of 'events' above relates to a causal theory which assumes an object ontology. Events, then, are particular configurations of energy whose form or state determines which objects and properties are involved in causal relations (Fair 1979, p. 233-4).

5 Fair (1979) doesn't seem to think much of them: "Bizarre possible worlds rarely concern us" (p. 232).

can answer the objection as follows: if we allow objects to be identifiable by space and time co-ordinates, then there are two objects (the object at time 1 and the object at time 2), and there *is* energy transference; in particular, all or most of the energy at the spatial co-ordinates is transferred from time 1 to time 2. Identity over time of the object is explained by identity of the energy transferred between these two times. That, then, is the causal relation demanded by Sosa and Tooley.

The third objection offered by Sosa and Tooley (1993, p. 4) is that energy relations don't have the right kind of directionality. Fair, upon realizing this, suggests that time plays the requisite role; time directs causes. However, if the direction of time is to be explained by the direction of causation, as Sosa and Tooley rightly suspect, Fair is relying on a vicious circularity. However, all is not lost. Prigogine (1996, especially chp. 3) argues that the 'arrow of time' is a *result* of the correlations between particles that exchange energy. If this is true, perhaps Fair was too quick to appeal to time to enforce a direction on causation. The direction of time *is* to be explained by causation as Sosa and Tooley suggest. However, the notion of causation that is used to explain time, is precisely that which identifies causes with energy transfer. So, in fact, Sosa and Tooley (and Fair) were wrong to think that the directionality of causation needs to be explained *in addition* to an energy transfer theory of cause.

So, with this slight modification of Fair's original theory, we have just the tool we need for understanding representational content. Lawful dependencies are directed probabilistic dependencies, and the direction is determined by energy transfer. Token causal events are instances of energy transfer. In sum, any two events are lawfully dependent if and only if they are statistically correlated, where the effect is the event that receives energy from the cause, and the cause is the event that loses that energy.

Now I have said something more specific about what I take causes to be. However, I still haven't said what role such causes play in a theory of mental meaning. That, at least partially, is the job of the next section.

4 A skeletal theory

The theory I outline here is only a skeletal theory of content. The purpose of this section is not to convince you that the theory I present will work as it stands, but rather to show the general shape of the theory, and convince you of *where* further detail is needed. For example, I will concern myself mainly with what I will call 'occurent intentionality'; that is, the aboutness and meaning of currently active perceptions of current states of the world. Concerns with occurent intentionality lead us to ask: why is my representation of this dog about *this* dog, and what does it mean? This is in contrast to what I will call 'conceptual intentionality' which leads us to ask: why is my concept 'dog' about dogs and what does it mean? How to get from a theory of occurent intentionality to one of conceptual intentionality will be a major theme of chapter 8 and partially addressed in chapter 7. But I presume that a theory of occurent intentionality should get us on the right track, so I'll begin with one. As well, the details of the single underlying factor that I propose (and call a 'computational factor') are left to the next chapter. And, finally, worries about more subtle aspects of the theory, and the integration of the theory with the details I present in chapter 6, are left to chapter 7.

To begin, then, there are four important theoretical objects for explaining the representation relation: vehicles, contents, systems, and referents. Three of these, vehicles, contents, and systems, are taken directly from the representation relation. The third, referents, may or may not be distinguished from contents in a theory of content (Fodor (1998) and Dretske (1995, p. 30), for example, don't whereas Block (1986) and Cummins (1996), for example, do). Roughly speaking, referents are the same as the sensory objects posited by the Stoics and Descartes. They are the things in the world that the representations are about. *Prima facie* reasons for thinking they might be distinct from contents are: 1) the kinds of reasons that give rise to questions about the relation between the environment and a representation, independent of its content (i.e., questions 10-13 in chapter 3); and 2) the kinds of reasons that have inspired conceptual role theorists and two-factor theorists to claim that meanings are determined (or partially determined) by internal inferential relations (see chapter 2, sections 4 and 6). So, what a representational theory must do is explain what these four theoretical objects are and what the relations are between them.

Of the four, identifying the system is perhaps easiest. It is easy because the definition of the problem at hand (i.e., the problem of neurosemantics), determines the system of interest; it is the nervous system. Explaining the other three is where the most work lies. And, because a theory of *content* is necessary for explaining the representation relation, and content is the most difficult of the three to explain, let me begin with the other two. What, then, are vehicles and referents?

Vehicles are the internal physical objects (i.e., 'representations') that lie in causal/computational relations to one another and to the outside world. I am interested in *internal* vehicles (where 'internal' simply means 'inside the nervous system') because internal representations are the focus of the problem of neurosemantics. There are many kinds of internal vehicles. In chapter 3, I divided them into two groups, basic vehicles and higher-order vehicles. It is important to note that the distinction between basic and high-order vehicles is an overly simplified one. There are, in fact, *many* orders of vehicles and thus many levels at which content may be ascribed. As well, a vehicle of a given order may participate in any number of higher-order vehicles. Analogously, letters of the alphabet are vehicles, words are vehicles, sentences are vehicles and books are vehicles; they may all have content and they may help constitute each others' content, but it might not be, in any obvious sense, the same content or even content of the same kind.

I take the basic vehicles to be neurons as *functional* units (i.e., including their output, or spike train). Neurons are *basic* vehicles in the sense that they are an agreed upon minimal functional unit necessary for understanding content generally. Notably, being 'basic' in this sense doesn't tell us much about the kinds of content that a particular neuron may be carrying. Initially, we might argue that basic vehicles carry a basic kind of content; content about their immediate input signal. The reason, we could argue, that they carry content *about* the input signal is because their output signal has the highest statistical dependency with their input signal. However, given that we are interested in explaining internal representation of the *external* world, this immediate input/output description will only be a valid one at the sensory periphery. Under this description, non-peripheral neurons would only ever carry content about other internal states, but that's not what we want to explain. So, individual neurons (i.e., basic vehicles) can have more complex kinds of contents as well; i.e., they can be about more than just their immediate input signal. If, for example, a neuron in late visual cortex has a highest statistical dependency with motion in a certain are of the visual field, then it is about motion in that part of the visual field.

Higher-order vehicles are, practically speaking, 'just' groups of neurons. However, theoretically speaking, they are not just any groups of neurons. They are groups of neurons that together ascribe complex statistical properties (perhaps even *incredibly* complex properties such as 'a situation that is just', though I will limit my discussion to perceptual examples). Which neurons comprise *the* higher-order vehicles is a question that must be left to empirical investigation and theory building. The right set of higher-order vehicles will be the set that allows us to accomplish our explanatory goals, such as explaining the system's behavior.

Now on to referents. As I noted earlier, referents are the external objects that representations are about. The traditional way of picking out referents has been to rely on causes. Not surprisingly, then, the theory of cause outlined in the last section will play an important role in determining referents. However, having a theory of cause is not all there is to finding the right referents. As Dretske (1981, p. 26-33) rightly notes, a theory of cause doesn't tell you *which* causes are important for representational content; i.e., which causes are referents. Given the probabilistic theory proposed above, we must ask: which of the statistical dependencies is important for representational content? But, this question has already been answered in chapter 4 by the statistical dependence hypothesis. The right causes are those that have the highest statistical dependence under all stimulus conditions. Referents, then, are those things to which the vehicles are causally related and with which the vehicles have the highest statistical dependence. So they are the things that transfer energy to vehicles and, in situations of a given type, have the highest statistical dependence with the vehicles.

Notably, the referent under this theory is not necessarily what is determined by the more traditional reference relation. Reference, for one thing, is traditionally used to describe a property of the meaning of *sentences* and *words*, not neural states. So, the sentences 'water is wet' and 'H_2O is wet' share their reference, although they differ with respect to sense. They share reference because both are about the same stuff in the world, but they differ in sense because someone agreeing to one of them may not agree to the other. It is not at all clear, however, that we can speak of the reference of neurons. Kripke (1977), for example, notes a difference between speaker's reference and semantic reference. If a speaker gestures towards a glass of vodka and says 'the glass full of water is mine', Kripke wants to say that the speaker's reference is the glass, but the semantic reference is indeterminate (or non-existent), since there is no glass full of water. At the level of neurons firing, however, such a distinction does not even begin to make sense (in a different sense of 'sense'). Perhaps it doesn't make sense in this case because Kripke is working with a entirely linguistic notion, and I am not (for reason argued in chapter 1 section 5). In any case, because reference can be so subdivided and referents cannot, referents are not what are determined by reference, so understood. Furthermore, even for those who use reference in a non-linguistic way, it is a relation that can hold between representations and things we are not in causal contact with (e.g., dogs beyond my light cone). Referents, in contrast, are always in causal contact with what refers to them. So, in some ways my notion of referent is more general than reference, since things like

neural firings can clearly have referents. And, in other ways my notion is less general because it depends strictly on causal connections, and not on a more general notion of 'aboutness'.

Finally, there is content. The notion of content, as Cummins (1989) points out, is neither clear nor simple in contemporary philosophy of mind. He suggests, that it can be (vaguely) understood as "whatever it is that underwrites semantic and intentional properties generally" (p. 12). Semantic properties are like Dretske's (1995, p. 3) "representational facts" in that they tell us about the semantic (or representation) relation, not about things that are semantic (or representations). Content, so understood, underwrites meaning and the representation relation in much the same way; so providing a theory of content goes hand in hand with providing a theory of the representation relation. In this sense, understanding content is central to understand the representation schema as a whole. Recall that in chapter 1, I argued that content is the set of properties ascribed by a representation to a referent. Explaining content, and thus the representation relation, comes down to explaining how a set of properties is ascribed by a representation to a referent.

The basic theory I hold of how this occurs is nothing earth shattering. I think that content ascription is determined by the relations between vehicles and referents and the relations between vehicles themselves. These vehicular relations include relations between and within the 'orders' of vehicles. This view is, of course, reminiscent of a standard two-factor theory view. I have identified an internal factor (relations between vehicles) and an external factor (relations between vehicles and referents). However, the big difference between this theory and others is that I explain both factors in terms of a single, underlying 'computational factor'. In other words, I show that causal relations to the external world and internal transformational relations between vehicles can be described in terms of a single computational framework. Showing this is the job of the next chapter.

For now, I would like to point out some ways in which the two factors I am interested in explaining are atypical. First, the external factor in this theory is strictly causal. If something hasn't transferred energy to the vehicle, it isn't a candidate referent. This, as I just noted, is quite different from what many philosophers have taken to be external factors. Standard external factors are taken to bind my representations with anything my thoughts can be about. Thus, external factors can normally be related things I've never been in causal contact with (e.g., dogs beyond my light cone). I'll address concerns that this stricter constraint on external factors may raise in chapter 8.

The internal factor is also somewhat non-standard. I have claimed that the internal factor must account for transformations relating internal vehicles. I've also said that I will show how we can describe transformations as computations. Thus, it shouldn't be surprising that the internal factor on this theory is a computational one that defines representational transformations. For some, making *computational* claims about *biological* systems is nonsensical. But, I don't hold a typical (at least in philosophical circles) notion of what counts as a computation. As often understood by philosophers, computation is something that must take place over discrete symbols (see e.g., Cummins 1989; van Gelder 1995; Fodor 1998). This 'definition' of computation can be traced back to the pioneering work of Turing (1950). However, as many familiar with the field of computational neuroscience will tell you, one of the working hypotheses of the field (as may be obvious from practitioners' self-designation) is that biological systems *are* computational even though they don't (at least don't obviously) rely on discrete symbol manipulation (Eliasmith in press). The growing field of analog computation would also be disappointed to learn that what they are interested in just isn't computation because it isn't about discrete symbols (Uhr 1994; Hammerstrom 1995). As Churchland and Sejnowski (1992) put the point: "once we understand more about what sort of computers *nervous systems* are, and how they do whatever it is they do, we shall have an enlarged and deeper understanding of what it is to compute and represent" (p. 61). The kinds of computational relations we posit might look quite different if we suppose computations can take place over noisy analog values, rather than over discrete symbols. For one thing, the relations are bound to look much less like those common to first-order logic or natural language. For another, the language of dynamic systems theory and probability calculus will play a central role in describing such relations.

Given these considerations, then, the content of an occurent mental state is determined by the causal statistical dependencies it bears to events and objects in the external world and the sorts of transformations it licenses under current stimulus conditions. Which dependencies are the strongest is determined by the internal transformations that occur between a vehicle and its referent. The reason we ascribe these occurent states the *kinds* of content we do is because of their statistical dependencies under *all* stimulus conditions and the transformations licensed. Transformations and causal relations align reasonably because there is a way of describing both in terms of a single computational factor.

That, then, is a very rough sketch of the theory I wish to defend. One of the most glaring omissions is that I've told you what the objects of the theory are, and I've said that there are certain relations these objects enter into, but I haven't said much about what those relations are (except the causal one). Describing the computational and vehicular relations is where some neuroscientific details become important. As I noted previously, those details are left for the next chapter.

Recall, for a moment, the discussion in chapter 2. It may have become disconcerting, in light of that discussion, how 'causal' the theory I am offering is. In particular, I am claiming to solve the alignment problem by using a single computational factor. That computational factor, I said, can be used to describe causal relations. Furthermore, I said that transformations are causal because they can be described naturalistically as well. Perhaps, this 'computational' factor is just a different name for a standard causal factor. If this is the case, I might have a hard time explaining misrepresentation since most causal theories do. To allay such fears, I would like to show, briefly, how misrepresentation can be accounted for on this theory (for further discussion see chapter 8).

Simply put, misrepresentation occurs when the properties of the referent and the properties ascribed to the referent by the content do not match (see Cummins 1996). If the content of my neural state is 'dog' and the thing that is the referent of that state is a cat, I have misrepresented the cat. As Dretske (1995, p. 27) notes, there is another kind of misrepresentation as well; representation of a non-existent thing. This, too, can be accounted for by a comparison of content and referent. So, if a neuron fires whose content is '4 or so photons in such-and-such a location' but there were no photons at all, that neuron has misrepresented the environment. Notice that the first example of misrepresentation is at a high level, while the second is at a low level. Misrepresentational mismatches, then, can happen at any vehicular order.

More specifically, the theory I have proposed suggests that these mismatches are a result of mismatches between the application of statistical dependence hypothesis and the application of its corollary. If, in other words, the highest statistical dependency under these stimulus conditions doesn't pick out the same thing as the highest statistical dependency under all stimulus conditions, we have a case of misrepresentation. Of course, for any given occurent representation there could be misrepresentation concurrent with correct representation. There could be some vehicles that get it right (ascribe the right properties) and others that get it wrong (ascribe the wrong properties). There could also be some orders of vehicles that get it right and others that get it wrong. Of course, misrepresentation in a deep and interesting sense depends on the possibility of representation in a deep and interesting sense. Those kind of (mis)representations are going to depend on complex computational relations and the ability to construct interesting higher-order vehicles.

5 Summary

I began this chapter by explicitly stating the assumptions I make in articulating my theory of content. These assumptions include materialism, naturalism, and a commitment to the central role of cause in determining content. In order to be precise about the nature of cause, I outlined and defended a physical reductivist theory articulated by David Fair. In the remainder of the chapter, I provided a skeletal theory of content based on these assumptions and some considerations from previous chapters.

It is important to note that I have not presented and defended a full theory of content in this chapter. What I have done, rather, is to provide a general outline of how the considerations in preceding chapters and this one come together to provide a first guess at how content is determined. Thus, I don't think what I have *so far* described is satisfactory as a theory of content. In particular, I need to provide neuroscientific details that underwrite this story. And, I need to provide a philosophically oriented defense of the theory in light of previous theories and standard concerns about content. The most obvious omissions in the theory I have outlined so far include specifications of: 1) computational descriptions of causal relations; 2) the nature of the inter-vehicular relations; and 3) the relations between high- and low-order vehicles. In the next chapter, I show how each of these relations can be understood using the tools of computational neuroscience.

CHAPTER 6

A Neurocomputational Theory

[H]ypotheses will be put forward that 'leave the details to the neuroscientists'…this is admittedly armchair science with its attendant risks. – Daniel Dennett (1969, p. 42)

1 Introduction

If we take the analysis in the previous chapters seriously, we don't want to the 'leave the details' to anyone. The neuroscientific details are just as important as the philosophical framework for a complete theory of content. Ignoring either discipline prevents us from solving the problem of neurosemantics. We should *look* to neuroscience and related disciplines to help determine the details, we can't just leave them up for grabs as Dennett suggests. The attendant risks he speaks of are unacceptable ones. We might, for example, just end up redescribing the problem (see chapter 2 section 3), or end up with a theory disconnected from the world (see chapter 2 section 5), or not be able to explain how content can be unified (see chapter 2, section 7), or only answering some of the important questions (see chapter 3), or vicariously adopting a limited perspective (see chapter 4). The details, then, are very likely to inform nearly all aspects of a theory of content. If the details are significantly flawed, or absent, then theory is likely to be problematic. What, then, are my details?

Recall that in chapter 4 I concluded that best way to solve the problem of neurosemantics is by adopting the perspective of the animal. In addition, at the end of the last chapter I issued promissory notes to provide details for three relations that are central to the theory of content I outlined:

1. Vehicle/world relation (i.e., a computational description of a causal relation);
2. Inter-vehicular relations (at a single level); and
3. Basic/higher-order vehicular relations (across levels).

A question that brings the animal's perspective and these relations together is: How do we characterize these relations from the animal's perspective? How, that is, do we characterize these relations while taking into account the epistemic position of the animal? That epistemic position is, I have claimed (in chapter 4), one that does not have access to the stimuli itself, but only to the neural firings that a stimuli causes.[1] Taking this perspective will undoubtedly change our characterization of some these relationships, just as it changes the characterization of the representational relation in neuroscience (as discussed in chapter 3).

Before providing the details of these three relations, I introduce the concepts of encoding and decoding, which play a central role in the characterization of the relations. I employ these concepts to describe how we can understand the informational properties of neural firings and relate these neural firings to causal descriptions of the same process. As well, I show how encoding and decoding relationships can be used to capture the relation between basic and higher-order vehicles and the transformational relations between vehicles at a single level. In recent work, Charles H. Anderson and his colleagues (Anderson 1994; Van Essen and Anderson 1995; Anderson 1998; Eliasmith and Anderson 1999; Hakimian, Anderson et al. 1999; Eliasmith and Anderson forthcoming; Eliasmith and Anderson in press) have combined these two means of characterizing neural information processing into a single theory; a theory that provides a unified account of the details needed for a theory of content.

2 Conceptual apparatus

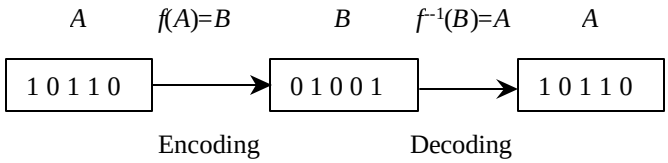
The terms 'encoding' and 'decoding' are used to describe mathematical functional mappings. Adopting these two terms implicitly assumes the presence of a *code* of some sort. Codes, of course, are used to communicate and carry content. Thinking of representation in neurobiological systems as a kind of code, allows us to conveniently adopt theories of communication – at least this far Dretske (1981) was on the right track. We can, of course, justify this convenience if the resulting theory is successful. For these reasons, I will adopt the standard

---

[1] Perhaps this can be seen as analogous to Kant's epistemological point that we can not know things in themselves (1787/1965, A299-300, A595-6), but perhaps not.

communications theory language of 'encoding' and 'decoding' to talk of functional mappings. Adopting these terms is important for another reason: they suggest a focus on the *implementation* of mathematical functions and neurobiological systems are clearly implementations. In particular, they are implementations to which we can ascribe certain contents. So, for the function f:$A \rightarrow B$ or $f(A)=B$ where A and B are sets of some kind, I will say f encodes A into B or f decodes B from A.

In an ideal case, the encoding and decoding functions are inverses of one another. In such a case, if an encoding converts a set of numbers (i.e., a signal) A into a set of numbers B, then the decoding will convert B back into A for any A and B. A simple example is depicted in Figure 2.



$$f(x \in A) = \begin{cases} 0 \text{ for } x = 1 \\ 1 \text{ for } x = 0 \end{cases} \qquad f^{-1}(x \in B) = \begin{cases} 0 \text{ for } x = 1 \\ 1 \text{ for } x = 0 \end{cases}$$

Figure 2: A simple example of an ideal encoding/decoding relation.

However, like most things ideal, it is safe to say that this kind of situation never occurs to an arbitrary degree of precision in real, implemented systems. The reason is the ubiquity of *noise* in real systems. If there was enough noise present in the example in Figure 2, the mapping between A and B might not satisfy the functional description. For example, B might have become {01011} if there was enough noise to 'flip' the fourth digit. Of course, noise is something that affects physical quantities, not numbers. Numbers provide a convenient and important way of labeling relative amounts of some physical quantity. The example of the encoding/decoding relation shown here is obviously an *abstraction* of a physical process of some kind. But, in real systems we must work hard to reduce the effects of noise. In the case of many engineered systems, like digital computers, that physical process is controlled in such a way that the numbering of the physical quantity (e.g., voltage) is very robust to noise. In other words, the effects of noise are more or less eliminated by establishing large differences between two neighboring values.[2] This is why there is a difference of about 5 volts between '1s' and '0s' on computer chips. Having established these relatively large voltage differences, we can forget about noise and get on with our programming; although we pay a price in relatively large energy consumption.

When we are faced with a system that we didn't engineer, or one that is subject to very large noise effects, we *can't* presume that noise has been nearly eliminated. For one thing, we may no longer be able to presume that the decoder is the inverse of the encoder. If, for example, we encode a low frequency signal (e.g., a voice) that is then transmitted over a channel (e.g., a telephone wire) that introduces lots of high frequency noise (e.g., a 'hiss'), the decoding that will give us the best reconstruction of the original encoded signal will not be the inverse of the encoding. The exact inverse would give us a reconstructed signal with lots of high frequencies (e.g., a 'hissy' voice) that weren't in the originally encoded signal. However, an *approximate* inverse that removed most of the high frequencies would give us a better reconstruction of the original signal. But, what do we mean by better?

Because we are concerned with implementations of encoding/decoding schemes, there are two important measures of 'goodness'. One measures how well we can reconstruct the original signal. The other, equally important measure, tells us how well we used the resources we have available. Let me begin by discussing the second measure. Luckily, we have a precise way of determining how well we are using the resources at our disposal; we can apply the tools of information theory (see e.g. Reza 1994). In particular, information theory will let us determine the maximum possible amount of information we can transmit over a given channel (the channel capacity). Information theory will also let us determine the amount of information in any given signal we transmit.

---

[2] In fact, there is still *some* probability that a 'high' voltage will fluctuate to the extent that it is read as 'low' or vice versa – this is one, very uncommon, way a computer may crash.

So, if the amount of information in a decoded signal is close to the information capacity of the channel, we have a good encoding/decoding scheme for that channel; the closer the better. Because many of the signals we are interested in characterizing are continuously transmitted, it's useful to have a 'per unit time' measure of this kind. In these cases, information theorists work with information *rates* (i.e., bits per second) rather than total information transmission. So a measure of the kind we are after is the ratio of the information transmission rate in a given signal over the maximum possible transmission rate. This is called the *efficiency* of the encoding/decoding scheme. The higher the efficiency, the better the use of resources by the scheme.

This measure of efficiency, like most things from information theory, is concerned only with the amount of information transmitted, not the content of that information. In fact, a scheme that inverted all of its input signals would have the same efficiency measure as one that didn't. So we need some way of determining that the content of the original signal is retained. The way to ensure that content is retained is simply to ensure that the original signal is reconstructed by the decoding. If a given input signal has certain content, then that content will be perfectly preserved in the case in which that exact signal is transmitted. In other words, you can't change the content if you don't change the signal.[3] Determining the discrepancy between two signals means minimizing the average difference between the original signal and the reconstructed signal (i.e., the signal estimate). A standard measure of this difference is the root-mean-square (RMS) error.[4] If the RMS error is low, then the reconstruction is a good one. In sum, we can know that an encoding scheme is good (i.e., preserves the amount and content of the information given the available resources) if it is highly efficient and has low RMS error.

With these conceptual tools from information theory, we can now look at some attempts to employ these tools to understand the characteristics of neurobiological systems. That is, we can look at the relation between the world and the systems we presume can represent it.

3 Basic level representation

To begin, I will present a characterization of the relation between the world and neurobiological systems at the level of the basic vehicles. In other words I will ask: how do the spikes at a peripheral neuron relate to the physical quantities impinging on it? Examples of this kind of transduction are retinal cells that detect photons, mechanoreceptors that respond to touch, and auditory hair cells that are sensitive to pressure waves. Much theoretical and experimental work has been done by Rieke, Bialek, Miller, and others that addresses just this kind of encoding in neurons (Theunissen and Miller 1991; Bialek and Rieke 1992; Rieke, Warland et al. 1997). These researchers have developed methods for decoding the information in neural spike trains using various tools, including those described in the previous section.

One clear example of employing these tools is provided by the work of Warland et al. (1992) on the cricket cercal system. Household crickets have two long thin appendages at the rear of their abdomen called cerci. These cerci have hundreds of tiny hairs, each of which acts to detect wind direction along a single plane. The hairs are arranged in various orientations, allowing the cricket to very accurately detect the direction and magnitude of small air currents. Warland et al. (1992) set up a delicate experiment in which he was able to manipulate a single hair and record the response of the corresponding sensory neuron. In this way, he is able to probe the relation between a known input signal (i.e., the mechanically controlled hair deflections) and its encoding into a neural spike train.

As discussed in chapter 4, a good way to characterize this relation is to adopt what I have been calling the animal's perspective. This means that although the researcher determines the precise input signals, the point of such experiments is to see how well a given signal can be reconstructed on the basis of just knowing the resulting spike train. Specifically, these experiments have focused on determining the decoding rule that can be used to reconstruct the original wind direction and magnitude changes given the spike train generated by the sensory neuron attached to the cercus hair. The experimental setup and results are shown in Figure 4.

---

[3] Presuming, of course, that you don't change the properties of the receiver. This is the case because the *a priori* knowledge of an observer helps determine the content of any signal.

[4] RMS error is technically only applicable as a measure of this kind *if* the noise we are dealing with is Gaussian; I will assume that it is.

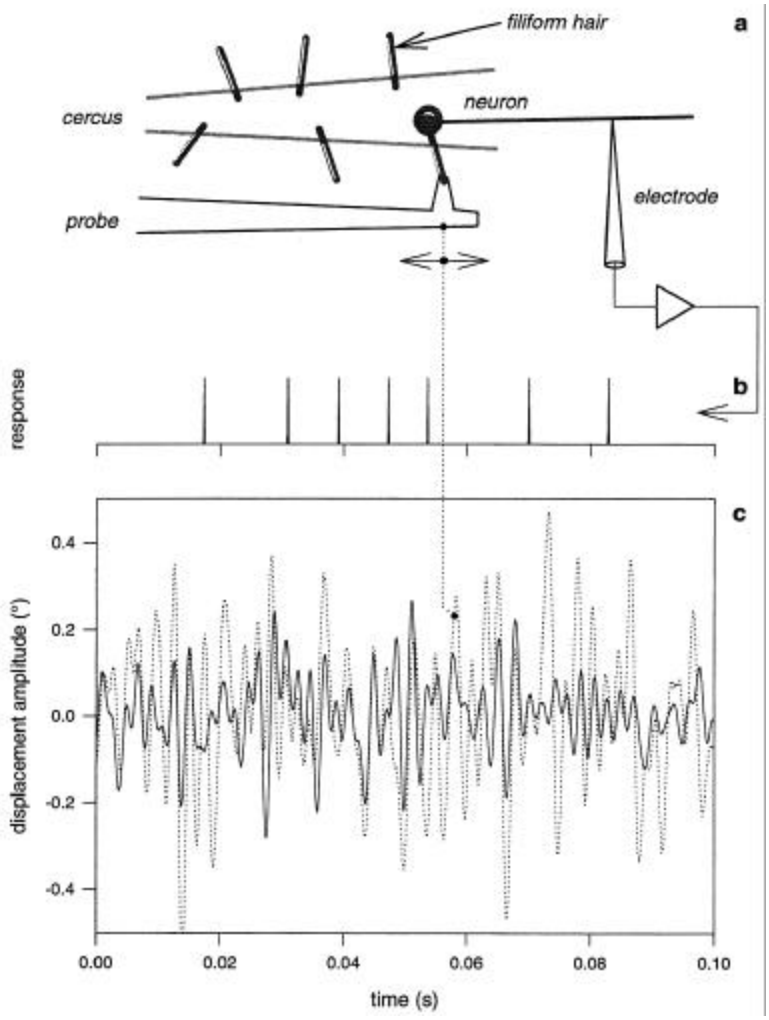


Figure 3: a) Experimental setup for testing the response of a sensory neuron attached to a cricket's cercus hair. b) Sample spike occurrences measured from the intracellular electrode inserted into the sensory neuron. c) A sample of the estimate (solid line) of the input signal (dotted line) generated from decoding the sensory neuron's spike train (taken from Rieke, Warland et al. 1997).

In this kind of experiment, the researchers are attempting to determine what information about the stimulus is available to a neuron that receives this spike train. The major achievement of this work is being able to decode the information the neuron has encoded through its receptive electrochemical mechanisms. Notably, determining what the right decoding rule is takes a lot of prior examples. In particular, researchers first must record the results of a large random sampling of signals the neuron is responsive to. This provides an estimate of the joint probability between a signal and a set of spikes (just as in chapter 4). Using this relation, the researchers can then make a good guess at what that rule is.

Surprisingly, the decoding rules tend to be quite simple – meaning that they are well modeled by a *linear* filtering of the spike train (Rieke, Warland et al. 1997, p. 170). In other words, we can use a linear filter to determine what, precisely, a neural spike train could be telling the animal about its input signal. This characterization of the decoding process can be expressed as:

$$x^{est}(t) = \int \sum_m h(t-t')d(t'-t[m])dt' \qquad (4)$$
$$= \sum_m h(t-t[m])$$

To begin to understand this equation, suppose there is a filter, $h(t)$, that looks like that drawn in Figure 4. Then, read from left to right, this equation says that the estimate of the input signal is equal to the signal you get when you place that filter at each spike time (i.e., $d(t[m])$) and then sum the result. So, as shown in Figure 4, this

equation says that to decode the signal encoded in a neural spike train simply replace each spike with the curve specified by the filter, and then sum those curves. The result is an estimate of the original signal.

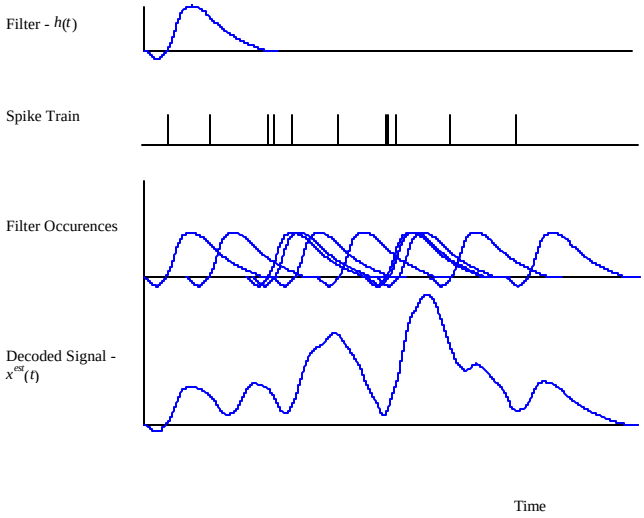


Figure 4: Decoding a temporal signal on a single neuron.

Another way of understanding this process is to think of the occurrence of a spike as indicating the presence of a signal with the form of the filter in the original signal. In a sense, the way to decode the signal is to assume that the signal is encoded by a filter-feature detector; the neuron fires whenever a certain feature (in the shape of the filter) occurs in the input signal.

There has been much work done justifying and closely examining the validity of this kind of decoder for spike trains (de Ruyter van Steveninck and Bialek 1988; Bialek, Rieke et al. 1991; Miller, Jacobs et al. 1991; Theunissen and Miller 1991; Rieke, Warland et al. 1997). Using these techniques, these researchers have shown that spike trains are within a factor of two of the maximum possible efficiency (Rieke, Warland et al. 1997, p. 173-4, 185). This kind of information transmission is impressive, but perhaps not unexpected from biology. The important point here is that these decoding procedures work quite well. So, assuming that biology actually has *perfect* information transmission, the reason for the factor of two difference would the encoding/decoding scheme. So, at a *minimum*, this decoding extracts about half of the *possible* amount of information in the spike train.

Notably, the possible information is information available *without any noise*. But there are lots of reasons to expect noise in biological systems. For one, synapses are unreliable in their release of vesicles into the synaptic cleft given the presence of an action potential in the presynaptic axon (Stevens and Wang 1994). Furthermore, the amount of neurotransmitter in each vesicle varies significantly, as does the ability of the presynaptic neuron to release the vesicles (Henneman and Mendell 1981). Lass and Abeles (1975) found that propagation along an axon introduces a few microseconds of jitter over a length of about 10cm of mylenated axon. Noise, then, is part of a neuron's environment. What this means for the encoding and decoding of information I have been discussing is that it is probably much closer than a factor of two to what biological systems actually do. If neurons operate in a noisy environment, they have no reason to decode information to the maximum possible limit – the 'extra' information they would be extracting is just noise, and is thus useless for the purposes of learning about the external environment.

But, how do we know that we are extracting the right information? In the case of the cricket, for example, Warland et al. (1992) found that the information rate in this experiment is around 300 bits per second. So, as Rieke et al. note (1997, p. 168-9), over a window of one second, the cricket could uniquely identify one out of about 10^{90} possible signals. Alternately, the cricket could uniquely identify one out of 2 possible signals every few milliseconds (positive or negative deflection for example). The information transmission rate itself is silent on the content of the signal but it certainly gives us some bounds. These bounds *help* restrict what information can *possibly* be represented by the neuron. If there are thousands of possible states every few milliseconds, the cricket *just can't* have access to all that content. Notice also that we are certain that in these experiments the reconstructed signals look like the original signals (i.e., the RMS error is small). Thus, even without knowing the details of how content is determined, we can preserve it. It is clear that a neuron *can* pass the signal itself, and hence (at least something of) the signal's content, reasonably completely. Of course, it would be odd if neurons

only transmitted information (so much for the transformations), but the point is that we have a means of reconstructing an aspect of the content (whatever it may be) when it's passed by a neuron.

What these results mean for the theory of content is that we can precisely understand the first relation in the trio identified above; the vehicle/world relation. In particular, we have a computational description of the causal process of neural firing. This, then, provides an explanation of the first factor for the theory of content I have been building. The reason we accept the computational description is because it is isomorphic to the causal process it is describing. There is, in other words, a one-to-one correspondence between, for example, mechanical input, neural transduction, and neural firings and the input signal, the encoding, and the output signal. We know the correspondence is isomorphic because we have a systematic mapping that is preserved in the face of transformations (i.e., relations in the causal description are captured by the computational description). In fact, the filters that are found are designed to do just that. They are chosen so that input signals can be recovered from output signals and, because this is successful, we know that the isomorphism is supported (see chapter 7 for further discussion).

Notice also that I have used a sensory neuron as my example, so the encoded signal is external (in this case, the wind direction). However, it could have been the case that the signal that the spikes were encoding was a signal from another neuron. From the perspective of the animal, there's no difference between the deflections of a hair causing a neuron to fire, and another neuron causing a neuron to fire. So, the second relation in the trio is also *partly* addressed; i.e., the inter-vehicular relations.

In fact both of these relations are only partly addressed, because this entire discussion has focused at the level of what I have been calling basic vehicles. As philosophers and scientists have realized for years, descriptions of phenomena at a low level may not be very practical or desirable at a higher level (e.g., quantum physics just won't do when it comes to understanding aerodynamics).[5] More complete characterizations of both of the relations addressed so far will only come with an understanding of how we can 'move up' from the world of neural firings in single neurons.

4 Higher-order representation

An important next step is to consider ways of modeling interacting populations of neurons. Laying aside, for the moment, the discussion in the previous section, I will consider a superficially contradictory means of understanding the behavior of neurons. Most current models of neural population behavior depend on average neural firing rates, rather than the precise timing of single spike transmission. Although the discussion in this section relies heavily on neural firing rates, I will return to the question of how average firing rates relate to encodings of precise spike timings in the next section.

Perhaps the simplest population level representation evident in the nervous system is that of an analog variable. The value of this kind of variable can be either transduced from the environment, as in the case of wind direction, or given by a previous ensemble of neurons as in the case of desired horizontal eye position (Eliasmith and Anderson 1999).[6] In the latter case, the desired horizontal eye position can be captured in a single analog variable (e.g., degrees from center) even though a large population of motor neurons is being used to determine the precise position (Seung 1996). We can thus consider this population to be encoding the eye position, which is then decoded by the interactions of the ocular musculature and the eye, and results in an actual eye position.

I will refer to the value of the analog variable, x, as the *higher-order* value (i.e., the value of the higher-order vehicle, x). This value is encoded by the *basic* level spike trains. Notably, the encoding and decoding relation between these levels is clearly *virtual.* In other words, the higher-order value isn't actually represented by the system separately from the basic values. It is more a matter of our *description* of the system that makes this encoding/decoding relation hold. Nevertheless, it is the same kind of relation, as I show. As well, this kind of encoding/decoding relation is not temporal, as in the example in the previous section, but population based. In fact, the temporal decoding of spike trains I discussed is relatively new. Because of this, it is a technique seldom

---

[5] I think this may be a reason why the adaptation of results from artificial neural networks to solving philosophical problems has been criticized. It just isn't made clear how high-dimensional vectors relate to language-like representations.

[6] Notably, questions of *what* crickets represent are independent of questions of *how* they represent. Whether crickets are said to represent 'wind direction', or something else, will depend on what is the best theory of cricket behavior but a general characterization of the representation relation can remain the same.

employed by neuroscientists. They have relied instead on neural firing *rates*. Firing rates are generally determined by counting the number of spikes in a relatively large window (e.g., 100ms) and dividing by the widow size (e.g., 10 spikes/ 100ms = 100 spikes/s = 100 Hertz).

Despite discarding information about the precise timing of individual spikes, much successful modeling has been done using spike rates instead of spike trains. So, rather than thinking of higher-order values as being encoded by spike trains, these researchers have considered spike rates to be encoding analog quantities. There are a number of neural systems that are very well modeled under this assumption (e.g., the nucleus prepositus hypoglossi used for controlling eye position (Seung 1996), macaque motor cortex used for controlling arm movements (Georgopoulos, Schwartz et al. 1986), and 'place cells' in the rat hippocampus (Wilson and McNaughton 1993)). For the time being, I will adopt this same assumption, with the understanding that discarding temporal information must be either justified or retracted.

More precisely, we can write the relation between the analog variable, its estimate, and the neuron properties as follows:

$$x^{est} = \sum_i a_i(x) k_i \qquad (5)$$

This equation states that the estimate of the encoded variable, x^{est}, is the sum over the neuron ensemble, i, of their weighted, k_i, firing rates, $a_i(x)$. The firing rate, $a_i(x)$, of any neuron, i, in the population is a function of the encoded variable, x. The function, $a_i(x)$, is called the neuron response function. The functions in Figure 5 are idealized versions of typical motor output response functions being used to encode an analog value, x, such as horizontal eye position.

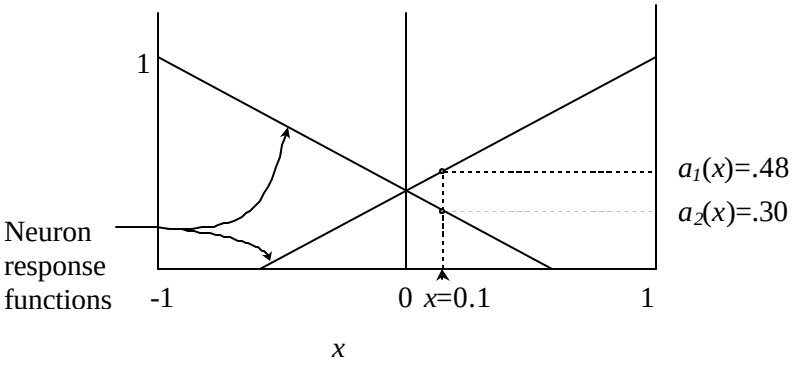


Figure 5: A population of two idealized neuron response functions where x is the analog variable to be encoded through neural firing rates, $a_i(x)$, as shown for a value of x=0.1.

Equation (5) is in a standard form for expressing a given variable or function in terms of other functions. Application of these equations to modeling neural systems has been well characterized and quite thoroughly explored (see e.g. Abbott 1994; Salinas and Abbott 1994). There are a number of properties of this kind of characterization that are particularly worthy of note.

First, in equation (5), there is no assumption about the particular form of the response functions that determine the firing rates. In practice, neuroscientists have found these functions to be generally non-linear. As well, they seem to come in many flavors, ranging from nearly linear to Gaussian (bell-curve shape) to multi-modal (multiple bell-curve shapes superimposed). The variety and complexity of the neural response function shapes makes it rather remarkable that linear decoding works so well (Rieke, Warland et al. 1997, p. 85-6). Nevertheless, if we determine the right set of weights, k_i, to use in our decoding, our estimates will be quite good.

This brings us to a second property of this equation: there are ways to determine good decoding weights. In particular, we can determine the weights by minimizing the difference between the estimated value and the actual value over all possible values of x. This will provide us with all the necessary weights for encoding the original input signal with little error. However, we often wish not only to transmit, but to transform the signal. In such a case, the weights can be quite different, but systematically so. They will be different in such a way that the estimate will now be a function of the original signal. Varying the decoding weights will thus change how the input signal is decoded. In other words, we can perform *transformations* by employing *biased decoding* (see section 5.2 for further discussion). It is important that these weights are not the same as standard connectionist network, or artificial neural network weights. Connectionists use the term 'weight' to refer to weights on connections

between neurons. The weights I am talking about are theoretical, higher-/lower-order decoding weights, not necessarily directly observable in networks of neurons (see equations (9) and (10) in section 5.2).

Third, this characterization generalizes extremely well. Though I will discuss this in more detail shortly (sections 5.2 and 5.3), it is worth pointing out here that just because I have used analog quantities in the above examples does not mean this kind of mathematical relation is so limited. In fact, we can use n-dimensional vectors, vector fields, or functions in place of the analog variables. Despite this generality such a formulation of neural function leaves a lot of neurobiology to be done. This kind of framework can help us understand and organize the results of neuroscience, but it won't determine those results. The shape of the neural response functions, for one example, is in no way dependent on this characterization. Neuroscientists must work hard to determine these response functions, and provide other biological constraints on the models that are proposed under such a framework.

Fourth, and last, one major reason this kind of characterization is so compelling is that it very naturally incorporates concerns with the effects noise. One way of thinking of neuron response functions is that they are a means of capturing the encoding process neurons use to convert analog quantities into neural spike rates. If the analog variable has a value of 0.1, for example, each neuron will fire at different rate, as determined by the response function. But why would we need more than one response function? As shown in Figure 5, each response function uniquely determines the value of x. But, because the brain is a physical system we know that the actual response functions won't, in fact, agree because of the ever-present effects of noise. Noise, then, is a major reason we need population codes at all. If we need to encode an analog value with greater precision than is possible by one neuron alone, we can pool neurons together to get a better estimate of that value. Since noise, by its very nature, is random, we can be sure that the effects of noise will cancel out over larger and larger populations. Biology clearly uses this tactic (Nicholls, Martin et al. 1992), and characterizations like those of equation (5) allow us to naturally incorporate the effect of noise on the quality of the encoding. The way noise plays a role in such a characterization is in the determination of the weights, k_i. Earlier, I over-simplified the procedure for determining the weights. These weights must, in fact, be determined by minimizing the input/output error given a certain noise profile. Again, there are well-established procedures for doing this (e.g., least squares, singular-value-decomposition). In other words, we can find those weights that give us the best approximation to the original value given our neuron response functions *and* their noisiness.

These four properties of the mathematical characterization of the encoding/decoding relation between neural firing rates and analog variables are useful ones. They allow us to begin to understand how higher-order vehicles are related to basic ones. Of course, the examples provided so far demonstrate only a small step in getting from neural firings to something 'more interesting' like a visual image, but it is often the first step that is the hardest.

To summarize thus far, we have initial characterizations of each of the three relations. The discussion of basic level representation showed how neural firings could be described computationally, essentially capturing the vehicle/world relation. I suggested that this same description could capture the relation between basic vehicles. In this section, I have discussed how we can understand the relation between these basic vehicles and higher-order vehicles. I have also given a beginning sense of how we can capture higher-order transformations (i.e., with biased decoding). In the remainder of this chapter I am concerned with building on these basics to give a fuller account of the second two relations.

5 A general theory

5.1 Putting time and populations together

To begin, I will return to the pressing question I posed at the beginning of the previous section: how do we amalgamate these two approaches I have discussed? How do we combine an approach concerned with the temporal encoding of individual neurons with an approach concerned with populations and somewhat atemporal rate codes, into a single theory?

As described above, firing rate is determined by counting the number of spikes in a "relatively large" window and then dividing by the size of that window. "Relatively large" here means large with respect to the average interspike interval (i.e., the mean time between two neighboring spikes). Rieke, et al. (1997, p. 31-2, 118-20) note that something very interesting happens as we change the size of this window. In particular, as the

window becomes very small, much *smaller* than the average interspike interval, we blur the distinction between a firing rate code and a timing code (i.e., a code which depends on the precise timing of individual spikes). They have shown, in other words, that there is a smooth transition between firing rates and timing codes.

As first recognized by Anderson (1998), the blurring of this distinction is important for combining the two encoding/decoding relations discussed in the previous sections. In particular, he considers the firing rates, $a_n(x)$, of equation (5) to be instantaneous firing rates, encoded by a temporal code like that of equation (4). Thus he derives the unified expression that captures the time-dependent *and* population characteristics of neural representation resulting in:

$$x^{est}(t) = \sum_n \sum_m h_n(t - t_n[m]) \qquad (6)$$

This equation, then, expresses the population level representation of an instantaneously time varying signal. It combines the important advances in reading neural codes of individual neurons with those of reading the codes of populations of neurons. The $h_n(t)$ in (6) is not the same as the $h(t)$ in (4). Rather, it is a combination of the weights, k_i, from (5), and the filters, $h(t)$, from (4). Notice that in (5) and (4), the weights and filters respectively are used to decode the signal encoded by the neurons. So, the weighted filters of (6) play exactly the same role; they are the decoders of the (population-temporal) neural code. The subscript, n, on the weighted filter means that each neuron in the population may have a different weighted filter. Together, these weighted filters (just 'filters' from now on) provide a means of reconstructing the time varying signal encoded by a neural population in a noisy environment. This equation, then, is the first step in precisely capturing and unifying all three of the relations identified in the introduction as being important to representational content: vehicle/world (temporal encoding); inter-vehicular (temporal encoding); and basic/higher-order vehicles (population encoding). I will say more on this in the next chapter.

5.2 Transformations

It may seem that all this talk of 'decoding' the signal from a neuron is rather strange. Am I really suggesting that a receiving neuron must decode a spike train to gain access to the signal at the other end of a transmitting neuron? Is there any neurobiological proof for this? Furthermore, does anyone really think that neurons simply pass *the same* signal that is their input? The answer to each of these questions is, as we would hope, a resounding 'No!'. The question, then, is what does this mathematical apparatus really provide? Essentially, I think it provides a powerful way to understand the *transformations* that signals in nervous systems can undergo.

Just as it was easiest to begin by introducing a simple higher-order vehicle (i.e., an analog variable), so it is easiest to begin with a simple transformation. The simplest of transformations is the one-to-one mapping of some value back onto itself. As mundane as this transformation may seem, it lays a foundation for understanding a huge variety of transformations in neurobiological systems.

What this simple transformation amounts to is the passing of a value from one neural population to the next. Taken together, these populations form a *communication channel* as depicted in Figure 6a. In this case, the transformation of a value onto itself could be expressed as:

$$y^{est}(t) = x^{est}(t) \qquad (7)$$

For simplicity's sake, we can assume that the neuron transfer functions are like those in Figure 5. The mathematical expression for such lines is shown in Figure 6b. What we need now is a means of determining the connection weights, ω_{ij}, between the two neural populations that will perform the desired transformation.

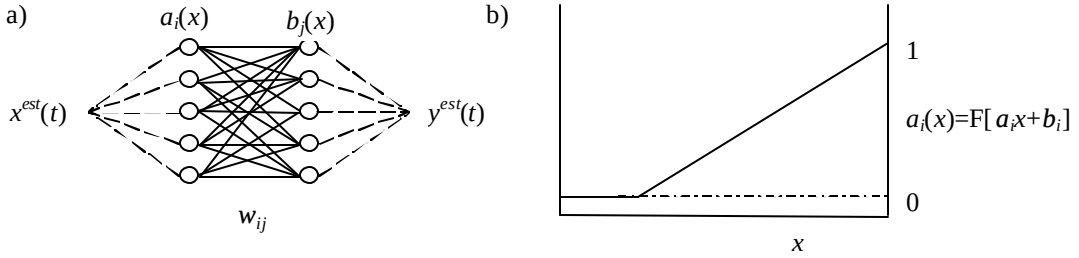


Figure 6: a) Communication channel in a neural ensemble where $x^{est}(t)$ and $y^{est}(t)$ are the population estimates for an analog quantity, $a_i(x)$ and $b_j(x)$ are the neural firing rates, and w_{ij} are the weights connecting the two populations; b) sample neuron response function $a_i(x)$ – a straight rectified line.

Though simple, this problem includes all the important ingredients of a more realistic model: 1) a nonlinear neuron transfer function (i.e., rectification, which is written in Figure 6 as $F[\ldots]$); 2) time dependent communication between two populations of neurons; and 3) a transformation between input and output.[7]

Working with firing rates, we can write the following expression for the firing rate of each neuron, j, in the receiving population:

$$b_j(x(t)) = F\left[a_j x(t) + b_j\right] \tag{8}$$

We can now substitute our expression for the decoding of $x(t)$ from equation (5), obtaining:

$$\begin{aligned} b_j(x(t)) &= F\left[a_j \sum_i a_i(x(t))k_i + b_j\right] \\ &= F\left[\sum_i a_j k_i a_i(x(t)) + b_j\right] \\ &= F\left[\sum_i w_{ij} a_i(x(t)) + b_j\right] \end{aligned} \tag{9}$$

The final expression here looks much like that for a standard connectionist network. As noted in the last section, however, we can further decode spike trains into neural firing rates by using equation (4), giving:

$$\sum_n h_j(t - t[n]) = F\left[\sum_i w_{ij} \sum_m h_i(t - t[m]) + b_j\right] \tag{10}$$

This gives us analytic means of calculating the weights, w_{ij}, to perform the desired transformation. Notice that in this expression, we need make no reference to the higher-order vehicles. That is, the transformations between higher-order vehicles are expressed solely in terms of basic, measurable quantities of our system. In other words, we have provided a reduction of a transformation expressed at the level of higher-order vehicles to one expressed at the level of basic vehicles. This is the real fruit of the previous work that relies on decoding neural encodings. Only if we can understand how to decode a signal can we precisely capture the transformations that a signal can undergo. Ultimately, it is these transformations that are important for understanding neurobiological systems. So, even though neurons themselves may not decode signals,[8] we must understand the decoding process in order to understand what the neurons are doing to the signals. Equation (10) doesn't express a means of decoding a neural signal; it only expresses a means of changing one set of spikes into another. This approach, then, is explicitly taking the perspective of the animal. This doesn't mean that the animal must extract the value, but rather that in order to characterize the kinds of processing (transformations) possible by a system using this encoding *we* must understand how to do so.

Of course, determining *which* transformations take place in a given neurobiological system is obviously an important task, but one for experimental neuroscience. What we need in order to understand transformations of representations in general is a means describing any given transformations in a neurobiologically reasonable fashion. Work from Anderson (1994; 1998), Hakimian and Anderson (1999) and Eliasmith and Anderson (Eliasmith and Anderson 1999; Eliasmith and Anderson forthcoming; Eliasmith and Anderson in press) details a

---

[7] As uninteresting as this transformation may seem, if, rather than connecting the first population to the second, we recurrently connect it to itself, we have constructed a form of memory – an extremely important function to realize (Eliasmith and Anderson 1999).

[8] However, it may also be the case that neurons, or more precisely dendritic trees, do decode spike trains in order to perform analog computations on them (Warland, Landolfa et al. 1992, p. 330).

means of doing just that. The general procedure is to take a neurobiological description of a system (e.g., cerebellum, eye position control circuit, lamprey locomotion system), mathematically describe its function, and determine the weights needed to reproduce that function. It is then possible to compare the properties of the model (e.g., connectivity and spiking patterns) to those found in the original neural system. If they match, the model is successful. If many models are successful, the framework seems to be a good one for understanding neurobiological systems. So far, the framework has been quite successful.

5.3 Extensions of the theory

As I have frequently noted, the examples I am presenting are limited in scope. I have only really addressed 'first steps' in getting from neural firings to interesting higher-order vehicles. However, there is reason to think that the methods I've introduced generalize quite well. The reason is that the process of expressing higher-order vehicles in terms of lower-order vehicles can be an iterative one.

Consider the following four equations by means of example:

1. $x(t) = \sum_n a_n(t) k_n$

2. $y(t) = \sin(\ x(t))$

3. $S(t) = \sum_n y_n(t) b_n$

4. $\mathbf{Z}(t) = \sum_n S_n(t) \mathbf{c}_n$

The first equation should look familiar; it is just like equation (5) in section 4. This then, is an expression for a higher-order vehicle, $x(t)$, in terms of neural firing rates (which can be expressed in terms of basic vehicles as discussed in section 5.1). The second equation in this list is a simple transformation between higher-order vehicles, which can be implemented as described in section 5.2. The third equation is an iterative application of the relation in the first equation *to* a higher-order vehicle. Thus, $S(t)$ is 'made up of' higher-order vehicles, $x(t)$. So, it is an even higher order, say third-order. In fact, it happens to be a lot like a Fourier series, which can be used to express *any* time-dependent signal. The process doesn't have to stop there. The fourth equation applies the same linear operation to the third-order vehicle, $S(t)$, to give $\mathbf{Z}(t)$, a fourth-order vehicle. We can think of $\mathbf{Z}(t)$ as a vector in an n-dimensional space that is a combination of time-dependant signals each describing different attributes of the input. This process can continue, but hopefully my point is made.

Using the encoding/decoding ideas, we can build up vehicles of any complexity we like. However, it is more likely that we will wish to go the other way around. Say, for example, we want to understand the neural representation of images. If we can express an image (or series of images) as a time-dependent signal of high-dimensionality (as we can, see e.g. (Rao and Ballard 1995)), we can then reduce this complex description to lower and lower order vehicles until we have expressions in terms of neural firings. Of course, it won't be an easy task to determine what high-dimensional expression best captures the kinds of representations used by neurobiological systems, but some progress is being made on the problem (Olshausen and Field 1996; Lewicki and Sejnowski 1998).

What is important for developing a theory of representation is that we *can* describe the kinds of representations used by neurobiological systems, no matter their complexity. This framework is flexible enough to do just that. We can accommodate vectors (e.g., a set of analog variables which represent horizontal and vertical eye position concurrently) and vector fields (e.g., sets of vectors which might specify color, intensity, etc. at some points in space). More importantly, we can generalize this formulation to represent functions over any of these kinds of representations (i.e., analog quantities, vectors, and vector fields). In place of the k_i weights, we can introduce weighting functions. These functions can be determined the same way we determined the k_i, that is, by forming and minimizing the analogous error function, or they can be learned from natural inputs (Lewicki and Olshausen in press). Being able to represent functions over variables introduces the ability to encode not just a value, but additional information such as the variance and uncertainty of our estimate.

Not only do we have to be able to describe the representations that are used by neurobiological systems, but also their transformations. Transformations between high-order representations will undoubtedly work together to give us the rich representations we, and other animals, use. In other words, it is equally important to be able to transform one space (e.g., that of light intensities and colors) into another (e.g., that of objects) as it is to be able to represent elements in those spaces. Transforming representations between spaces is important for at least two reasons. First, it can make further transformations quicker or easier given available computational resources. Second, representation in a given space will support the extraction of certain kinds of information better than others. For example, representing an image verbally may help for the classification of objects, while representing an image pictorially may help for the extraction of relations between objects.

Again, work in Anderson's lab has shown that many different kinds of transformations can be supported, including: simultaneous top-down and bottom-up inference (Eliasmith and Anderson forthcoming); general Bayesian inference (Barber 1999); recurrent transformations (Eliasmith and Anderson 1999); and transformations describable by differential and difference equations (Eliasmith and Anderson in press). Given these successes, it seems that this framework is a good one for using to tackle the problem of neurosemantics. In addition, it provides powerful tools for 'filling in the details' for a theory of representation. Now that we have those tools, it is time to see what a representational theory of content that employs them looks like.

6 Summary

I have provided the details necessary for understanding the three relations presented at the beginning of the chapter: the vehicle/world relation, the inter-vehicular relation, and the higher-/lower-order vehicle relation. Specifically, I have presented selective results from recent work in computational neuroscience to characterize these relations. I have shown how the vehicle/world relation can be characterized as an isomorphism between a computational description in terms of basic vehicles and a causal description of underlying neural processes. The inter-vehicular relation is captured by the computational formalism because it provides a way of understanding transformations at any level of description. The relation between higher- and lower-order vehicles can similarly be understood as a virtual encoding relation. That is, an encoding dependent on populations of lower-order vehicles that are taken to encode a higher-order vehicle. These characterizations provide all the ingredients needed for a theory of content. In the next chapter I show, more precisely, that this is the case.

CHAPTER 7

A Neurocomputational Theory of Content

Let no one mistake it for comedy, farcical though it may be in all its details. – H. L. Mencken, about the Scopes Monkey Trial

1 Introduction

In section 4 of chapter 5 I outlined a skeletal theory of content that I claimed needed details. I have just finished presenting the details that I think, excuse the gory turn of phrase, can flesh out this skeleton. As you will recall, the skeletal theory identified four theoretical objects: vehicles, referents, contents and systems. Vehicles are the internal physical things (or combinations of them) that we can poke, prod, and measure. Referents are the external physical things (or combinations of them) that we can also poke, prod, and measure. Contents, I claimed, are determined by causal relations and transformations. The system is the least problematic because it is determined by the problem we are trying to solve; the system is the neurobiological system.

As noted, my characterization of the theory in chapter 5 was an attempt to give just a flavor of the theory to come. As a result, a number of these claims are, strictly speaking, wrong. However, now that we have seen the details that are needed to precisely characterize the relations between these entities, it is also possible to better characterize the entities themselves. My strategy here is to first explain in detail what the relations between the objects are, taking into account both the theory of cause presented in section 3 of chapter 5, and the theory of neurobiological computation presented in the last chapter. I will then revisit vehicles, referents, and contents, and give them a more precise treatment. Finally, in section 4, I will concern myself with how this theory answers the representational questions proposed in chapter 2.

2 Relational details

At the beginning of the last chapter, I identified three relations that need to be explained by any theory of content: the vehicle/world relation; the inter-vehicular relations (at a single level); and the basic/higher-order vehicular relations (across levels).

At the basic level, i.e., the level of neural firings of individual neurons, the first two relations (vehicle/world and inter-vehicle) are closely linked. I assume that neuroscience has clearly demonstrated that neurons encode their input, be it from the world, or from other neurons, into voltage changes. In the case of periphery neurons, the encoding process generally begins with a non-electrical signal (e.g., photons, pressure), which is transduced and converted into a spike train. The encoding process is a causal one. A photon striking a retinal rod, for example, causes rhodopsin to partially separate into opsin and retinene, which causes cyclic GMP to be activated, which then causes sodium ion channels to close, which causes a depolarization in the cell. All of these causal relations can be identified by statistical dependencies and the relevant energy transfer. The relation between the world and the vehicles at this level of description, then, is a causal encoding one: fluctuations in light levels are encoded by neurons into voltage changes (and eventually spike trains). Given the discussion in chapter 5, the term 'causal encoding' makes sense because the causal and computational descriptions are isomorphic (also see below). Recall that the details of this relation depend directly on the statistical dependence hypothesis. That is, the decoders are determined by constructing the joint probability distribution between neural events and certain external causes (those that have the highest statistical dependency).

Staying at this level, there is no strict distinction to be made between this kind of transduction and the passing of neurotransmitters between cells. The way neurons generally influence one another is to send chemical signals, in the form of neurotransmitters, to each other. Just as photons induce a chemical reaction resulting in modulations of voltage levels, so neurotransmitters induce chemical reactions that generally modulate spike trains. It is the correlations between the presence of neurotransmitters, and their transfer of energy through dendritic trees to the neuron soma that help determine when a spike is fired and when it is not. So, we can give a similar causal encoding description inside the nervous system as well: fluctuations in neurotransmitter levels are encoded by neurons into neural spike trains. The relation between neurons or between neurons and the external environment *can* be characterized in exactly the same way. It doesn't matter if we are talking about 'external' causes or 'internal' ones: causes are causes, and causes are all there are at the basic level. *However* (and this is

a big however), this description doesn't have a direct relation to the *content* of these internal neurons. Describing the relations between neurons in this way just shows that computational/causal isomorphism holds everywhere (as we would hope). In order to understand *content* we need to characterize these relations as transformations; it is, after all, these transformations that set up the statistical dependencies relevant to content determination.

How, then, do we characterize *transformations* (i.e., inter-vehicular relations) between basic vehicles? Recall that characterizing a neuron as encoding its input signal is dependent on the possibility of decoding the output to regain that input signal. Suppose, now that we have a peripheral retinal ganglion neuron that encodes photon impact rates in an area of the visual field. Suppose that a connected neuron is 'interested in' occasions when the impact rate goes above some threshold. The output of this second neuron would transform the output of the first into 'above' or 'below' the threshold. There is a rule (i.e., a transformation) that relates the photon impact rate with this categorization. How are we to understand this kind of relation between basic vehicles? We can characterize the transformation as a kind of *biased decoding*. That is, the second neuron decodes the photon impact rates as a signal of '1' when the rate is above the threshold and '0' otherwise. It is a bias because it isn't decoding the input signal, but rather a *function of* the input signal. This sort of bias is implemented in neurobiology by the connection strengths between neighboring neurons. Given the computational characterization presented in the last chapter, we can determine what these connection strengths must be to implement a given transformation (see section 5.2). These transformations help determine how the behavior of neurons relates to the external world.

The vehicle/world and inter-vehicular relations can be similarly characterized for higher-order vehicles. In particular, the vehicle/world relation is a causal encoding dependent on the highest statistical dependency and the inter-vehicular relations are transformations described at the level of the higher-order vehicles. But, there is an important difference as we move to higher-order vehicle. Higher-order vehicles are, in some sense, up to us. We posit variables like 'horizontal eye position' and try to provide good explanations of the behavior we see using these kinds of vehicles in our explanations. We will have to justify choosing a certain set of neurons that make up this vehicle. We might have physiological reasons, functional reasons, and pragmatic explanatory reasons. The more these reasons converge, the more likely our theoretical posits are real.

A consequence of the comparatively theoretical nature of higher-order vehicles is a *seemingly* more complex theory of content. In discussing content ascriptions for basic vehicles, I concentrated on their relation to the external environment. This clearly leaves out one of the two factors I suggested would be important in chapter 2; something like conceptual role. However, when we consider the content of higher-order vehicles, it becomes clear that our intuitive ascriptions of content depend on the transformations such vehicles enter into. For example, we presume that the vehicle 'horizontal eye position' has the content it does because it is the result of transforming (in this case integrating) the vehicle 'horizontal eye velocity' and/or because it is decoded by the muscles into an actual horizontal eye position. The positing of vehicles depends on their relations with other vehicles.

If there were no other high-order vehicles, it would make little sense to identify 'horizontal eye position' (or 'edge at location (x,y)'), because such an identification would need to be justified. Any such justification would identify computational relations because it would be necessary to answer relevant why-question: e.g., why does that vehicle carry horizontal eye position contents? In answering such questions, we have two choices: 1) we can say 'because it's decoded into horizontal eye position' i.e., identify the computational relation to states of the world; or 2) we can say 'because it integrates horizontal velocity' i.e., identify a computational relation to another higher-order vehicle. The most complete answer would be one that provides both characterizations. But, in any case, the question can't be answered without specifying how the vehicle is *used*. This means that these theoretical posits are (at least partially) identifiable by their computational/transformational relations. Having relations determine object identity really shouldn't be surprising, as it happens all the time in science (e.g., centers of mass, electrons, black holes).

Such considerations don't mean, of course, that the relation between higher-order vehicles and the external world isn't causal; it is. As important as the transformational relations are, the precise identity of higher-order vehicles can't be pinned down without reference to what they cause, or what causes them. 'Edge detectors' are so-called (erroneously, perhaps) because they are thought to be normally causally related to edges. The content 'horizontal eye position' is carried by a horizontal eye position vehicle because it is normally causally related to horizontal eye positions. What such considerations *do* mean is that we should be more careful about how we characterize the content of *basic* vehicles. In particular, we now have reasons for thinking that statistical/causal relations aren't all there is to content determination. How the basic vehicles are used, that is, the transformations they give rise to, are also relevant. Edge detectors (or, better yet, orientation filters) are, after all,

basic vehicles (Felleman and Van Essen 1991; Kandel, Schwartz et al. 1991; Callaway 1998; see section 3.3 for further discussion).

So far, I have shown that the vehicle/world and inter-vehicular relations can be characterized in the same way at different levels of organization. Along the way, I have noted that not only can they be characterized in the same way, but that they seem to be similarly relevant to content ascriptions at the different levels. I will discuss the role of these relations in content ascription in more detail in section 3.3. Recall that the third relation that the neuroscientific details can help characterize is that *between* basic and higher-order vehicles.

The importance of the basic/higher-order relation shouldn't be underestimated; if we couldn't abstract over basic vehicles, we would be in the position of having to describe neurobiological systems only in terms of the functioning of individual neurons. However, there are numerous reasons that such descriptions are not likely to be adequate. First, such a description would provide no indication of what relation neuroscientific theories have to psychological ones. Second, such descriptions of interesting cognitive phenomena are likely to be far too complex to be satisfactory *explanations*. Third, neuron level descriptions are so wedded to implementation as to not suggest a means of abstracting the theory to understand non-biological representation. Fourth, we may be able to extract important cognitive laws at higher-levels of description that simply aren't evident at lower levels. In general, any reasons that have been offered for wanting higher-level descriptions of a natural system, be they epistemological or metaphysical, are reasons for needing a good description of the basic/higher-order vehicle relation in this theory (Fodor 1975; Wimsatt 1980; Bechtel 1986; Bechtel and Richardson 1993). Thus, just because there is currently less agreement on higher-order vehicles than on basic ones, and just because this framework provides a means of reducing the former to the terms of the latter, doesn't mean higher-order vehicles are, in any obvious sense, dispensable. Once we have a good understanding of the relations between higher- and lower-order vehicles, we can know precisely in what ways these higher-order activities approximate or mirror underlying neural function. Talking of images and objects may make it easier to explain certain avoidance behaviors than talking of millions of neural spike trains would, even though a basic-level description may produce better predictions. Higher-order vehicles are, at the very least, epistemically unavoidable posits.[1]

There is an important difference, then, between the two relations I have already discussed, and the relation between basic and higher-order vehicles; the latter is a relation between levels of organization. A precise formulation of this relation is given in equation (6) in chapter 6. This equation says that we can talk about sets of neurons as encoding analog quantities with certain levels of precision (due to noise effects). As I noted in section 5.3 of the last chapter, this relation is recursive, meaning that it also tells us how to talk about analog quantities as signals, vectors, vector fields, images, etc. Each of these representational levels is related to lower levels by a weighted linear combination. Surprisingly, then, higher-order vehicles *really are* a sum of their parts. However, depending what parts you sum, you will get very different kinds of transformations that apply.

Complex behaviors evident at higher levels may not be evident at lower levels because of the particular kinds of organization evident at different levels. Transformations evident at one level likely won't be evident at other levels. But, more than this, transformations that can be performed at one level might depend for their performance on a particular organization of components. This is reflected in the kinds of descriptions we use. It makes sense to talk of integrating 'eye velocity' commands to determine eye positions. It doesn't make sense to talk of integrating 'eye muscle rate change' commands to determine eye positions because the same 'eye muscle rate change' command can be present at many different eye positions. If we are interested in explaining *eye positions* only the former description will be adequate, despite the fact that an 'eye velocity' command is a sum of 'eye muscle rate change' commands. These same kinds of considerations hold at even higher levels of abstraction, like those that describe our representing a dog.

3 Details for objects

Now that I have given a better sense of the nature of the relations between objects in the theory of content I defend (in the next chapter), it is useful to revisit the objects taking part in those relations. These objects include vehicles, referents, contents, and systems. I take it, as I did in chapter 5, that the last object, 'systems', is uncontentiously defined by the problem of neurosemantics: the system is the nervous system. Each of the other objects is the subject of one of the following subsections.

---

[1] More likely, higher-order vehicles are *real*, in the same sense that water is real even though we understand that it is composed of parts. However, metaphysical claims of this sort aren't particularly important for this theory of content.

3.1 Vehicles

Vehicles are the internal physical objects, or 'representations', that carry representational contents. Basic vehicles are neurons, as functional units. Higher-order vehicles are sets of neurons. The theory I have presented does not place a lot of constraints on the nature of higher-order vehicles. I think that this is a good thing, since we quite clearly have some way to go before we know what the vehicles are in neurobiological systems. The vehicles that give us a good (or the best) explanation of such a system's behavior will be the right vehicles. It is not my purpose to say what the vehicles are, but rather to provide a framework for ascribing content to the vehicles, *whatever they may be*. Determining the right vehicles is a matter for psychology and neuroscience, and we won't know *all* the right vehicles until we can provide *all* the explanations we want.

One example of an on-going debate about vehicles is the descriptionist/pictorialist debate in psychology (Pylyshyn 1973; Kosslyn 1994). Descriptionists think the vehicles of mental imagery have the property of being discrete. Pictorialists think the same vehicles are continuous. If we can specify vehicles that have one of the properties and not the other, the debate should be resolved.[2] Similarly, in neuroscience there is a debate about whether neurons in visual area V1 are 'edge detectors' or 'orientation filters' (see Van Essen and Gallant 1994). A theory of representation should not resolve these debates by fiat, but should be able to accommodate the best explanation, whatever it may be.

This second example raises related concerns about the nature of vehicles. Notice that the terms 'edge detectors' and 'orientation filters' are 1) descriptions of single neurons and 2) picking out vehicles based on their contents. More subtly, 3) use of such terms also assumes certain background theory. I will come back to this last point shortly. First, consider 1) and 2).

The fact that terms like 'edge detector' are descriptions of the content of single neurons has been noted already. Since there are debates over such terminology, it's clear that the identities of basic vehicles really aren't uncontentious after all. Rather, neurons are uncontentiously basic vehicles only as functional units, not as carriers of content. That is, everyone (more or less) agrees that single neurons act in certain ways and are a basic functional unit underlying cognition. However, everyone *doesn't* agree on what the behavior of single neurons *means*. This leads us to 2).

The reason that the identity of vehicles isn't easy to settle is because vehicles are generally named after the contents they carry. Edge detectors are so-called because they are thought to be activated by edges and used to detect edges. Orientation filters are so-called because they are thought to be activated by spatial orientations and used to analyze orientation gradients in visual images. In some ways this seems unnecessarily confusing, but it just goes to show how important content determination actually is to theories in *neuroscience* as well as philosophy. This should also make it clear why drawing a sharp distinction between basic and higher-order vehicles is an oversimplification. Both are individuated in terms of the theory in which they play a part. So, although the physical objects that we call vehicles (both basic and higher-order) can be uncontentiously picked out by our physical theories, *qua vehicles* they can only be picked out by our theories of content. And, our theory of content is still pretty much up for grabs, unlike much of our physical theory. However, we can *use* this well-established physical theory to support and establish a good theory of content. This is one way of seeing what the details in the last chapter are about. Theories of content shouldn't be disconnected from our physical theory. This leads us to 3).

In a footnote in chapter 4 I noted that making claims about a system representing 'velocities' instead of 'nearby flashes' raises a deep philosophical worry. The worry is that we are simply mandating what is being represented (i.e., velocities) in an unprincipled way, since we can't *really* tell the difference between velocities and nearby flashes. Given the considerations in the previous paragraph, it should now be clear why I don't think this is a serious worry. Following Quine (1960; 1969; 1981), I take it that our theories about the world are not disconnected sets of propositions. Terms in one theory depend for their meaning on terms in others. In this case, 'velocities' are the best kinds of things to quantify over because *so many other of our successful theories do so*. A theory of content, in other words, has a right, if not an obligation, to connect with our other successful theories. This means a theory of content can and should quantify over the kinds of things other theories do. This will serve

---

[2] Keeping in mind the subtleties of claims concerning continuity and discreteness (see Eliasmith in press). Note also that this presents a way of deciding between the two possibilities despite Anderson's (1978) having shown that descriptions and analog representations are equivalent. This is because Anderson doesn't consider implementational issues when making his argument.

to minimize our overall ontology and give us a good theory of the world, and thus a good theory of content. Furthermore, realizing the importance of adopting current terminology and assumptions in constructing new theories goes a long way to giving a clearer understanding of the problem of misrepresentation. I will discuss this last point in more detail in the next chapter.

3.2 Referents

In chapter 4 I suggested that we individuate referents by using the statistical dependence hypothesis. However, the hypothesis, as it stands, is inadequate. The purpose of this section is to be more precise as to how the referent of a vehicle is determined.

> To begin, recall that the statistical dependency hypothesis is:
>
> *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.*

This hypothesis makes two important assertions: 1) the highest statistical dependency picks out referents; and 2) referents are *causes*. Given the theory of cause I outlined in chapter 5, the second part of the hypothesis has a specific interpretation: referents have energy transfer with the relevant vehicles. This is extremely important. If we solely depended on statistical dependencies for determining referents of a representation, we would be forced to make many odd referent ascriptions. For example, if you and I were both viewing a cat, and perhaps representing (or misrepresenting) it in exactly the same way, it could well be the case that the highest statistical dependency between our internal states were with *each other's* neural states, not the cat. In this case we would have to say that the referent of our representations were each other's neural states, not the cat. However, given that there is no causal relation (i.e., energy transfer) between our neural states, we can safely say that the referent of such representations is the cat. In general, statistical dependencies are too weak to properly underwrite a theory of content. Cause, then, is central to referent determination.

Despite incorporating causes, the statistical dependence hypothesis is still inadequate. In particular, the hypothesis will result in a kind of solipsism. This is because the highest dependency of any given vehicle is probably with another vehicle that transfers energy to it, not with something in the external world. How can we solve this problem? Before suggesting a solution, I would like to consider an example that makes the problem clearer and motivates the solution. As well, the example highlights an important strength of the hypothesis.

Consider, then, the example of neural emulators. In particular, it has become clear that nervous systems use 'emulation' strategies to aid precise motor control (Grush 1997). These strategies result in one part of the nervous system emulating the expected response of another part of the nervous system in order to facilitate quick changes to a motor plan. For example, if I reach for a pen sitting in front of me, the motor command sent to my arm is also sent to an emulator. It is the emulator's job to predict the proprioceptive feedback that will result from the motor command. Because the emulator doesn't need to wait for the kinematics of my arm to take effect, it will generate a result sooner than *actual* proprioceptive feedback is available. These results can then be used to begin appropriately slowing down my arm before pressure sensitive neurons tell me I have touched the pen. Emulation thus allows for quick corrections to the motor program if necessary. How would we determine the referent in the emulator in this case? Notice that in such a case the highest statistical dependency for the emulator will be with the proprioceptive feedback. But, proprioceptive feedback is an internal state and we are trying to figure out how we can avoid having internal states as referents. Perhaps we can avoid this by noting that the proprioceptive feedback comes *after* the emulation, and thus rule out proprioceptive feedback as the referent on this basis.

However, things aren't so easy. Notably, the statistical dependence hypothesis *doesn't* mandate that a referent precedes the representation. In fact, this is an important *strength* of the hypothesis. Mandik (1999) convincingly argues that a general flaw of teleo-informational accounts of content (like those of Dretske, Fodor, and Millikan) is precisely that they can't account for this kind of representation relation. However, there is nothing in the account presented here that demands either 'forward' or 'backward' representation; both are acceptable. In fact, the 'horizontal eye position' vehicle is precisely of this 'backward' nature. The neural state has horizontal eye position as its referent because that is horizontal eye positions to which it is most highly statistically correlated and *to* which it transfers energy.

So, the emulator still seems to have the proprioceptive state as its referent because of the high dependence and the energy flow from the emulator to the proprioceptive state. Given the considerations in the previous paragraph, we can't rule out the proprioceptive state as a referent merely because it succeeds the emulator. However, there *is* something odd about calling the proprioceptive state the referent of the emulator's representation. Namely, the proprioceptive/emulator relation is a causal relation that *falls under our computational description*. But, we can't have our computational relations determining the referents of our vehicles; this is precisely why conceptual role theories can't account for truth conditions. We need to ensure, then, that referents don't fall under computational descriptions. Given this additional constraint, the referent of the emulator would be the arm position, not the proprioceptive feedback. Arm position doesn't fall under our computational description and so it is a candidate for referent status.

Consider another example. How do we know that the referent of the horizontal eye position vehicle is the eye position, and not the velocity command that precedes it? If all we had to go on was energy transfer and statistical dependence, we could not distinguish between the velocity command and the actual eye position as being the referent of the horizontal eye position vehicle. In general it seems that whenever we attempted to determine referents, we would end up with referents that were always *nearby* their vehicles. This kind of referent relation would not be able to support misrepresentation as outlined in chapter 5. In the eye position case, as in the case of the emulator, the referent is the eye position and not the velocity command because the vehicle's relation to the velocity command is a computationally specified one. These kinds of considerations introduce a new constraint on referent determination: referents don't fall under the computational description. Taking this new constraint into account, referents satisfy 3 constraints: 1) they have the highest statistical dependency with the vehicle; 2) they either transfer energy to or from vehicles; and 3) they do not fall under computational descriptions.

This additional constraint raises the question: does the third constraint guarantee that the referents of vehicles are external? If candidate referents can't fall under computational descriptions, and all internal events (i.e., events in the nervous system) do fall under computational descriptions don't all referents have to be external? Surprisingly, perhaps, the answer is 'no'. Referents can be internal as long as the computational links between the referent and the vehicle are broken. For example, suppose I am looking at a real-time brain scan of my own brain, and then think a thought about one of the neural states I see. Suppose also that the highest statistical dependence is between my thought and the neural state. Clearly, the neural state is transferring energy to this internal vehicle. In such a case, however, there is no internal computational description that relates the neural state to my thought that can account for the statistical dependence, so the internal state *is* the referent of the vehicle. More accurately, then, the third constraint should read: referents do not fall under computational descriptions that account for the statistical dependency.

3.3 Content

In chapter 5 I noted that referents and contents may or may not be distinguished on a theory of content (recall that Fodor (1998) and Dretske (1995, p. 30) don't, although Block (1986) and Cummins (1996) do). I think that contents do, in fact, need to be distinguished from referents. The reasons are essentially those given for motivations of a conceptual role theory (see chapter 2): 1) non-referent based meaning is needed to explain behavior; 2) non-referent based meaning is needed to satisfactorily handle Frege cases; and 3) transformations are relevant to understanding content.

The first of the reasons was discussed in some detail in chapter 2. In particular, I noted there that Twin Earth cases show exactly how referents can't serve to explain behavior. The referents on earth and Twin Earth are, *ex hypothesi*, different yet the twin's behaviors are the same. If contents and referents are equated, then we have to ask: how can *differences* in meaning explain the *sameness* of behavior? There is no aspect of meaning that can be appealed to in a "content=referent" theory (as in any causal theory) that remains the same in Twin Earth cases.

The second reason (i.e., that non-referent based meaning is necessary to explaining Frege cases) has been denied by causal theorists like Fodor (1998). Fodor, in fact, stipulates that all co-referential terms are synonymous: "I'm assuming that coreferential representations are *ipso facto* synonyms" (p. 15). There is, according to Fodor, *no difference in meaning* for any two co-referential terms. This is indeed a strange claim. As unsophisticated as dictionaries may be about meaning, if there is any sense in which they elucidate the meanings of terms, we can see why this claim is strange just by looking at one. Pick your favorite dictionary (or encyclopedia), look up 'Hesperus' and 'Phosphorus' (or 'morning star' and 'evening star') and you will see that they *mean different*

things: one is seen at sunset, the other is seen at sunrise.[3] In fact, they aren't synonymous. Fodor, then, has a unique understanding of meaning if he takes it that coreferential terms are synonymous. Worse, it is a deficient understanding of meaning because it clearly can't account for the pervasive intuition that coreferential terms aren't synonyms. Although I don't want to be a meaning rationalist (i.e., hold that we have special, introspective access to the nature of meaning; see Millikan 1993), and as slippery as the meaning of 'meaning' may be, a theory that comes closer to explaining the pre-theoretic notion of content is going to be a better (more complete) theory than one that doesn't.

Lastly, recall that the third reason for thinking that contents are different from referents is that transformations have, literally for ages, seemed to be relevant to determining contents. On the current theory, *referents* are determined by transformations, so perhaps this could be explained by a theory that equated contents and referents. However, there are *more* transformations than just those that determine the statistical dependencies that underwrite referent determination. Consider, for example, a neural state that is tokened whenever there is blue in the visual field. But suppose that this neural state is never *used* to pick out blue things. Certainly, as external observers, *we* are in a position to use the state to know when there is something blue in the visual field, but it would be odd to say that this neural state *means* 'blue in the visual field' *to the system in which the state is tokened*. This 'blue' state is the kind of state Dretske (1988) calls a natural sign and that Grice (1957) says has a natural sense. Dretske, following Grice, claims that such states don't "mean" in the same sense that languages and thoughts "mean".[4] Natural signs, for example, can't mean something other that what is actually the case; they can't be wrong (p. 55). This is because such states aren't necessarily taken to be about anything. We can see, then, that the problem of neurosemantics is really a problem of what neural states mean *to nervous systems*. In order to solve *this* problem, we have to take meaning to be (at least partially) determined by how neural states are used. Such considerations are reminiscent of 1) above. If a neural state can't possibly affect the behavior of the system it is in, how can it be said to *mean* anything? Meanings, after all, help to explain behaviors.

These, then, are reasons for thinking that contents aren't the same as referents. So, what *is* content? Or, more specifically, what is the part of content not captured by the referent of a representation? In order to answer this question, I have to say what those transformations are that are left out of referent determination. My suggestion is this:

> *The set of relevant transformations is the ordered set of all those causally related transformations that succeed (or precede) the vehicle.*

In other words, the transformations that can be caused by the vehicle capture how it is used and thus what it means. The ordering of the set is determined by the likelihood that the transformation is effected over all stimulus conditions. The reason the set is ordered is that the more common the transformation, the more relevant it is likely to be to what we intuitively call meaning. The most common transformations thus help determine something like the "core" meaning of the vehicle (Smith 1989).

The picture of content determination that I have provided can be summarized as follows: For any vehicle, both the transformations that precede it and succeed it are relevant to content determination. One of the antecedent or consequent set will determine the highest statistical dependency that the vehicle holds with a non-computationally described, causally related event or object in the environment (i.e., the referent). The other set determines the relation the vehicle has with other vehicles. The entire set of transformations can be described in terms of a computational factor. The computational factor can be identified with a causal description because an isomorphism holds between the computational description and neurons as functional units.

A number of comments are in order. First, transformations can be identified at the various levels of description. Only some of these may capture what we intuitively take to be the transformations relevant to meaning (e.g., inferences). Perhaps at the level of 'dog' vehicles we can find one or a few transformations that results in the ascription of the property 'has four legs'. But quite likely at lower levels no reasonably *small* set of transformations could account for this inference.

---

[3] I've checked Webster's, the Oxford English Dictionary, and Encyclopedia Britannica.

[4] In fact, Dretske is careful to use a term other than 'meaning' or 'sense'. He prefers to say that such states 'indicate' (p. 55).

Second, the story of content determination I have provided here is applicable to both conceptual and occurent content. I have concentrated on the former, but the latter can be determined in the same manner, substituting 'current stimulus conditions' for 'all stimulus conditions' (see chapter 8 for further discussion).

Third, the whole set of transformations relevant to content determination is likely to be very large for any given vehicle, no matter the level of description. This inherent complexity of content determination may explain some of the conflicting intuitions about the meaning of terms. For instance, looking at only half of the transformations as relevant, as both causal theorists and conceptual role theorists do, greatly simplifies the problem. The examples each tend to chose further simplifies the set of transformations. Causal theorists look to perceptual representations, where there are fewer transformations underlying the nomic relations, making external causes seem rather immediate. Conceptual role theorists look to abstract representations, decidedly removed from causes and intuitively definable in terms of only the most common inferences drawn on the basis of those representations. Both stories are accused of not accounting for the insights of the other. My story, though perhaps more complex, can satisfactorily account for both sets of insights, in a *unified* way.

3.4 A detailed example

In this section I present a sample of the kind of account of content ascription this theory provides. I would *like* to present an account that can equally address abstract and sensory representations. Unfortunately, the neuroscientific details that are available are most complete near the sensory periphery. So, the story I tell focuses on an instance of sensory representation. Hopefully, I tell enough of a story to convince you that there is a similar story to tell for higher cognitive areas as well.

In fact, things are worse, we don't really know much about how brains work in general. However, we do know a fair amount (relatively) about the visual system. Even so, I will be leaving out more details than I put in. I won't explicitly use the equations I have outlined (see Eliasmith and Anderson (forthcoming) for that degree of detail), and I won't be too concerned about missing some relevant neurophysiological facts in the story I'll tell. I'm most interested in giving a sense of the sorts of content ascriptions that come out of this theory and what justifies those ascriptions.

That being said, consider someone watching a dog run through a field. Light reflected from various physical objects in this scene over a period of time enters the eye and stimulates the retinal cones. The cones systematically encode the occurrence of photons into voltage changes. The content encoded by these voltage changes is something like 'photon impact rate'. The number of photons striking a cone cell determines how it will respond, just as for the cricket mechano receptor. Both statistical dependences and energy transfer are straightforward in this, basic level, case. Even here we could call these neurons 'photon detectors,' bringing in background assumptions and clarifying how this information is used by the nervous system.

Equally important, are the kinds of claims we would not make at this stage. For example, we wouldn't call any set of cone cells a 'dog detector'. In some sense, they do detect the presence of the dog; without cones the dog would not be visually detected. However, positing such a higher-order vehicle would fail under this theory for at least two reasons. First, the referent of such a vehicle would not be dogs, so the content won't be dogs, so we can't call the vehicle a 'dog' vehicle. The referent wouldn't be dogs because, although dogs don't fall under a computational description, and dogs do transfer energy to the cones, this vehicle would not be most highly correlated with dogs. In fact, it isn't particularly correlated with anything, just with photons hitting the retina.[5]

Second, there are no other theories that could support the identification of the cones with 'dog detectors'. Calling the cones 'dog detectors' would be fine if it cohered with a whole story of neural function, but it doesn't. Cones aren't used by the nervous system (just) to detect dogs. And, insofar as they are used to detect dogs, it's not because they have a special relation to dogs, i.e., they don't have dogs as referents. For these reasons, this vehicle would not prove explanatorily satisfactory, and is a bad candidate for a higher-order vehicle.

Back to our dog in the field. Once the photon rates have been encoded into spike trains by the retinal ganglion cells, the information is transmitted in two main streams to the lateral geniculate nucleus (LGN) in the thalamus. These two streams of information are called the magno (M) and parvo (P) streams. The M stream

---

[5] The notion of correlations and probability I'm using is a realist one. So, showing a retina only dogs, still won't make the correlations between such a vehicles and dogs the highest one.

carries high frequency information at low spatial resolutions and the P stream carries lower frequency information at higher spatial resolutions, although there is some overlap (Van Essen and Gallant 1994). From the thalamus, both streams are projected to visual cortex area V1.

Before I discuss the major connections to V1, there is a secondary pathway of interest. From the retina, a subset of the retinal ganglion cells also projects to the superior colliculus, a small area of the midbrain. This area is known to be important for visual orienting behavior. Furthermore, it projects directly to preoculomotor neurons in the brain stem (Everling, Dorris et al. 1999). These neurons are just those kinds that encode things like horizontal eye position (Moschovakis, Scudder et al. 1996). This secondary pathway provides a reasonably short route from input to output and helps show the kinds of content ascriptions this theory would make along the way.

In the superior colliculus (SC), the retinal ganglion cells provide a low spatial resolution map of the visual field that is updated very frequently. The referent of this high-order vehicle (i.e., the map) is average photon impact rates over large areas of the visual field. We can't say that the map picks out objects, because objects *or* anything else that causes the photon impact rates to vary (e.g., lights shone on the retina) will be picked out by the SC representation. The SC isn't particularly sensitive to objects. In any case, this kind of representation is the right kind for making fast decisions about where to look given gross properties of the visual field. If something reasonably large is looming in one part of the field, or quickly traversing a part of the field, it will be salient in such a representation. Notably, this representation is in retinal coordinates, meaning that the map is centered on the center of the retina. So, if a particular area of the representation 'lights up', to indicate a rapid change in some part of the map, the direction of eye movement can be determined by taking the difference (a simple transformation) between this position and the center of the map. Wherever the map lights up is where the eyes should move next to get a sharper, foveal picture of what's happening there. The superior colliculus thus generates a desired eye position command as a displacement from the current retinal position. That, then, is the central way in which the map is *used*. So, the content of the map is something like 'photon densities indicating the next eye position'.

Like most retinotopic maps, this one can be defined as a scalar or vector field. Thus, neurons represent the value of some set of variables (just contrast in this case) at a particular location. This, then, is a description of a lower level of organization of the same set of neurons. *Qua* map, a set of neurons have the whole retina as a referent, but *qua* set of vector representations, various subsets have limited parts of the visual field as a referent. Notice also that a particular set of neurons represents contrast at a given retinal location even though the location itself isn't represented explicitly. Rather, location plays a role in referent identification because of the organization of the system. Without taking this organization into account, the referent couldn't be correctly identified. In fact, there are computational advantages to retaining the nearness of representations of near retinal neighbors. In particular, it saves having to encode and decode a representation of the position of a variable whose retinal location is relevant to its representational content. Likewise, it simplifies wiring circuits to implement certain kinds of transformations; again, those for which retinal location is relevant. The kinds of transformations that are supported, for example, are those that help support contiguity of spatial representations. These considerations highlight two important points. First, at this level of organization the higher-level transformations (e.g., the 'difference' transformation) aren't always obvious, although other kinds of transformations are. Second, referent determination doesn't necessarily depend on what is *encoded* by a given representation.

To continue, commands from the SC eventually reach the neural integrator as a velocity command (e.g., move the eye in such and such a direction at 500 degrees per second) that lasts for a certain length of time (e.g., 20ms). This is an eye velocity command because it has highest statistical dependence with eye *velocities*. The neural integrator integrates this command, resulting in a particular eye position that it stores and sends on to the ocular motor neurons that connect directly to the ocular muscles. The combination of these muscles and the kinematics of eye itself acts to decode the represented eye position into an actual eye position.[6] The cells in the neural integrator can thus be taken as a higher-order vehicle that carries content about eye positions. In fact, this vehicle has a fairly high statistical dependency with some set of moving objects, but it has a higher one with eye positions because there are fewer confounds with eye positions than with moving objects (given that 'objects' and 'eye positions' are the kinds of theoretical entities our theories talk about). Given the simplified story I've told so far, that is, assuming that the SC pathway was the only one to eye positions, the content of the neural integrator would be something like 'eye position to salient stimulus'. However, eye positions can be determined in lots of

---

[6] As I warned, this description does not even begin to do justice to the complex circuits involved in the saccade system. For an in-depth treatment see Moschovakis et al. (1996).

ways (not just by salient stimuli) and, since we have no way of picking out those different ways uniquely (i.e., with one theoretical term) we say the integrator just carries the content 'eye position', leaving such complexities out of the picture.

The transformation from this kind of map to a desired eye position is relatively simple, but still instructive. It shows just how invisible the dividing line between 'forward' and 'backward' representation really is. Even though the details of these transformations may be buried deep in the brain in many cases, there is no need for a special kind of intervention for turning input into output. And, there is no drastic difference between one and the other. Indeed, we could have described the entire process as causal or as a result of certain transformations. However, choosing one or the other soon conflicts with background considerations and principles of good theory building (e.g., simplicity, coherence, etc.). Rather, the best explanations rely on both cause and transformation, and both causes and transformations can be understood in terms of the theory of representational content I have been developing.

Returning to the primary projections to V1 from the LGN in the thalamus, it quickly becomes clear that much less is certain about the nature of the representations and transformations these representations undergo. However, there is much that *is* known (see e.g. Felleman and Van Essen 1991; Kandel, Schwartz et al. 1991; Callaway 1998). The M stream synapses in the upper half of layer 4 of V1 while the P stream projects to the lower half. These layers both project to more superficial layers of V1 cortex that are sensitive to such things as color, edge orientation, motion, and microfeatures such as 'T' junctions (Das and Gilbert 1999). The motion sensitive parts of these layers project to a number of areas including V2, V3, V3A and V5. Area V2 is divided into thin, thick, and inter- stripes. These regions are most sensitive to color, binocular disparity, and form respectively, and project to areas V4, V3 and V5, and V4 respectively (DeYoe and Van Essen 1988; Nicholls, Martin et al. 1992). Areas V3, V3A, and V5 are part of what has been dubbed the 'where' pathway that processes spatial information. Whereas area V4 is the beginning of the 'what' pathway that processes information about form (Ungerleider and Mishkin 1982).[7] V4 projects to area PIT (posterior inferotemporal cortex) which projects to CIT and AIT (central and anterior inferotemporal cortex) (Van Essen and Gallant 1994). These two areas project to area TE that then projects to the hippocampus, amygdala, and areas 28 and 36 of temporal cortex (Saleem and Tanaka 1996). In almost every case, these projections are reciprocal.

Much of the work on cortical projections has been done in monkeys, and raises some worries as to their relevance to humans (DeYoe, Carman et al. 1996). However, functional magnetic resonance imaging (fMRI) and single cell studies in humans support these general projections, and also provide further insight into how language areas may be connected to areas of IT. Ojemann and Schoenfield-McNeill (1999) have shown that middle areas of temporal cortex on both sides of the brain are active in object naming tasks and not in similar non-naming tasks. These same areas are just before lateralized areas that have been implicated heavily in language processing (Binder, Frost et al. 1997). These areas lie near the homologous areas 28 and 36 in monkeys, which has been implicated in recognition memory of objects (Saleem and Tanaka 1996). In addition, areas of AIT just preceding TE, which projects to 28 and 36, have been shown to have view independent neurons (Booth and Rolls 1998).

These cortical connections and their functional properties help give a sense of the kinds of transformations and the kinds of representations that seem to be present in processing a visual scene, such as the dog in the field. The theory I have outlined provides a method for determining contents, referents, and vehicles that provide good explanations of behaviors and mental meaning. Even though some of the transformations, referents, and vehicles may be empirically disproven, how we can use the theory I have been presenting should be coming clearer. In particular, I have been showing how we can relate a theory of mental representation to neurobiological details to generate constrained explanations of behavior and meaning. In other words, by considering these sorts of details, we can posit likely high-order vehicles that can be used to explain both cortical processing and neural meaning. Let me continue with this example.

Consider area V1 at the beginning of the cortical visual processing stream. There are a number of guesses we can make about possible higher-order vehicles. One guess is that V1 consists of many vehicles that are have graded sensitivity to stimuli features in that they are single neurons and have statistical dependencies with various external features (e.g., orientation filters, color filters, motion filters). A second guess is that multiple neurons form higher-order vehicles that encode a more accurate estimate of some feature (e.g., edge orientation,

---

[7] The division into these two pathways is clearly an over-simplification (Felleman and Van Essen 1991; Goodale and Milner 1992; Van Essen and Gallant 1994), but one that I adopt to simplify my discussion.

color value, motion direction, etc.), just as multiple neurons encode horizontal eye position. There is evidence that near neighbor neurons tend to have similar response properties, which would support this kind of vehicle. Such guesses can be easily generalized to neurons that are tuned along multiple dimensions (e.g., motion *and* orientation). In these cases, single neurons and populations of neurons would encode vectors along both dimensions. A third guess is that even higher-order vehicles (i.e., larger areas of V1 cortex) encode all of the dimensions simultaneously. In fact, it seems that every square millimeter or so of V1 might be a complete representation of a small area of the visual field, i.e., a representation that describes the visual information along all relevant dimensions. A fourth guess would be that an even higher-order vehicle encodes a vector field. That is, for every point in the retinal image, there is a set of cells that describe each visual dimension. Together, these points form a map (a more detailed version of the superior colliculus map) of the retinal image incorporating all information that can be used in further processing.

Notice that despite the fact that each guess is at an increasingly higher-order of representation, they are not mutually exclusive. Undoubtedly, there are other guesses that *do* conflict with these. Nevertheless, this example demonstrates how different levels of description that can be applied to a single cortical area. Notice also that these different orders of vehicle have different contents. At the lowest order, neurons such as orientation filters have contents such as 'degree of similarity to 45 degrees off vertical at such and such a retinal location'. They have these kinds of contents because they are statistically correlated most strongly with this kind of feature at a specific location in retinal coordinates and are used to make decisions about the features at that retinal location. Unlike the SC map, this kind of content does reach into the world because photon densities can't describe the feature to which this kind of neuron is sensitive. Illumination doesn't particularly matter to these cells, there being the right orientation gradients is what matters (Callaway 1998). Orientation gradients are found on things in the environment, and are not directly reflected in photon emissions. The relevant photon impact rates depend on those gradients *and* the particular lighting conditions. Moving to higher-order vehicles, we find different dependencies. For instance, once we have the complete multi-dimensional representations, the strongest dependencies will be between sets of features that include things like color, depth, orientation, and contrast at any part of the visual scene that lies at certain retinal coordinates.

Let me now consider downstream cortical areas. Determining the kinds of vehicles these areas support helps us identify the kinds of transformations that the vehicles in V1 must undergo. In particular, this helps identify the transformations that map V1 vehicles onto V2 vehicles. Knowing the *precise* nature of neighboring vehicles can help us determine the transformations. However, a complimentary way to figure out what the vehicles are is to guess at the kinds of transformations that are present and make predictions about the vehicles you would expect to find. Like most things in science, the process is one of mutual refinement. For now, I will make some very unrefined guesses about the kinds of transformations and vehicles in higher cortical areas.

In general, the cortical areas in the 'what' pathway, including V1, V2, V4, PIT, and AIT, have projections to the next higher area, which tend to converge. Back projections, in contrast, tend to diverge (Van Essen and Anderson 1995; Felleman, Xiao et al. 1997).[8] This supports a kind of 'information pooling followed by top-down feedback' processing strategy. For example, it seems quite likely that orientation filters may have subtle responses that are sensitive elements of the visual field outside of what is called its 'classical' receptive field. These responses are probably in part due to feedback from higher areas (Knierim and Van Essen 1992). At this early stage in cortical processing, receptive fields tend to be quite small. They get larger as we move higher in the cortical hierarchy. By the time we are near the end, in IT, single neurons respond to features anywhere in the retinal image. And, some neurons in AIT respond to objects regardless of their orientation in three dimensions (Booth and Rolls 1998).

This general pattern supports a shift in the kinds of content we would ascribe at higher levels of cortical processing. In particular, there is a general trend of transformations from retinotopic representations to position independent representations along the ventral processing stream. So the content we ascribe shifts from 'such and such features at such and such retinal location' to 'such and such features in the visual field'. Given the presence of view independent neurons that seem to be part of distributed representations of individual objects, by AIT we may be able to ascribe contents like 'a dog with such and such features' or perhaps even 'such and such dog with such and such features'. This dog (anywhere in the visual field) will be the thing that has the highest statistical dependence with this higher-order vehicle. As well, there is an energy transfer to the vehicle from the dog, and the dog does not fall under our computational description. Thus, the dog is the referent of the vehicle. The

---

[8] In this context, 'convergence' means many neurons project to few neurons. 'Divergence' means the opposite.

transformations from V1 to IT have abstracted the features that support the identification of the dog. Furthermore, the vehicle is used by the system to identify this dog, perhaps classify and name it, perhaps orient towards it, etc.; these are just some of the many transformations that are licensed by this vehicle. This vehicle, then, is a vehicle that specifies an individual point in the huge space of objects. Given the nearness of areas involved in naming and language to the areas that support this kind of vehicle, subsequent transformations might map that kind of complex object representation onto a lexical one. A similar kind of story can be told about the lexical representations being mapped to motor commands that result in verbal behavior: 'there's a dog'.

Though this specific example has been a visual one in humans, there is evidence that many of the organizational properties I have identified are quite consistent across modalities and animals (Parker and Newsome 1998). Many of the features of content ascription evident in this example are thus general ones as well. We can expect many levels of higher-order vehicles, with the same neurons participating in different vehicles. We can expect the transformations to abstract more complex features independent of certain less biologically relevant variances (e.g., the form of objects). We can expect the kinds of content we ascribe to vehicles to pick out more complex statistical properties eventually picking out 'common-sense' objects in the world. We can expect the dependencies of vehicles to be with more complex features the farther in the vehicle is (i.e., the more transformations that precede it). We can expect to be able to ascribe different contents to different orders of vehicles, so a single neuron may participate in multiple content ascriptions.

4 Answers to the representational questions

That, then, is the theory of content I wish to defend. As I argued in chapter 3, a theory of representational content must answer the 13 representational questions, in order to properly address the problem of neurosemantics I identified in chapter 1. The purpose of this section is to answer those questions succinctly, based on considerations I have provided above.

4.1 What are the basic vehicles?

Basic vehicles are neurons as functional units. They are uncontentiously basic vehicles under the computational isomorphism description. As carriers of content proper, their description depends on their role in a whole theory of neural function.

4.2 What are the higher-order vehicles?

Higher-order vehicles are the theoretical objects we posit to provide good explanations of the operations of a neurobiological system. The right ones will be posited by the best theory.

4.3 What is the relation between basic and higher-order vehicles?

In regards to content, this relation is the encoding/decoding relation specified in chapter 6. As physical objects, the relation is mereological (i.e., higher-order vehicles are groups of basic vehicles).

4.4 What is the system?

Given the problem of neurosemantics, the system is the nervous system.

4.5 What is the relation between the basic vehicles and the system they are in?

Basic vehicles are parts of the nervous system (i.e., they are in a mereological relation).

4.6 What is the relation between the higher-order vehicles and the system they are in?

Higher-order vehicles are also parts of the nervous system. Higher-order vehicles provide various levels of description of the system to aid our understanding of its operation.

4.7 What gives a basic vehicle its content?

The content of a basic vehicle is determined by its causal (referent) and transformational (use) relations, both of which can be described computationally. Taking the perspective of the animal is important for generating this description.

4.8 What gives a higher-order vehicle its content?

The content of a higher-order vehicle is determined by its causal (referent) and transformational (use) relations, both of which can be described computationally.

4.9 What is the relation between a basic vehicle's content and the system it is in?

Although a basic vehicle will fall under the same computational isomorphism regardless of where it is in the system, the individuation of basic vehicles *as carriers of content* depends on transformation relations to other elements of the system.

4.10 What is the relation between a higher-order vehicle's content and the system it is in?

The individuation of higher-order vehicles as carriers of content depends on transformation relations to other elements of the system.

4.11 What is the relation between the basic vehicle and the external environment?

Basic vehicles are in a referent relation with items in the external environment. The external relata of the referent relation is determined by energy transfer (i.e., cause) and the highest statistical dependency with basic vehicle responses.

4.12 What is the relation between the higher-order vehicle and the external environment?

Higher-order vehicles are in a referent relation with items in the external environment.

4.13 What is the relation between the system and the external environment?

Under a strictly causal description, there is no special relation to speak of. Energy flows between environments and nervous systems as it flows anywhere else. However, as regards content, the external environment doesn't fall under computational descriptions. This helps determine what the referents of representational states are, as well as what counts as a representational state and what doesn't.

5 Summary

I argued in section 5 of the first chapter that theories of content might fair better by actually attempting a neuron-by-neuron characterization and building up from there. The details in the last chapter in conjunction with considerations I have presented in this chapter show precisely how such a theory can be constructed. By finding a computational/causal isomorphism at the level of single neurons, and knowing how to relate sets of neurons to construct more complex vehicles we can unify causal and conceptual role factors in a general theory of content. This theory is applicable at any level of description of the behavior of neurobiological systems.

The theory of content I have offered combines a causal factor, which determines a representation's referent, with a transformational factor, which determines how a representation is used. However, unlike standard two-factor theories, on this theory the factors are aligned in virtue of being descriptions of the same underlying process. I have shown how such processes can be described as purely causal or as purely transformational. In all likelihood, the most satisfactory explanations (i.e., those that fit well with our current theories and background assumptions), will be those that combine causal *and* transformational descriptions. This combination is justified by the fact that these descriptions are unified by the underlying computational theory.

That, then, is the neuron-by-neuron characterization that Lycan (1984) suggested we shouldn't bother with. Whether or not such a theory can fair better than other theories is the focus the next chapter.

CHAPTER 8

Concerns with Content

I think we ought always to entertain our opinions with some measure of doubt. I shouldn't wish people dogmatically to believe any philosophy, not even mine. – Bertrand Russell (1872-1970)

1 Introduction

For many philosophers, the holy grail of philosophy of mind is to "somehow [get] from motion and matter to content and purpose – and back" (Dennett 1969, p. 40). Answering the 'thirteen representational questions,' as I have called them, is far from likely to convince anyone that the theory presented here can do this. Scores of questions remain that I have not yet explicitly addressed in presenting the theory. The purpose of this chapter is to try and anticipate some of the more pressing questions that may be raised, and to provide answers to those questions that are consistent with the theory I have described.

I begin by addressing the concern that statistical dependence is unable to support an interesting and complete theory of content. This discussion supports the extension (in the subsequent section) of the account of occurent content that I have so far concentrated on, to an account of conceptual content. As I noted in chapter 5, a satisfactory account of content must support a robust notion of misrepresentation. In section 3 of this chapter I show the novel way in which the theory accounts for misrepresentation.

2 Statistical dependence and representational content

2.1 Statistical dependence and causes

A statistical dependence between two events means that the occurrence of one event changes (either increasing or decreasing) the probability of the occurrence of the other event. As I noted earlier (chapter 4, ff. 5), one natural measure of the strength of the statistical dependency between two events is mutual information (Hyvarinen 1999). When events are not so related, they are called *independent*, and the probability of their both occurring is the product of the probability of each occurring individually.

Statistical dependence in general, I take it, is a good reason to think events are causally related (see chapter 5). If the probability of a book falling to the floor depends on whether or not I loosen my grasp, then we would conclude that these two events are causally related. Of course, statistical dependence is not *enough* because two events may be statistically dependent if they are the *result of* the same cause (in which case neither is the cause of the other). This is where the addition of energy transfer to the theory of cause (and content determination) becomes important. Without energy transfer between two statistically dependent events, we wouldn't claim that one is the cause of the other.

Statistical dependence can also help us know when events aren't causally related. In particular, statistical independence means there is *no* causal relation between two events. If, for example, striking a match is independent of whether or not that match lights then we would have no basis for thinking that striking the match and its lighting are connected. This, of course, is something like Hume's point. On the theory I have presented, it is such causal relations that help fix content.

An apparent concern for any causally based theory of content is: What if the underlying theory of cause is wrong? In this case, the question would be: What if statistical dependence and energy transfer are not necessary or sufficient for actually *being* (metaphysically) a cause? There are, after all, many philosophers who think that even the basic assumptions of this kind of Humean analysis are wrong (see Sosa and Tooley 1993, pp. 1-33 for a review). If these philosophers are right, is this theory of content doomed to failure? The answer is no.

While I have defended this causal theory (chapter 5, section 3) from the kinds of concerns Sosa and Tooley (1993) discuss, the success of the theory of content does not rest on the success of this theory of cause. The reason is quite simple: this theory of content doesn't depend on there being a causal relation between the referent and the vehicle, it depends on there being a statistical/energy transfer one. We can use the dependencies

and energy transfers to identify the presence of a cause *whether or not* cause is identical to such energy transfers and dependencies. For this theory of content, identification is the key. In other words, we have to *know* when the vehicle and referent are causally related. This, then, is a weaker form of a causal theory of content. It is weaker because it only depends on knowing when there is a cause, it does not specify the metaphysical nature of cause.

Even if this weaker form is found to be unsatisfactory (i.e., if these relations don't help identify causes), the theory of content is still not in jeopardy. If it is discovered that, metaphysically *and* epistemologically speaking, cause is not related to statistical dependence and energy transfer this doesn't mean the right theory of content doesn't depend on exactly that relation; maybe that aspect of the theory just shouldn't be called *causal*. Therefore, as far as getting content determination right, getting the right theory of cause isn't too important. Of course, it would make for a more convincing theory of content if the causal theory were right. And, there are good reasons to think that this causal theory is likely right, but this isn't a *necessary* condition on understanding representational content.

2.2 Getting the right statistical dependence

None of these considerations speak to determining *which* dependencies are the right ones for referent determination. A vehicle will have many things on which it is statistically dependent. Statistical dependence is cheap: of course, this is where the statistical dependence hypothesis comes in. On this theory, the dependence that counts for content determination is the one that is the highest i.e., the referent of a vehicle is that set of causes on which the vehicle is *most* statistically dependent. That, as I explained in the last chapter, is how we determine the referent of the vehicle.

But counter-examples come racing to mind. Consider again looking at a dog in a field: Why is my higher-order vehicle, whatever set of neurons it may be, about the dog itself and not, for instance, the set of photons that intervenes between me and the dog? On first glance it seems that these (non-computationally described) photons will covary just as well, or perhaps better, than the dog with my internal vehicle. The photons are, after all, the things that directly affect my neural firings; so we might expect them to be less susceptible to intervening distortions and therefore have a higher statistical dependence with the content of my 'dog' vehicle.

But we have to be careful about what is being dependent on what. It clearly isn't the case that my 'dog' vehicle is statistically dependent on any particular photon, or even a bunch of photons that recently bounced off the dog. In fact, any given photon bouncing off the dog will be nearly independent of the neural firings that occur when the *dog* is in front of me. After the photon bounces off the dog it is absorbed by retinal cells and converted to chemical energy. Life as a photon is over, but the neural firings certainly don't stop. Notably, then, my 'dog' vehicle doesn't particularly depend on the physical properties of one photon, or even a bunch that recently left the dog. Rather, the higher-order 'dog' vehicle particularly depends on the physical properties *of the dog* and that is why it represents that dog.

We can make the counter-example more sophisticated by allowing that it is not particular photons (or a recent bunch) on which the vehicle statistically depends, but a more appropriately delineated set of photons. Such a set has to be something like the photons in the set of time-slices just before hitting the retina that have ever bounced off the dog. Assuming the 'dog' vehicle is as dependent on this set as on the dog, there are at least three good reasons that this set is a bad candidate for the referent of the vehicle. First, it is far *simpler* to use the dog as the referent. Ockham's razor should convince us to choose the dog as referent over this complex, difficult-to-specify set. After all, one thing that the set of photons has in common is their having bounced off of the dog.

Second, the dog vehicle is dependent on far more than just photons. Dogs are, after all, multi-modal stimuli. They are furry, noisy, and smelly. All of these modalities help determine how much the vehicle depends on something in the world. A set of photons, no matter how well delineated, will only explain the dependence of part of the total dog vehicle. Even if we only *see* the dog, and we don't touch, hear, or smell it, the vehicle on which that seeing depends will have these other modes inferred (through appropriate transformations). Since these inferred values for non-visual parts of the vehicle statistically depend on the dog, and not the set of photons, it is the dog that is the referent. The obvious rebuttal to this response is to insist that it's not just the appropriate set of photons, then, it's the set of photons, plus the set of electrons between skin and fur, plus the set of sound waves, plus the set of molecules that adhere to olfactory receptors, etc. on which the vehicle is most statistically dependent. To answer this, I can again appeal to the fact that it is the dog that is the source of all of these sets, so we should choose the dog as the referent.

Third, we must recall that the content of vehicles depends not only on causes, but also on uses. That is, the kinds of transformations that result from tokening the 'dog' vehicle support ascriptions of properties that only make sense for dogs, such as 'is vicious'. Sets of photons, electrons, molecules, etc. just aren't the kinds of things that can be vicious. Notably, this point reduces to the previous one. That is, if we did describe the referents as photons, electrons, molecules, etc. then the properties we took the system to be ascribing probably wouldn't be things like 'is vicious'. Of course, as I mentioned last chapter, we want a theory of content that is consistent with other theories. In many of these theories (e.g., biology, macro-physics, astronomy, etc.) what are quantified over are objects with properties; objects like dogs, tables, airfoils, stars, and so on. The best theory of content, then, is one that quantifies over the same sorts of things. This is especially true when quantifying over such things helps make the theory simpler, to the extent of being actually manageable.

2.3 Some consequences of using statistical dependence

Adopting statistical dependence as a means of referent determination has a number of beneficial consequences. First, statistical dependence comes in degrees. The highest dependence of some vehicle with its referent can be higher than the highest dependence of another vehicle with its referent. The strength of the dependence maps nicely on to the precision of the representation in question. If, for example, the occurent dependence of my representation of the dog's position with the dog's position is nearly perfect (which means any changes in my relation to the dog are reflected in changes in the contents of my vehicle) we know that my representation is precise (in a technical sense, see section 4).

A second benefit of this approach is that we don't have to rely on intuitions about what is represented by cognitive systems. Rather, we have to systematically examine the features of the environment on which the vehicle is statistically dependent; we will *discover not stipulate* what the referent of the vehicle is. Of course we can still use our intuitions to generate hypotheses about vehicles and referents. If we have a set of neurons we believe to be a vehicle, we can construct the joint probability histogram between those neurons and features of the environment (by taking the animal's perspective), and thus discover the referent of that vehicle. In other words, we can test our hypotheses about possible vehicles. If we suppose that a certain set of neurons acts as a 'dog' vehicle, we can test that hypothesis by seeing how statistically dependent the vehicle and referent are. If they have no dependencies, or have better dependencies with other vehicles or referents, our hypothesis about the nature of the vehicle will change. The converse may occur as well; i.e., what we take to be good referents will help us determine vehicles. As I discussed in chapter 4, both methodologies serve to give us a good idea of what vehicles and referents there are. What is most important, is that this theory allows for these to be *discoveries* (unlike the traditional neuroscientific approach; see chapter 4) – no doubt we will be surprised by some of the things biological systems depend on for representing the environment.

The last consequence of adopting statistical dependence I will discuss is a subtler one. Because referents and vehicles are picked out by statistical dependences, they are obviously mutually dependent. Furthermore, referents help determine contents, as I discussed in the last chapter, so contents depend on referents as well. Therefore, content depends on vehicles. This shouldn't be too surprising. If you know all there is to know about a vehicle you will know its possible content. However, the vehicles we discover won't *determine* content, but rather *constrain* possible content. Psychophysics, for example, tells us that there are certain dynamic ranges over which our retinal cells can encode brightness. Outside of those ranges, differences in intensity are indistinguishable; that's a fact about physiology. If the vehicles can't carry such differences in intensity, then those differences can't be used by the system to react to the environment, so those vehicles can't carry content about those differences.

What this means, then, is that vehicles and contents aren't independent. In a more traditional turn of phrase: syntax and semantics aren't independent.[1] The stuff that carries meaning (syntax/vehicles) also helps to determine meaning (semantics/content). As powerful as natural languages are, they are stuck with certain vehicles. There are contents that these vehicles carry well and there are contents these vehicles don't carry well. Imagine that I'm looking at a flashing red light. I now write down: "There is a flashing red light". I can also take a video of the flashing red light. I can show both of these representations to you. The video is a better

---

[1] For arguments as to the benefit of blurring the distinctions between vehicles and contents, see e.g. Langacker (1987) and Mohana and Wee (1999). See Eliasmith and Thagard (in press) for an example of how blurring the distinction may help explain the nature of high-level cognitive processes.

representation than the sentence because it has the higher of the dependencies with the referent of the two representations (i.e., the flashing red light). The syntactic properties (i.e., physical structure) of each of these vehicles helps determine, very differently in this case, the precise meaning (i.e., the exact properties assigned to the referent) of the representations.

I take it that each of these consequences goes towards showing that relying on statistical dependence and energy transfer is a good idea. In the next section I show how adopting this method changes our understanding of the nature of concepts. Subsequently, I will defend this method against concerns that it cannot convincingly explain misrepresentation.

3 Occurent and conceptual content

3.1 Introduction

I introduced the distinction between occurent and conceptual intentionality (or content) in chapter 5. The term 'occurent content' denotes the content of currently active vehicles. It can be thought of as the encoded value of the variables (or more complex structures) of the neurobiological system at any one point in time. This, then, is the kind of content I have been mostly concerned with up until now: it is the kind of content I discussed in the visual system example of the dog in the field in the last chapter; it is the kind of content had by the 'horizontal eye position' vehicle; and it is the kind of content determined by the causal encoding processes I detailed in chapter 6.

However, this is not the kind of content that philosophers have traditionally focused their attention on. Rather, philosophers have been interested in the content of our concepts; content that has a kind of stability. Those who hold teleological accounts of content, including Dretske (1988) and Millikan (1984), hold that our 'dog' concept is about dogs because the part of our brain that represents dogs has the evolutionarily determined function to represent dogs. Others, like Fodor (1998), believe that our 'dog' concept represents dogs because there is a unique law-like relation between our 'dog' concept and dogs. These theorists are centrally concerned with what is called the *reference* relation (not to be confused with the *referent* relation I have been discussing). I briefly discussed this relation in chapters 1 and 4, but I did not show how it is accounted for by this theory; I do that in section 3.4. First, however, I extend the theory I have presented so far to account for the conceptual content, and present some examples of how it applies to some challenging examples of content determination in section 3.3.

3.2 Conceptual content

The account of conceptual content that follows from what I have so far presented is not so much an extension of the current theory as a reconsideration of relations already introduced; in other words, nothing new is added. Before providing this account, it is important to get a sense of what a theory of conceptual content must be able to explain.

Fodor (1998, chp. 2) argues for five criteria that must met by a theory of concepts. Four of these are relevant to this discussion:[2] 1) concepts are mental particulars and function as mental causes and effects; 2) concepts are categories; 3) many concepts must be learned; and 4) concepts are public and people share them. In this section I show how the theory I have proposed can meet each of these criteria.

---

[2] The criterion I do not discuss is that concepts must be compositional. That is, "mental representations inherit their contents from the contents of their constituents" (Fodor 1998, p. 25). I don't discuss compositionality for two reasons. First, it is a prime example of misplaced intuitions about language. If we think animals have concepts, as Fodor (1987) does, and we have no evidence that these concepts are compositional, why make compositionality a necessary condition on theories of concepts. Furthermore, basic compositionality seems to fail even in language: consider 'couch potato', 'raining cats and dogs', and 'unravel' (which is synonymous with 'ravel'). Second, a suitably careful consideration of compositionality (which would have to include a consideration of systematicity and productivity as well) and its relation to the theory of content presented here would be too lengthy.

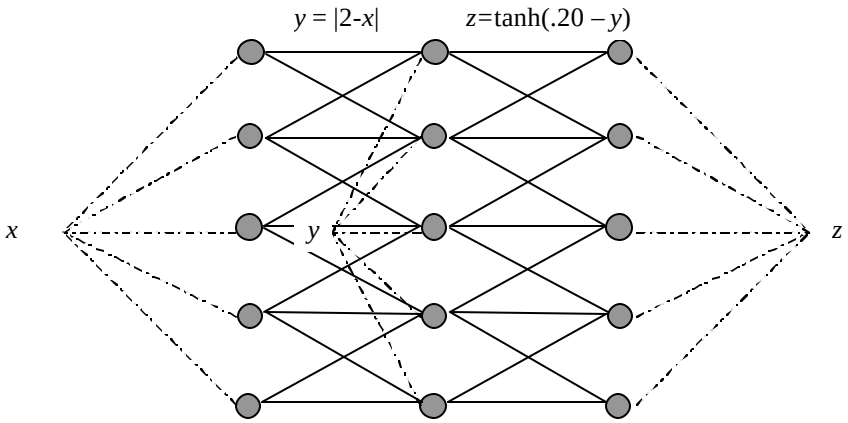


Figure 7: A simple three layer network that acts as a categorizer.

In order to explain how the theory meets these criteria, consider the simple three layer network depicted in Figure 7. The first layer encodes an analog value, *x*, between 1 and 10 to two decimal places, and the second layer encodes, *y*, the nearness of that value to 2. We could write this transformation as $y = |2-x|$. The vehicle in both layers is an analog variable. The nature of each of these vehicles is determined by the properties of the neurons over which the vehicle is defined (the individual layers in this contrived case). The weights that perform this transformation between the two layers can be determined as explained in chapter 6. These weights are determined by a combination of the properties of neurons in both layers (and they can be changed by appropriate learning rules). The weights are thus relations between neural properties that determine how input is transformed. Suppose also that there is a third layer, *z*, that 'thresholds' the output of the second layer. So, any value greater than 1.80, say, would count as 2 and cause this layer to encode a 1, anything less and the layer encodes a -1.

This network instantiates a simple categorizer. In this example, we can think of the transformations as categorizing all values in the first layer between 1.80 and 2.20 as the value 2. The transformation, then, picks out the property of 'twoness', in the sense that neurons in the third layer output a 1 when something sufficiently 2-like appears on the input. A large number of stimuli on the first layer are mapped to the same output on the third layer; this is the essence of categorization. And, categorization is the thread that unites conceptual behaviors.

I think this admittedly simple case can help show how the theory I have been discussing accounts for the criteria that Fodor (1998) claims must be met. 1) Clearly, this account shows how we can explain mental causes and effects. After all, *x* causes *y* to have the value it has. The value of *x* determines the value of *y* via the transformations relating *x* and *y*. Of course, 1) is a conjunctive claim. The first part says that concepts are *mental particulars*. The mental particulars at work in this theory are vehicles. If vehicles are to be concepts, I must explain how they can meet the remaining criteria. 2) As the example of the three-layer network demonstrates, these vehicles are quite naturally understood as categorizers. The vehicle *y* categorizes values according to their distance from 2 (so both 3 and 1 will be in the same category). And, even more obviously, *z* categorizes all values of *x* as either 2-like or not. 3) The categories can be adjusted just by changing the relations between the neurons. This, then, is the sense in which the vehicles are learnable (see Bishop 1995 for an overview of some learning procedures for such networks).

The last criterion is, *prima facie*, the most difficult for this theory to satisfy. How can *vehicles* be shared if they are individuated based on (unique, individual) transformations? Recall that a similar question arose concerning the relativism objection to two-factor theories in chapter 2 (How can meanings ever be the same if conceptual role helps determine meaning?). There, as here, we simply needed to realize that 'exact sameness' is simply too strong a criterion. Fodor's (1998) own example gives this away: he claims that he and Aristotle share the concept 'food' (p. 29). If this is true, sharing a concept must be a rather easy criterion to satisfy. If both he and someone from a very different culture are said to share the concept 'food' (as Fodor demands), yet he and that other person differ on what counts as food (he may count fish eggs, the other may count ants), then they only have to *partly* agree on the *application* of the concept to be said to share it. So, the three-layer network could have counted values between 1.81 and 2.20 as 2-like and it should still be said have the 'same' categorization. In almost all cases, after all, that network will categorize like the original. So, criteria 4) can be satisfied as long as vehicles in two different people have similar categorization results. In all likelihood, sharing concepts will be a matter of degree; and degrees of sharing *are* captured by this theory.

So, vehicles can satisfy each of the criteria Fodor proposes for concepts. However, this may seem an odd state of affairs. Vehicles, after all, can carry content about edges at some location in the visual field. Is this what we mean by the term 'concept'? Probably not, but it is likely just that Fodor's criteria are not sufficient conditions on something's being a concept. However, this might be a result of the vagueness of the term, rather than any failing of Fodor's. The wise tactic, it seems to me, for a theory like the one I have proposed is simply to explain whatever people use the term 'concept' for in terms that are well defined in this theory. To this end, in each of the next sections, I address some of the more difficult questions about content ascription that are explained in terms of concepts.

3.3 Conceptual content in action

The question I would like to answer in this section is: How can I think about a dog without there being a dog in front of me on this theory? Answering this question shows how my theory of content addresses aspects of concept employment not captured by meeting Fodor's (1998) criteria. In the next section I tackle related issues by using this theory to analyze the reference relation. Together, these two sections provide a picture of how to explain conceptual content is on this theory.

The above question addresses the heart of what many take to be a central benefit of having concepts; namely, 'offline' mental manipulation. Let me limit the discussion to visual properties of objects for the moment. I described the visual system in chapter 7 as being essentially feed-forward. This assumption was carried over to the simple example of categorization in the previous section. However, I was careful to note the ubiquity of back projections among visual areas. Let me suggest at least a partial reason why these are so important.

Suppose that someone says, "Picture a dog," when there are no dogs present in my immediate environment. What happens in the visual system? Kosslyn (1994; Kosslyn, Thompson et al. 1995; Kosslyn and Thompson 1999) provides evidence that visual areas used to process visual input are also active during such imaginings. Presumably, they are activated by back projections from higher visual (or even non-visual) areas. This suggests that the same vehicles that are statistically dependent on certain environmental properties during visual perception are activated by other higher brain areas (e.g., language areas). This activation can then spread (through various transformations) to visual areas, giving dog-related vehicles occurent content. It is this occurent content that can be used for offline reasoning.

But what is the referent in such a case? An application of the constraints on referent determination provides the answer. That which correlates best with the 'dog' vehicle and transferred energy to it is presumably the dog that caused me to activate the vehicle. And, this dog is external to me and thus does not fall under a computational description. Notably, at least some of the properties that I ascribe to this dog depend on the properties of dogs I have already been in causal contact with. Properties that aren't determined directly by the causal interaction with this particular dog are inferred as needed through various transformations. If most of the properties assigned to the referent have to be generated by 'default' transformations we will get insight into what is normally called the 'dog' concept. This content, then, is generally (i.e., in the cases of learned concepts) determined by those things we have previously been in causal contact with. I won't call something a dog if doesn't have any properties that dogs I've been in causal contact with have had.

It may seem odd that an agent has to come in causal contact with a dog in order to represent (or conceptualize) one, but recall that causal contact simply means 'having energy transferred to'. Thus, even in the case when you utter the words 'the dog' to me, or provide a description of some dog, I come in causal contact with that dog. And, if your description is a good one, I may have a vehicle that is highly statistically dependent on that dog. This is how we can learn about things (such as moon rocks) that most of us have never been in 'direct' causal contact with. Energy transfer is ubiquitous. When I touch moon rocks, and then describe the sensation to you, there is a determinate amount of energy transferred from the moon rock to you. The causal chain from moon rocks to you just happens to have me in the middle (rather than air, or photons, or what have you). Similarly, we can learn about dogs by seeing them (as described in chapter 7), but we can also learn about dogs by having someone tell us about them. This may seem like a difficult story to tell, but it would be more surprising if the story we need to tell of offline reasoning was any simpler. And, there is nothing in this story that can't be captured by the details given in chapters 6 and 7.

3.4 Traditional reference

The previous example has shown how this theory handles some of the more difficult cases of conceptual content ascription. However, some of the most difficult cases of all to explain convincingly are those that arise from a consideration of the conventional notion of *reference*. As I noted in chapter 1, reference is traditionally distinguished from *sense*. Simply put, the sense of a representation is the set of properties ascribed by that representation (see chapter 1). The reference of a representation is the *relation*[3] between the representation and the object those properties are ascribed to. On the face of it, reference seems much like the referent relation I have been discussing so far. However, as I showed in chapter 5, it is not.

It is important to note here that I have purposely not adopted the reference/sense distinction because reference, as I will show, is poorly defined. Nevertheless, I have to be able to explain the phenomena this distinction is used to explain. In particular, I have to explain our ability to ascribe properties to all kinds of objects. I have discussed how we ascribe properties to objects in our immediate environment and to objects not in our immediate environment that we are in causal contact with. However, there is another kind of object to which we ascribe properties: those we have never been, or can never be, in causal contact with.

There are two kinds of property ascription that meet this criterion: 1) property ascriptions to objects that do not exist; and 2) property ascriptions to objects that we have never been in causal contact with. An example of the first is our ascription of various properties to unicorns (e.g., horned, horse-like, etc.). An example of the second is our ascription of various properties to a dog outside of our light cone (e.g., small, brown, fuzzy, wet nose, etc.). In other words, I must explain how we can 'refer to' unicorns and also how we can 'refer to' a dog outside our light cone. Under the traditional understanding of the reference relation, this means I must explain what the relationship is between: 1) our 'unicorn' representation and unicorns; and 2) our 'dogs outside our light cone' representation and dogs outside our light cone. The traditional answer would be, of course, that our representation bears the reference relation to these things. However, we clearly do not bear the *referent* relation to them – since we are not in causal contact with them.

Let me begin by examining the case of our 'unicorn' representation. It is immediately apparent that there is a difficulty with the claim that our representation of unicorns bears the reference relation to unicorns; unicorns simply don't exist. It is not possible to define a relation when one of the relata does not exist. Relations are, after all, mappings from one thing (our representations) to another (what those representations are about). If there is no element in the range for the element in the domain to be mapped to, the mapping is undefined. Therefore, the reference relation is undefined in this case.

How, then, does the referent relation handle such cases? We can begin by asking: What is the content of our 'unicorn' representation? On the theory presented here, the content of our 'unicorn' representation is determined by some referent signal and its licensed transformations. It is the 'signal' that is the problem when the 'right' causes aren't obviously available. But, we do have *related* referents in our environment; we have seen horns and horses, and we have even seen them combined in pictures, paintings, and in movies. The referent of the 'unicorn' vehicle *itself* will be those depictions – they are, after all, the things on which our vehicle is most statistically dependent. They would, of course, be more statistically dependent on real unicorns, but they are not causally connected to such beasts, because such beasts don't exist. The relation our representations bear to *unicorns* is thus undefined, but the relation our representations bear to depictions of unicorns is the referent relation. Such depictions are all we can appeal to when explaining inferences and behavior. Furthermore, such depictions are all we can appeal to when evaluating the truth or falsity of unicorn claims.

Things are more complicated in the second example; dogs really exist. It is quite possible that there are dogs beyond our light cone (or, if not dogs, stars, or hydrogen atoms, or the like). What, then, is the relation between these dogs and our representations of them? Notice, first, that we can never *know* if there are dogs beyond our light cone or not. There is *absolutely no way* we can be in *causal* contact with such dogs (not even with quantum 'action at a distance' (Zeilinger 2000)). Nevertheless, if we want to be realists, we must presume that such dogs may exist. If we are realists, and we can't be in causal contact with something we are making claims about, all such claims must be hypothetical; i.e., they are about unconfirmed states of the world.

---

[3] Some, like Brentano (1874) didn't think that reference was a relation, but rather relation-like (as in Chisholm 1957, p. 146). For those who agree with Brentano, this section can be considered a description of precisely how reference is like and unlike a relation.

Notice that a problem with the reference relation again becomes evident. Since the reference relation, like all relations, depends on the existence of the relata for its definition, it runs into problems when the existence of the relata are indeterminate; even though only epistemically indeterminate. The reason is that misapplications of the relation are indistinguishable from proper applications. Thus, the relation will never be properly definable. The best that can be said is that *if* dogs beyond our light cone exist, then they bear the reference relation to our representations of dogs beyond our light cone. If such dogs don't exist, the relation is undefined; but there's no way to tell the difference. This, it seems to me, is a highly undesirable state of affairs, especially if the relation is supposed to *explain* something (like content). So, again the reference relation seems to not be a proper relation after all. Rather, we have a set of conditional statements that *may or may not* help determine a relation.

However, these conditional statements themselves are important. They are the truth conditions of the representation; i.e., they determine if the statement is true or false. The statement 'there's a dog beyond my light cone' is true if and only if there really is a dog, out there, beyond my light cone. This truth conditional relation is a logical relation and that is why it is not fettered by causal boundaries. What really needs to be explained by a theory of content, then, is not the reference relation, but the derivation of such truth conditions.

The problem we are left with bears some striking similarities to the problem of explaining the projectibility of predicates (Goodman 1955). The problem with projectibility is to figure out how to distinguish projectible predicates (e.g., 'all emeralds are green') from unprojectible predicates (e.g., 'all emeralds are grue', where 'grue' means 'green before some future time and blue afterwards'). The problem for content ascription in non-causal cases is to explain how truth conditions are determined for such projections. A quick and easy answer to this latter problem is simply to say that my 'dog beyond my light cone' representation is about the dog beyond my light cone because I am disposed towards that dog in various ways (i.e., I would assent and dissent to claims about its brownness, location in space, etc.). The relevant truth conditions, in other words, are determined by my dispositions. The problem with this answer is that all the work is being done by my mysterious dispositions. What are dispositions?

I think the solution to the predicate problem suggests a solution to the content problem by explaining dispositions. In the case of projectible predicates, discerning projectibility comes down to determining the current status of the predicate. In particular, a hypothesis containing a non-grue-like predicate is projectible if it is *currently* supported and unviolated. And, a hypothesis containing a grue-like predicate is projectible if it is unviolated and each of the conjuncts of its equivalent expression in non-grue-like predicates is supported (Johnson 1995). Dispositions can be understood in the same way. In particular, wherever we would say 'I am disposed to assent to '*X*'' we can say, instead, 'I think that the predicates *Y, Z,* … in comprising '*X*' are projectible'. For example, we could translate 'I am disposed to assent to 'The dog beyond my light cone is brown'' to 'I think that the predicate 'brown' in comprising 'The dog beyond my light cone is brown' is projectible (in particular, projectible beyond my light cone)'.

The important point is that projectibility depends on the *current* status of the predicate. So, similarly, the projected predicates and their objects have contents that depend on their current features. Those features include the referent of the relevant representations. Thus, the representations retain their referents even under projections. Truth conditions, in contrast, can vary between projections and non-projections. Thus, 'this emerald is green at time t' (non-projection) and 'this emerald will be green at (future) time t+1' (projection) have very different truth conditions; one set of conditions could be met while the other fails. But, in both cases, the referent of the terms stays the same.

Clearly, satisfaction of truth conditions is not limited by causal relations. The statement 'there are dogs beyond my light cone' is true or false regardless of my causal connection to those dogs. What the truth conditions *are*, however, is causally determined. Furthermore, the determination of these conditions can be explained by this theory of content. We can understand projections as hypothetical assignments of properties by a vehicle to a referent. These properties can be unattained, or unconfirmable, there are no restrictions (save those imposed by the nature of the vehicle). The transformations that license these assignments can be the same ones that usually perform this task (e.g., if the value *was* 1.9 then it *would be* a 2). The subjunctive nature of the transformation is irrelevant to the transformation itself. The content of the representation and its referent, just like in cases of misrepresentation, determines the conditions under which the representation is true. So, representations of hypothetical (or non-comfirmable) situations have contents that depend entirely on causes, despite the fact that truth conditions so determined (and their satisfaction) don't.

I think this discussion shows three important things. First, reference is not a well-defined relation. This, of course, means that reference is a poor candidate for providing good explanations of content. Second, the referent

relation *is* a well-defined relation. Thus, the referent relation is a good candidate for providing a strong basis for a theory of content, including conceptual content. Finally, the referent relation can explain how truth conditions are determined. And, truth conditions seem to be all that is really left of reference once it is divested of its status as a viable relation. So, the referent relation as employed in this theory of content can more effectively account for the conceptual phenomena that 'reference' has traditionally been used to explain.

4 Misrepresentation revisited

In chapter 5, I briefly commented on the ability of this theory to handle the problem of misrepresentation. The time has come to see, in more detail, how this theory deals with the various kinds of misrepresentation. Recall that Dretske (1995) claims that the problem of misrepresentation has at least two sub-problems. The first is related to the kind of misrepresentation I discussed in chapter 2. Dretske (Dretske 1994, p. 472; Dretske 1995) considers this the problem of explaining how we assign the wrong properties to an object. This is the same as the problem Fodor (1987) calls the 'disjunction problem'. It is the problem of explaining how a representation can assign a specific set of properties to a referent even though the representation can be caused by a disjunction of referents that don't all have those properties. The example from chapter 2 is that of my 'dog' vehicle being caused by a cat under some circumstances that, according to a naïve causal theory, should mean the vehicle is actually a 'dog or cat' vehicle. So, the problem is how do I explain ascribing the property 'dog' to a cat in certain cases (rather than just tokening a disjunctive vehicle)?

The second problem of misrepresentation that Dretske (1995) identifies is that of ascribing properties to an object when there is no object at all; e.g., representing that dog to be fuzzy even though there is no dog to represent as such. These kinds of misrepresentations occur in cases of hallucination, imagining, and dreaming.

A third problem of misrepresentation that Dretske doesn't address, but is important, is that misrepresentations *can* be more common than a correct representation but we still want to classify them as *mis*representations. Ruth Millikan (1993, p. 62-3) points out that it is quite possible that a correct representation is the exception rather than the rule; i.e., that, statistically speaking, correct representation is less likely than misrepresentation. This kind of concern is particularly important to a theory, like the one I have presented, that relies heavily on statistical relations.

In this section, I address each of these problems and show how the theory I have presented can solve or avoid them. After considerable background discussion, I show that the *important* kinds of disjunctions can be handled properly by this theory. I also argue that Dretske's distinction between these two different kinds of misrepresentation does not stand up to scrutiny. Finally, I show how Millikanian concerns about common causes versus correct causes can be handled.

To begin, it is important to distinguish three kinds of misrepresentation; 'personal', 'social', and 'absolute' misrepresentation. Personal misrepresentation occurs in cases of poor performance. That is, in cases when I could have (with the same skills, experience, etc.) represented better than I did. If I call a dog 'brown' that, under other circumstances I would say is 'black', I have personally misrepresented the dog. Social misrepresentation occurs in cases when my property ascriptions don't agree with social norms. If I call a dog 'brown' that *everyone else* calls 'black', then I have socially misrepresented the dog. Subtleties of what counts as a social norm is clearly beyond the scope of this project, but I am happy to say that, *whatever they are* if my representations don't agree with them, it is a case of social misrepresentation. Absolute misrepresentation occurs in cases when my representation doesn't agree with the metaphysical fact of the matter. So, if I call a dog 'brown' and everyone else calls it 'brown' but it's *really* black, then I have absolutely misrepresented the dog. Of course, to think that this kind of misrepresentation occurs is to be a realist, and perhaps committed to the existence of natural kinds. Notably, I don't need to espouse such a position in order to explain how the consequences of such a position can be explained by a theory of misrepresentation. As regards the relations between these three kinds of misrepresentation, it is important that any instance of representing can fall under any of the categories or not *independently* of which other categories that instance falls under. So, I can personally represent while socially misrepresenting and absolutely representing, or I can personally misrepresent while socially representing and absolutely representing, etc. I think that neglecting this distinction has lead to much of the seeming difficulty with misrepresentation. Nevertheless, all three kinds of misrepresentation need to be accounted for. I will focus on personal misrepresentation, but will also show how the explanation of personal misrepresentation generalizes to account for the other two kinds of misrepresentation.

First, however, I think it is important to critically examine how cases of (mis)representing are understood by philosophers. In particular, philosophers tend to presume that cases of representing fall into one of two categories: right or wrong. This, it seems to me, is a mistake. Dretske (1994, p. 472), for example, defines misrepresentation as the saying of something that does not have a property, that it has the property. For example, saying that a black dog is brown. In other words, misrepresentation is representing one thing (a black dog) as another (a brown dog). Unfortunately, under a strict application of this definition, it is not clear that we ever get anything right – that we *ever* represent.

Consider my representing a black dog standing in front of me. Suppose that after 3 minutes the dog is removed from my sight. I would then be able to answer all sorts of questions about its shape, size, length, etc. based solely on my representation of the dog. However, each of my answers would be inaccurate in some way. Consider the line of questioning: What color was the dog? Answer: Black. Dark or light black? Answer: Dark black. This color (showing a color patch)? Or this color (another patch)? etc. There is little doubt that I would eventually answer incorrectly. Does that mean I am attributing to this dog a property it doesn't have (i.e., a certain shade of black)? Yes, it does. Does it mean we should say I am misrepresenting the dog? According to Dretske's definition it does, but I think that such an answer is hasty. In order to say why, consider some elementary distinctions in measurement theory.

Measurements are said to be *accurate* if they are near the right value. If I measure darkness and get the right answer, I have made an accurate measurement. Measurements are said to be *precise* if they are reproducible. If I measure the darkness of a color patch over and over again, and get the same answer every time, I am making precise measurements of darkness. If my measurements are precise and accurate, they are said to be *exact.* Notably, precision is a property of a set of measurements, while accuracy is a property of a single measurement. But, we can define the accuracy of a *set* of measurements to be the *average* nearness to the right value. In statistical terms, precision is measured by the variance of a set of measurements, while accuracy is the difference between the average measurement and the correct answer.

What the line of questioning above is doing, is probing the representer with increasing degrees of precision. Although the representer may be perfectly accurate at one degree of precision (black versus white) the representer may be inaccurate at another (one color patch versus another). We know already that neural representations have a limited degree of precision; only about three bits of information are transmitted per spike (see chapter 4). So we shouldn't be surprised that we will *eventually* misascribe properties; it isn't possible to be consistently accurate to a degree of precision greater than our 'measuring device' can provide.

What is important here is that representations are best characterized as 'better' or 'worse', not 'right' or 'wrong'. 'Better' means high accuracy with high degrees of precision. 'Worse' means low accuracy with low degrees of precision. We may want to make 'right' and 'wrong' claims at a given level of precision, but in doing so we'd have to make an argument why being below some standard of accuracy at a given precision is a good criteria for making this distinction. Dretske's definition clearly does nothing of the sort. But, perhaps arguments that look to pragmatics (i.e., 'good enough to help the animal survive') would be convincing. Even so, there are going to be representations that come closer to that standard and some that are farther away. Using the term 'misrepresentation' to divide representations into two groups obscures important subtleties of representation. We will have a more general understanding of misrepresentation (i.e., one that doesn't depend on choosing particular standards and can account for degrees of deviation from any given standard), if we accept that representations come in degrees; i.e., that they lie on a continuum from good to bad.

The preceding discussions have laid the groundwork for addressing the disjunction problem head on. Recall that I am primarily interested in personal misrepresentation. The challenge of the disjunction problem, then, is to explain how my representation of something (a cat) could mean to me that it had some property (the property of being a dog) even though it was caused by something that I would say didn't have that property (a cat). The solution I offered in chapter 5 was that we can explain this case of misrepresentation by noting that the statistical dependence hypothesis and its corollary give different answers. In particular, under all stimulus conditions, this vehicle has the highest statistical dependency with dogs even though, under this condition, this vehicle has the highest dependency with a cat.

But, does this really solve the *disjunction* problem? Won't it be the case that the highest statistical dependency holds between dogs-or-this-cat under all stimulus conditions? In fact, no. This cat under *all* stimulus conditions will not have a high statistical dependency with my 'dog' vehicle, it will have a highest dependency with my 'cat' vehicle. It is only under *this* stimulus condition that it has a high statistical dependency with my 'dog' *vehicle*. So, perhaps the problem is that my vehicle has the highest statistical dependency with dogs-or-this-cat-

under-these-conditions. Luckily, this disjunction can be ruled out because it includes a specification of stimulus conditions. We can't find a dependency between a vehicle and something-under-a-stimulus-condition under all stimulus conditions since most of the stimulus conditions are ruled out by such a characterization. It would be self-contradictory to try and determine such a dependency.

Clearly, the concept of 'stimulus conditions' is doing a lot of work in this explanation. I need, then, to provide a means of distinguishing stimulus conditions. I have at least two options here. First, I can appeal to other scientific theories to help me individuate stimulus conditions. I can say that a difference in stimulus conditions is a difference in any physical variable relating the referent and my representation. Such variables include things like distance, illumination, relative velocity, etc. However, this definition makes it difficult to distinguish incremental changes in these more intuitive stimulus condition variables, from incremental changes in variables that specify the physical properties of the referent itself. This doesn't seem to be a problem for the theory I have presented, although it may be a problem for determining what the referent is (i.e., in answering the question "is it a dog or a cat?"). Some 'stimulus conditions' so defined would be difficult to realize because we can't construct 'dogcats'. That is, we can't "morph" dogs into cats in the real world. This may limit our ability to systematically examine 'all possible' stimulus conditions. Although this may seem problematic, development of our concepts would also be so limited so it may not be a hindrance after all. Furthermore, representation on this theory is still perfectly well defined for such 'non-actualizable' situations.

Rather than accepting this as a difficulty for the definition of stimulus condition, I could make a more direct appeal to other scientific theories that quantify over certain kinds of objects (e.g., biology and biological kinds) to rule out the problem. The properties of such objects, then, would be ones that should not be changed when changing stimulus conditions. In this case, individuation of stimulus conditions would be determined by changes in physical variables relating referent and vehicle such that the physical properties of (say) biological objects (e.g., an individual dog) are not affected.

Second, I can take a weaker position and note that 'stimulus condition' is a technical term in a number of scientific enterprises (e.g., psychophysics, neuroscience, psychology, etc.) that I can leave up to *them* to define. In other words, I can claim it is a practical problem beyond the theoretical considerations I am concerned with here. In some ways, this is a weak response – it seems that I'm simply passing the buck. However, I can point to the successes of such sciences in providing results and explanations that depend on individuating stimulus conditions. In this way there is clearly reason to think that 'stimulus condition' is a scientific term whose definition may be forthcoming. Better yet, I can point to the specific results of the blowfly and cricket experiments I have already discussed (see chapters 4 and 6). Such experiments have methods for determining exactly the kinds of statistical dependencies I am talking about. So, a notion of stimulus conditions that is consistent with these procedures and results will probably do, however vague it may be. In all likelihood, a more precise understanding of how we should individuate stimulus conditions will evolve with more experiments of this kind (in neuroscience, psychology, psychophysics, or wherever). For present purposes, either of these two understandings of 'stimulus conditions' will suffice.

Given a satisfactory understanding of stimulus conditions, a generalization of my characterization of personal misrepresentation to include social and absolute misrepresentation is straightforward. Cases of social misrepresentation can be picked out by noticing when my referent of doesn't match a socially defined one. In other words, we can define a 'social' statistical dependence hypothesis analogously. It will be:

> *The set of causes relevant to determining social content is that set that has the highest statistical dependence with the representations of socially relevant observers under all stimulus conditions.*

Of course, determining who are the socially relevant observers will not be easy, but this is just the problem of determining social norms. And that, of course, is a problem that I can ignore (since whatever the definition, this theory applies), as I noted earlier.

Consider the case in which I can't tell the difference between dogs and cats (say I call them all dogs). Then, I don't personally misrepresent when I call cats dogs, since my statistical dependence under all conditions matches that under these conditions. However, I do socially misrepresent when I call cats dogs because socially

relevant observers (namely everyone else) have a different statistical dependency under all stimulus conditions than I do.[4]

> The generalization to absolute misrepresentation will be similar:
>
> *The set of causes relevant to determining absolute content is that set that has the highest statistical dependence with the representations of relevant possible observers under all stimulus conditions.*

The addition here of 'relevant possible observers' moves this kind of misrepresentation to the realm of the metaphysical. To understand this case, consider Twin Earth cases once again (see chapter 2). Suppose no one knows the microstructure of water (H_2O), or of twin water (XYZ). In this case, if a sample of XYZ made it to earth, and I represent it as water, then the only sense in which I am misrepresenting twin water is *absolutely*. That is, sometime in the future when chemistry is invented and we can look at molecular microstructure (i.e., for these *possible* observers), there will be observers that have statistical dependencies that contradict mine. The reason we accept these as being *relevant* in advance of their being realized is because these dependencies result in more *precise* and presumably more *accurate* categorizations than ours. This kind of dependency will only be 'absolute' in the metaphysical sense if we think that metaphysics can't go beyond empirical data (see e.g. Quine 1960).

In any case, notice that what each of these different versions of the statistical dependence hypothesis is doing is changing the standard for accuracy. Recall that accuracy is deviation from some 'correct' answer. The correct answer in each of these cases is different. For personal content, correctness depends on how well the individual *could* do with their current resources. For social content, correctness depends on what the socially relevant observers (experts, perhaps) determine is correct. And, for absolute content, correctness depends on what really is the case (or perhaps what the case is that we would *ever* determine given any resources).

These differences in standards mean that we should *allow* some kinds of disjunctions. I have already explained how disjunctive representation can be avoided at the personal level on this theory. However, I then provided an example in which I couldn't tell the difference between dogs and cats. That is, I provided an example in which my representation was of the disjunction dog-or-cat. So, are disjunctions permitted as referents on this account or not? Well, some are and some aren't.

It is important to realize that 'disjunction' is a *formal* notion. That is, the notion doesn't apply to physical objects, it applies to linguistic expressions. So to say that my representation is of dogs-or-cats is really just to say that that set of objects is picked out by some language using the disjunction of the terms 'dog' and 'cat'. But, in my representational scheme, that same set of objects is picked out by the term 'dog' alone, so it isn't a disjunctive concept in that representational scheme. So, if one language (say personal or social) can only make co-extensive representations with another (say social or absolute) if the second language uses disjunctions to describe referents of the first, we shouldn't worry. We shouldn't worry because we need *other reasons* to say that one of these languages has precedence to carving up the world over the other. These other reasons don't matter for what representation relations hold within *one particular* representational scheme.

Disjunctions of these kinds are perfectly fine for two reasons. First, these disjunctions are innocent from an 'epistemic responsibility' point of view. The subject (or group) *couldn't do better* at representing the world given current resources. We wouldn't hold someone responsible for a misrepresentation if they couldn't help but misrepresent (relative to some other language).[5] Second, these kinds of disjunctions are innocuous because they aren't really the kind that caused the problem in the first place. The disjunctions that were problematic were those that challenged a causal theory of content. These disjunctions were within the same representational scheme that *could* make the relevant distinctions. In particular, I needed some way to show why the vehicle 'dog' could have a referent of a cat and still be about dogs (when I can, presumably, distinguish dogs and cats). Well, I have already provided the explanation we need in terms of statistical dependencies. In fact, I provided an explanation

---

[4] To be even more precise, we should distinguish conceptual social misrepresentation from occurent social misrepresentation (personal misrepresentation obviously doesn't need this distinction). I take it, however, that conceptual social misrepresentation is what is of the most interest.

[5] Generally an agent isn't considered responsible for something they can't help. In fact, there are laws to this effect in many countries. Of course, if we think the agent is in this 'helpless' state due to negligence, responsibility ascriptions may once again be made.

for this situation no matter how it is interpreted (i.e., whether we are talking personally, socially, or absolutely). I take it that I have thus solved the disjunction problem with this theory of content.

The second kind of misrepresentation that Dretske (1995) identifies is the representation of some object as having a property (say blackness) even though there is *no object* that has the property. The typical example of this kind of misrepresentation is a hallucination. If I 'see' a dog where there is none, I am hallucinating that dog, and assigning properties to something non-existent. The natural way of phrasing this problem in the context of the theory presented here is to say that I am representing something that has no referent. But, this formulation makes clear the strangeness of positing this kind of misrepresentation for a causal theory. If representations are supposed to be caused, how can there be no cause of a representation as supposed in the hallucination example? Well, of course there is *some* cause. We couldn't be naturalists and think otherwise.

In fact, exactly the same kind of story can be told for hallucinations as for misrepresentations generally. The difference is simply that, in the case of hallucinations, one of the (many) properties that are wrong about the referent is its location at a particular point in visual space. The referent will still be the thing in the world that caused the hallucination; it will just be that the properties ascribed are *very* wrong. The representation, then, is a highly imprecise and non-accurate one, but it's not one that is *not caused* or otherwise needs some special explanation.

Lastly, consider the Millikanian problem of correct representations being rare. This is an important problem because it clearly might be adaptively advantageous to adopt such a representational strategy. If predators were *really important* to avoid, then lots of false positives (i.e., representings of predators when there are none) would be advantageous. How can a theory that relies on *highest statistical dependencies* account for such situations? This is where it is important that what is highest are *statistical dependencies*. Consider a case in which a bird represents a cat (its predator) (see also Usher unpublished). Let us presume that the bird classifies its world into cats and non-cats. In this case, if the probability that 'cat' is tokened given a cat in the environment is greater than the probability that 'cat' is tokened given a non-cat in the environment, then the statistical dependency between 'cat' and a cat in the environment is greater than that between 'cat' and a non-cat in the environment (see appendix A). This is true even if cats are very uncommon and non-cats are very common. This is also true *even if the probability of calling a non-cat a cat is greater than the probability of calling a non-cat a non-cat.* In other words, statistical dependencies hold, as we would expect, even in the face of many false positives.

So, the theory of content I have presented can solve each of these problems of misrepresentation. The disjunction problem is accounted for by both a careful analysis of kinds of misrepresentation, and the ability of the theory to avoid 'within language' disjunctions. The problem of absent referents is explained away by appealing to degrees of misrepresentation. Lastly, the problem of accounting for many false positives is taken care of directly by the theory's reliance on statistical dependencies.

5 Conclusion

That concludes my defense of this theory and, along with it, my project in general. I take it that I have shown how to avoid the problems of current contemporary theories of content that I discussed in chapter 2. As well, I have answered the thirteen questions about content that I argued must be answered by a satisfactory theory of content (chapter 3). Presumably, answering these questions, and showing why the answers should be convincing, provides a good solution to the problem of neurosemantics I set out in chapter 1. Nevertheless, the theory of content I have spent the last four chapters describing is woefully incomplete. While I have insisted on providing some detail, there is still much that needs to be explored further. In particular, neuroscientists and psychologists have the momentous task of deciding which are the right vehicles to ascribe to a given neurobiological system. Philosophers must decide if the theoretical consequences of such a theory are viable ones; the implications of this theory for folk psychology, theories of concepts, and, of course, consciousness are entirely unclear given my discussion. I'm not sure which job is the greater.

APPENDIX A

Statistical Dependence and False Positives

The following shows that the statistical dependency of tokenings of 'cat' with cats is always greater than tokenings of 'cat' with non-cats if and only if the probability of 'cat' given a cat is greater than 'cat' given a non-cat.

Let

$x = P(\text{cat}_{\text{token}}|\text{cat})$,

$y=P(\text{cat}_{\text{token}}|\text{non-cat})$, and

$a=P(\text{cat})$.

Then,

$P(\text{cat}_{\text{token}}) = (x+y)a$ and

$P(\text{non-cat}_{\text{token}}) = 2a-(x+y)a$.

This results in the following four statistical dependencies as defined by mutual information ($M(A,B) = P(A|B)/P(A)$):

$M(\text{cat}, \text{cat}_{\text{token}}) = x/(x+y)a$;

$M(\text{cat}, \text{non-cat}_{\text{token}}) = (1-x)/(2a-(x+y)a)$;

$M(\text{non-cat}, \text{cat}_{\text{token}}) = y/(x+y)a$; and

$M(\text{non-cat}, \text{non-cat}_{\text{token}}) = (1-y)/(2a-(x+y)a)$.

In order to satisfy the statistical dependency hypothesis for the 'cat' representation, we want:

$M(\text{cat}, \text{cat}_{\text{token}}) - M(\text{cat}, \text{non-cat}_{\text{token}}) > 0$;

i.e., $x/(x+y)a - y/(x+y)a > 0$.

This is satisfied iff $x > y$. Notice also that:

$M(\text{non-cat}, \text{non-cat}_{\text{token}}) - M(\text{cat}, \text{non-cat}_{\text{token}}) > 0$ iff $x > y$.

References

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

Abbott, L. F., E. T. Rolls, et al. (1996). "Representational capacity of face coding in monkeys." *Cerebral Cortex* **6**(3): 498-505.

Akins, K. (1996). "Of sensory systems and the "aboutness" of mental states." *Journal of Philosophy*: 337-72.

Andersen, R. A. and D. Zipser (1988). "The role of the posterior parietal cortex in coordinate transformations for visual-motor integration." *Can J Physiol Pharmacol* **66**(4): 488-501.

Anderson, C. H. (1994). "Basic elements of biological computational systems." *International Journal of Modern Physics* **5**(2): 135-7.

Anderson, C. H. (1998). "Modeling population codes". *Computational Neuroscience 98 (CNS *98)*, Santa Barbara, CA, Elsevier.

Anderson, J. R. (1978). "Arguments concerning representations for mental imagery." *Psychological Review* **85**: 249-277.

Anscombe, G. E. M. (1993). Causality and determinism. *Causation*. E. Sosa and M. Tooley. Oxford, UK, Oxford University Press.

Aronson, J. (1971). "The legacy of Hume's analysis of causation." *Studies in the History and Philosophy of Science* **7**: 135-6.

Atherton, M. (in press). Instigators of the sensation/perception distinction. *Perception theory: conceptual issues*. R. Mausfeld and D. Heyer, John Wiley and Sons.

Barber, M. (1999). Studies in neural networks: Neural belief networks and synapse elimination. *Physics*. St. Louis, Washington University.

Bechtel, W. (1986). Teleological functional analyses: Hierarchical organization of nature. *Current issues in teleology*. W. Rescher. Lanham, University Press of America.

Bechtel, W. and R. C. Richardson (1993). *Discovering complexity: decomposition and localization as strategies in scientific research*. Princeton, NJ, Princeton University Press.

Bialek, W. and F. Rieke (1992). "Reliability and information transmission in spiking neurons." *Trends in Neurosciences* **15**(11): 428-434.

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

Binder, J. R., J. A. Frost, et al. (1997). "Human brain language areas identified by functional magnetic resonance imaging." *The Journal of Neuroscience* **17**(1): 353-362.

Bishop, C. (1995). *Neural networks for pattern recognition*. Oxford, UK, Oxford University Press.

Block, N. (1986). Advertisement for a semantics for psychology. *Midwest Studies in Philosophy*. P. French, T. Uehling and H. Wettstein. Minneapolis, University of Minnesota Press. **X:** 615-678.

BonJour, L. (1985). The structure of empirical knowledge. Cambridge, MA, Harvard University Press.

Booth, M. C. A. and E. T. Rolls (1998). "View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex." *Cerebral Cortex* **8**: 510-523.

Bower, J. M., Ed. (1998). *Computational neuroscience: Trends in research 1998*, Elsevier.

Brentano, F. (1874). *Psychology from an empirical standpoint*, Routledge & Kegan Paul.

Burge, T. (1979). Individualism and the mental. *Midwest Studies in Philosophy IV: Studies in Metaphysics*. P. e. a. French. Minneapolis, University of Minnesota Press.

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

Castaneda, H. (1984). Causes, causity, and energy. *Midwest studies in philosophy*. P. French, T. Uehling Jr. and H. Wettstein. Notre Dame, IN, University of Notre Dame Press. **9**.

Chisholm, R. (1955). "Sentences about believing." *Proceedings of the Aristotelian Society* **56**.

Chisholm, R. (1957). *Perceiving: A philosophical study*. Ithaca, NY, Cornell University Press.

Chomsky, N. (1986). *Knowledge of language*. New York, NY, Praeger.

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

Churchland, P. (1981). "Eliminative materialism and the propositional attitudes." *Journal of Philosophy* **78**: 67-90.

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

Churchland, P. M. and P. S. Churchland (1990). "Could a machine think?" *Scientific American* **262**(1): 3207.

Churchland, P. S. (1993). "Presidential address: Can neurobiology teach us anything about consciousness?". *American Psychological Association*.

Churchland, P. S. and T. Sejnowski (1992). *The computational brain*. Cambridge, MA, MIT Press.

Crick, F. and C. Koch (1998). "Consciousness and neuroscience." *Cerebral Cortex* **8**: 97-107.

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

Cummins, R. (1996). *Representations, targets, and attitudes*. Cambridge, MA, MIT Press.

Das, A. and C. D. Gilbert (1999). "Topography of contextual modulation mediated by short-range interactions in primary visual cortex." *Nature* **399**: 655-661.

Davidson, D. (1987). "Knowing one's own mind." *Proceedings and Addresses of the American Philosophical Association* **60**(441-458).

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. (1987). *The intentional stance*. Cambridge, MA, MIT Press.

Dennett, D. (1995). "The unimagined preposterousness of zombies: Commentary on T. Moody, O. Flanagan and T. Polger." *Journal of Consciousness Studies* **2**(4): 322-326.

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

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

Descartes, R. (1641/1955). *The philosophical works of Descartes*, Dover Publications.

Descartes, R. (1641/1955). *The philosophical works of Descartes*, Dover Publications.

Descartes, R. (1641/1990). *Meditations on first philosophy*. Notre Dame, IN, University of Notre Dame Press.

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.

DeYoe, E. A. and D. C. Van Essen (1988). "Concurrent processing streams in monkey visual cortex." *Trends in Neuroscience* **11**: 219-226.

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

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

Dretske, F. (1994). If you can't make one, you don't know how it works. *Midwest Studies in Philosophy*. P. French, T. Uehling and H. Wettstein. Minneapolis, University of Minnesota Press. **XIX:** 615-678.

Dretske, F. (1995). *Naturalizing the Mind*. Cambridge, MA, MIT Press.

Dretske, F. and A. Snyder (1972). “Causal irregularity.” *Philosophy of Science* **39**: 69-71.

Dummett, M. (1978). *Truth and other enigmas*. Cambridge, Harvard University Press.

Eccles, J. C. (1974). Cerebral activity and consciousness. *Studies in the Philosophy of*

Biology. F. Ayala and T. Dobzhansky, University of California Press.

Einstein, A. (1961). *Relativity: The special and the general theory*. New York, Crown.

Eliasmith, C. (1996). “The third contender: a critical examination of the dynamicist theory of cognition.” *Philosophical Psychology* **9**(4): 441-463.

Eliasmith, C. (1997). “Computation and dynamical models of mind.” *Minds and Machines* **7**: 531-541.

Eliasmith, C. (in press). “Is the brain analog or digital?: The solution and its consequences for cognitive science.” *Cognitive Science Quarterly*.

Eliasmith, C. and C. H. Anderson (1999). “Developing and applying a toolkit from a general neurocomputational framework.” *Neurocomputing* **26**: 1013-1018.

Eliasmith, C. and C. H. Anderson (forthcoming). *Neural engineering: The principles of neurobiological simulation*. Cambridge, MA, MIT Press.

Eliasmith, C. and C. H. Anderson (in press). "Rethinking Central Pattern Generators: A General Framework". *Computational Neuroscience 99 (CNS *99)*, Pittsburgh, PA, Elsevier.

Eliasmith, C. and P. Thagard (in press). “Integrating structure and meaning: A distributed model of analogical mapping.” *Cognitive Science*.

Evans, G. (1982). *Varieties of reference*. New York, Oxford University Press.

Everling, S., M. C. Dorris, et al. (1999). “Role of primate superior colliculus in preparation and execution of anti-saccades and pro-saccades.” *The Journal of Neuroscience* **19**(7): 2740-2754.

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.

Felleman, D. J., Y. Xiao, et al. (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.

Field, H. (1977). “Logic, meaning, and conceptual role.” *Journal of Philosophy* **74**: 379-409.

Fine, K. (1975). “Vagueness, truth and logic.” *Synthese* **30**: 265-300.

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

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

Fodor, J. (1987). *Psychosemantics*. Cambridge, MA, MIT Press.

Fodor, J. (1990). *A theory of content and other essays*. Cambridge, MA, MIT Press.

Fodor, J. (1995). “West coast fuzzy: Why we don't know how brains work (review of Paul Churchland's The engine of reason, the seat of the soul).” *The Times Literary Supplement*(August).

Fodor, J. (1998). *Concepts: Where cognitive science went wrong*. New York, Oxford University Press.

Fodor, J. (1998). Review of Paul Churchland's The engine of reason, the seat of the soul. *In critical condition: Polemical essays on cognitive science and the philosophy of mind*. J. Fodor. Cambridge, MA, MIT Press.

Fodor, J. (1999). “Diary.” *London Review of Books* **21**(19).

Fodor, J. and E. Lepore (1992). *Holism: A shopper's guide*. Oxford, UK, Basil Blackwell.

Fodor, J. A. (1994). *The elm and the expert*. Cambridge, MA, MIT Press.

Frege, G. (1892/1980). On sense and meaning. *Translations from the philosophical writings of Gottlob Frege*. P. Geach and M. Black. Oxford, UK, Basil Blackwell.

Fuhrmann, G. (1988). "Fuzziness of concepts and concepts of fuzziness." *Synthese* **75**: 349-72.

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

Goodale, M. A. and A. D. Milner (1992). "Separate pathways for perception and action." *Trends in Neuroscience* **15**: 20-25.

Goodman, N. (1955). *Fact, fiction and forecast*. Indianapolis, Bobbs-Merrill.

Goodman, N. (1968). *Languages of art*. Indianapolis, IN, Hackett Publishing Company.

Grice, P. (1957). "Meaning." *Philosophical Review* **66**: 377-388.

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.

Grush, R. (1997). "The architecture of representation." *Philosophical Psychology* **10**(1): 5-23.

Haack, S. (1993). *Evidence and inquiry: Toward reconstruction in epistemology*. Cambridge, MA, Blackwell Publishers.

Hakimian, S., C. H. Anderson, et al. (1999). "A PDF model of populations of purkinje cells." *Neurocomputing* **26**.

Hammerstrom, D. (1995). Digital VLSI for neural networks. *The handbook of brain theory and neural networks*. M. Arbib. Cambridge, MA, MIT Press.

Harman, G. (1982). "Conceptual role semantics." *Notre Dame Journal of Formal Logic* **23**: 242-56.

Harman, G. (1987). (Nonsolopsistic) conceptual role semantics. *Semantics of natural language*. E. LePore. New York, Academic Press**:** 55-81.

Haugeland, J. (1991). Representational genera. *Philosophy and Connectionist Theory*. W. Ramsey, S. Stich and D. Rumelhart. Hillsdale, NJ, Lawrence Erlbaum.

Henneman, E. and L. Mendell (1981). Functional organization of motoneuron pool and its inputs. *Handbook of physiology :The nervous system*. V. B. Brooks. Bethesda, MD, American Physiological Society. **2**.

Hofstadter, D. and D. Dennett, Eds. (1981). *The mind's I*. New York, Basic Books.

Hume, D., Ed. (1739/1886). *A treatise of human nature*. Darmstadt, Scientia Verlag Aalen.

Hutto, D. D. (1999). *The presence of mind*. Philadelphia, J. Benjamins Publishers.

Hyvarinen, A. (1999). "Survey on independent component analysis." *Neural Computing Surveys* **2**: 94-128.

Jackson, F. (1986). "What Mary didn't know." *Journal of Philosophy* **83**: 291-5.

Johnson, D. A. (1995). Grue paradox. *The Cambridge dictionary of philosophy*. R. Audi, Cambridge University Press.

Kandel, E., J. H. Schwartz, et al., Eds. (1991). *Principles of neural science*. New York, NY, Elsevier Science Publishing.

Kant, I. (1787/1965). *Critique of pure reason*. London, MacMillan.

Karni, A., G. Meyer, et al. (1995). "Functional MRI evidence for adult motor cortex plasticity during motor skill learning." *Nature* **377**: 155-8.

Kim, J. (1995). Causation. *The Cambridge dictionary of philosophy*. R. Audi. New York, NY, Cambridge University Press.

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.

Koch, C. (1998). *Biophysics of computation: Information processing in single neurons*. Oxford, UK, Oxford University Press.

Kosslyn, S. (1994). *Image and brain: The resolution of the imagery debate*. Cambridge, MA, The MIT Press.

Kosslyn, S. M. and W. L. Thompson (1999). Shared mechanisms in visual imagery and visual perception: Insights from cognitive neuroscience. *Handbook of cognitive neuroscience*. M. S. Gazzaniga. Cambridge, MA, MIT Press.

Kosslyn, S. M., W. L. Thompson, et al. (1995). "Topographical representations of mental images in primary visual cortex." *Nature* **378**: 496-498.

Kripke, S. (1977). Speaker's reference and semantic reference. *Midwest studies in philosophy*. P. French, T. Uehling Jr. and H. Wettstein. Notre Dame, IN. **2**.

Langacker, R. W. (1987). *Foundations of cognitive grammar*. Stanford, CA, Stanford University Press.

Lass, Y. and M. Abeles (1975). "Transmission of information by the axon. I: Noise and memroy in the myelinated nerve fiber of the frog." *Biological Cybernetics* **19**: 61-67.

Lehrer, K. (1974). *Knowledge*. Oxford, Clarendon Press.

Lepore, E. (1994). Conceptual role semantics. *A companion to the philosophy of mind*. S. Guttenplan. Oxford, UK, Basil Blackwell.

Lewicki, M. and B. Olshausen (in press). "A probabilistic framework for the adaption and comparison of image codes." *Journal of the Optical Society of America*.

Lewicki, M. and T. Sejnowski (1998). "Learning overcomplete representations." *Neural Computing*.

Loar, B. (1981). *Mind and meaning*. London, UK, Cambridge University Press.

Locke, J. (1700/1975). *An essay concerning human understanding*. Oxford, UK, Oxford University Press.

Long, A. A. and D. N. Sedley (1987). *The Hellenistic philosophers*. Cambridge, Cambridge University Press.

Lycan, W. (1984). *Logical form in natural language*. Cambridge, MA, MIT Press.

Mackie, J. (1974). *The cement of the universe*. Oxford, Clarendon Press.

Mandik, P. (1999). "Qualia, space and control." *Philosophical Psychology* **12**(1): 47-60.

Miller, J., G. A. Jacobs, et al. (1991). "Representation of sensory information in the cricket cercal sensory system. I: Response properties of the primary interneurons." *Journal of Neurophysiology* **66**: 1680-1703.

Millikan, R. (1993). *White queen psychology and other essays for alice*. Cambridge, MA, MIT Press.

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

Mohana, T. and L. Wee (1999). *Grammatical semantics: Evidence for structure in meaning*. Stanford, CA, CSLI Publications.

Moschovakis, A. K., C. A. Scudder, et al. (1996). "The microscopic anatomy and physiology of the mamallian saccadic system." *Progress in Neurobiology* **50**(2): 133-254.

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

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA, Harvard University Press.

Nicholls, J. G., A. R. Martin, et al. (1992). *From neuron to brain*. Sunderland, MA, Sinauer Associates Inc.

Nova (1997). "Secret of the wild child." *Public Broadcasting Service* **#2112G**(March 4, 1997).

Ojemann, G. A. and J. Schoenfield-McNeill (1999). "Activity of neurons in human temporal cortex during identification and memory for names and words." *The Journal of Neuroscience* **19**(13): 5674-5682.

Olshausen, B. and D. Field (1996). "Emergence of simple-cell receptive field properties by learning a sparse code for natural images." *Nature* **381**: 6-7-609.

Parker, A. J. and W. T. Newsome (1998). "Sense and the single neuron: Probing the physiology of perception." *Annual Review of Neuroscience* **21**: 227-277.

Peacocke, C. (1986). *Thoughts: An essay on content.* Oxford, UK, Basil Blackwell.

Place, U. T. (1959). "Is consciousness a brain process?" *British Journal of Psychology* **47**.

Prigogine, I. (1996). *The end of certainty*. New York, NY, The Free Press.

Putnam, H. (1975). The meaning of 'meaning'. *Mind, language, and reality*, Cambridge University Press**:** 215-71.

Pylyshyn, Z. (1973). "What the mind's eye tells the mind's brain: A critique of mental imagery." *Psychological Bulletin* **80**: 1-24.

Quine, W. V. (1981). *Theories and things*. Cambridge, MA, Harvard University Press.

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

Quine, W. V. O. (1969). *Ontological relativity and other essays*. New York, Columbia University Press.

Quine, W. V. O. and J. Ullian (1970). *The web of belief*. New York, Random House.

Rao, R. and D. Ballard (1995). "An active vision architecture based on iconic representations." *Artificial Intelligence Journal* **78**: 461-505.

Rauschecker, J. P. (1999). "Auditory cortical plasticity: A comparison with other sensory systems." *Trends in Neuroscience* **22**: 74-80.

Reza, F. M. (1994). *An introduction to information theory*. New York, Dover.

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

Saleem, K. S. and K. Tanaka (1996). "Divergent projections from the anterior inferotemporal area TE to the perirhinal and entorhinal cortices in the macaque monkey." *The Journal of Neuroscience* **16**(15): 4757-4775.

Salinas, E. and L. Abbott (1994). "Vector reconstruction from firing rates." *Journal of Computational Neuroscience* **1**: 89-107.

Sarris, V. (1989). "Max Wertheimer on seen motion: Theory and evidence." *Psychological Research* **51**: 58-68.

Searle, J. (1992). *The rediscovery of the mind*. Cambridge, MA, MIT Press.

Seung (1996). "How the brain keeps the eyes still". *National Academy of Science USA, Neurobiology*.

Shannon, C. (1948/1949). A mathematical theory of communication. *The mathematical theory of communication*. C. Shannon and W. Weaver. Urbana, IL, University of Illinois Press**:** 623-656.

Smith, E. (1989). Concepts and induction. *Foundations of cognitive science*. M. Posner. Cambridge, MA, MIT Press**:** 501-526.

Sosa, E. and M. Tooley, Eds. (1993). *Causation*. Oxford, UK, Oxford University Press.

Stevens, C. F. and Y. Wang (1994). "Changes in reliability of synaptic function as a mechanism for plasticity." *Nature* **371**: 704-707.

Strawson, G. (1987). "Realism and causation." *Philosophical Quarterly* **37**: 253-77.

Thagard, P. (1986). "The emergence of meaning: how to escape Searle's Chinese room." *Behaviorism* **14**: 139-146.

Thagard, P. (1992). *Conceptual revolutions*. Princeton, Princeton University Press.

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.

Turing, A. M. (1950). "Computing machinery and intelligence." *Mind* **59**: 433-460.

Uhr, L. (1994). Digital and analog microcircuit and sub-net structures for connectionist networks. *Artificial intelligence and neural networks: Steps toward principled integration*. V. Honavar and L. Uhr. Boston, MA, Academic Press**:** 341-370.

Ungerleider, L. G. and M. Mishkin (1982). Two cortical visual systems. *Analysis of visual behavior*. D. J. Ingle, M. A. Goodale and R. J. W. Mansfield. Boston, NY, MIT Press**:** 549-586.

Usher, M. (unpublished). "Conceptual representations need probabilistic categorization: an informational-theoretical approach to representation and misrepresentation." .

Van Essen, D. and J. Gallant (1994). "Neural mechanisms of form and motion processing in the primate visual system." *Neuron* **13**: 1-10.

Van Essen, D. C. and C. H. Anderson (1995). Information processing strategies and pathways in the primate visual system. *An introduction to neural and electron networks*, Academic Press.

van Gelder, T. (1995). "What might cognition be, if not computation?" *The Journal of Philosophy* **XCI**(7): 345-381.

van Gelder, T. and R. Port (1995). It's about time: An overview of the dynamical approach to cognition. *Mind as motion: Explorations in the dynamics of cognition*. R. Port and T. van Gelder. Cambridge, MA, MIT Press.

Wall, J. T., J. H. Kaas, et al. (1986). "Functional reorganization in somatosensory cortical areas 3b and 1 of adult monkeys after median nerve repair: possible relationships to sensory recovery in humans." *J Neurosci*(1): 218-33.

Warland, D., M. Landolfa, et al. (1992). Reading between the spikes in the cercal filiform hair receptors of the cricket. *Analysis and modeling of neural systems*. F. Eeckman. Boston, MA, Kluwer Academic Publishers.

Williamson, T. (1994). *Vagueness*. London, UK, Routledge.

Wilson, M. A. and B. McNaughton (1993). "Dynamics of the hippocampal ensemble code for space." *Science* **261**: 1055-58.

Wimsatt, W. C. (1980). Reductionist research strategies and their biases in the units of selection controversy. *Scientific discovery: Case studies*. T. Nickles, D. Reidel Publishing Company**:** 213-259.

Yolton, J. W. (1993). *A Locke dictionary*. Oxford, UK, Blackwell.

Zeilinger, A. (2000). "Quantum teleportation." *Scientifica American* **282**(4): 50-59.

Zipser, D. and R. A. Andersen (1988). "A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons." *Nature* **331**: 679-84.