An Introduction to Contemporary Theories of Content

Chris Eliasmith
University of Waterloo

> *Do not believe in anything merely on the authority of your teachers and elders.* – Buddha (Anguttara Nikaya, Tika Nipata, Mahavagga, Sutta No. 65)

1 Introduction

This paper is intended to introduce advanced students of philosophy of mind to central theories and concepts that have been used to characterize mental meaning, content, and representation. I initially provide relevant historical background. Then I discuss the ubiquitous sense/reference distinction suggested by Frege. Finally, I outline the three major kinds of contemporary theories of mental representation and discuss the difficulties with each. Much of this material has been extracted from Eliasmith (2000).

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. Here, I examine a small subset of these attempts, inspired mainly by work in philosophy, psychology, or neuroscience. 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).[1]

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

---

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

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.[2] 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

---

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

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 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.

3 Mental content

In the last twenty or so years, there have been a plethora of philosophical theories trying to answer questions about mental content. 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). In this section I provide a brief characterization of what content *is*, such that it is assignable not only to psychological states (or concepts), but to neurons as well.

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 *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 standard claim would be that they *still* have different '*senses*' – which seems incongruous with the claim that content is property ascription. 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 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.

4 Language and meaning

Notably, the term 'content' is often reserved solely for language, or language-like mental structures. Indeed, 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.[3] Furthermore, there *is* clear evidence that language is *not necessary* for content.[4] 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,

---

[3] 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.

[4] 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

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

---

yet not have content. Many others disagree (Turing 1950; Hofstadter and Dennett 1981; Thagard 1986; Churchland and Churchland 1990).

majority. Starting at the "neuron-by-neuron" level avoids such difficulties. We have quite good 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 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).

5 Contemporary theories of content

In the remainder of this paper I survey a variety of theories of content proposed by contemporary philosophers. I 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 theory 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.

5.1 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.

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).

5.2 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.[5] 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.

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

---

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

more concerned about the viability of diachronic theories than with the applicability of intuitions to fantastic situations.

5.3 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.

5.4 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.[6] 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

---

[6] 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).

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.

5.5 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.

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.

5.6 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.

6 Conclusion

I have surveyed the history, and conceptual apparatus of 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.

7 References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

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.

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

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

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

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.

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

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

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

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

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

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

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

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