A Dynamic Account of the Structure of Concepts

by

Peter Blouw

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Arts
in
Philosophy

Waterloo, Ontario, Canada, 2011

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Concepts are widely agreed to be the basic constituents of thought. Amongst philosophers and psychologists, however, the question of how concepts are structured has been a longstanding problem and a locus of disagreement. I draw on recent work describing how representational content is ascribed to populations of neurons to develop a novel solution to this problem.

Because disputes over the structure of concepts often reflect divergent explanatory goals, I begin by arguing for a set of six criteria that a good theory ought to accommodate. These criteria address philosophical concerns related to content, reference, scope, publicity, and compositionality, and psychological concerns related to categorization phenomena and neural plausibility.

Next, I evaluate a number of existing theoretical approaches in relation to these six criteria. I consider classical views that identify concepts with definitions, similarity-based views that identify concepts with prototypes or exemplars, theory-based views that identify concepts with explanatory schemas, and atomistic views that identify concepts with unstructured mental symbols that enter into law-like relations with their referents. I conclude that none of these accounts can satisfactorily accommodate all of the criteria.

I then describe the theory of representational content that I employ to motivate a novel account of concept structure. I briefly defend this theory against competitors, and I describe how it can be scaled from the level of basic perceptual representations to the level of highly complex conceptual representations. On the basis of this description, I contend that concepts are structured dynamically through sets of transformations of single source representation, and that the content of a given concept specifies the set of potential transformations it can enter into. I conclude by demonstrating that the ability of this account to meet all of the criteria introduced beforehand. I consider objections to my views throughout.

Acknowledgements

First, I owe a great deal of thanks to my supervisor, Chris Eliasmith, whose guidance and expertise have made the process of writing this thesis both enjoyable and highly instructive. I am grateful to Chris for introducing me to such an interesting area of research.

Thanks are also due to the readers of this thesis, Paul Thagard and Tim Kenyon, whose comments and suggestions have resulted in an improved final product and provided me with a number of pointers for future work.

Finally, I wish to thank both my family and my friends. In particular, I am grateful to my fellow graduate students for countless interesting philosophical discussions, and for making my experience in the MA program such a positive one.

Table of Contents

Chapter One

Concepts, Cognitive Science, and Explanatory Goals

1.1 Introduction

By common agreement in philosophy and psychology, concepts are the basic building blocks of thought (Murphy, 2002; Laurence & Margolis, 1999). They play a foundational role in almost all explanations of the nature of our mental lives, and they are invoked ubiquitously in discussions of a broad range of psychological phenomena including categorization, language use, learning, and inference. In addition to their use in explaining mental phenomena, concepts have also been a focus of philosophical analysis since antiquity, with scholars ranging from Plato to Gettier preoccupying themselves with the clarification of concepts like JUSTICE and KNOWLEDGE.[1] Accordingly, the stakes in both philosophy and the cognitive sciences are high when it comes to the correct characterization of such things as the ontology, structure, and origin of concepts.

It is likely for this reason that concept research has attracted so much attention and controversy (Laurence & Margolis, 1999). The debates are complex. With respect to the ontology of concepts, scholars have argued alternately that concepts are mental representations, cognitive abilities, and abstract objects (Margolis & Laurence, 2005). With respect to the structure of concepts, treatments are similarly diverse: classical views identify concepts with definitions (e.g. Peacocke, 1992); similarity-based views identify concepts either with prototypes or sets of exemplars (e.g. Smith & Medin, 1981); theory-based views identify concepts with coherent sets of explanatory beliefs about the categories they represent

[1] Small capital letters are used to refer to concepts throughout.

(e.g. Medin & Murphy, 1985); and atomistic views identify concepts with semantically unstructured mental symbols that enter into law-like relations with things in the world (e.g. Fodor, 1998). Finally, with respect to the origin of concepts, opinion is divided between empiricism (whose proponents claim to varying degrees that concepts are derived from perceptual experiences) and nativism (whose proponents claim to varying degrees that concepts are innate). It is helpful to recognize that a commitment in one realm of these debates often entails commitments in another. For example, the view that concepts are cognitive abilities is arguably incompatible with the view that concepts are structured as prototypes (and hence are representational).

Of these three broad topics in concept research, the problem of conceptual structure is arguably the most theoretically pressing. For one thing, there is a fairly general consensus amongst cognitive scientists and philosophers of mind that concepts are mental representations (Margolis & Laurence, 2005), and so I accordingly assume as much. For another, the question of whether empiricism or nativism is correct will in some sense depend upon how one addresses the problem of structure. To explain with an example, if perceptual representations figure in the best available theory of conceptual structure, then one will have reason to be suspicious of strong nativist claims, all else being equal. Another way to put this point is to say that by providing a thorough discussion of structure, one will likely either end up defending certain views about concept acquisition at the same time, or end up remaining neutral on the matter. Thus, giving primacy to the problem of structure is advantageous from the perspective of efficiency and presents no immediate drawbacks.

More importantly, by focusing on the problem of conceptual structure and assuming that concepts are representational, one can develop a highly useful strategy for adjudicating amongst different characterizations of concepts. Put simply, if concepts are mental representations, and mental representations are instantiated in neurobiological systems,[2] then one should expect a good theory of concepts to be consistent with a successful account of how representational content is ascribed to the different parts of such systems. In other words, a theory of concepts that is consistent with a theory of content responsive to both philosophical and neuroscientific considerations ought to be preferred over a theory of concepts that is not, all else being equal. My purpose, accordingly, is to employ an existing, neurocomputationally grounded two-factor theory of content to motivate an account that identifies concepts with a specific type of content-bearing neural vehicle. This account builds on recent but distinct theories offered by Barsalou (1999, 2003, 2008), Eliasmith (2000, 2005 forthcoming), and Prinz (2002), and it essentially claims that concepts are structured *dynamically* through sets of transformations of a single source representation called a 'semantic pointer' (Eliasmith, forthcoming), whose content is specified holistically. Where the present account departs meaningfully from these precursors is in articulating a unique dynamic view of concept structure through a careful consideration of both philosophical concerns (e.g. reference determination) *and* neuroscientific concerns (e.g. neural plausibility, empirical adequacy).

[2] The idea that mental representations are instantiated in neurobiological systems is uncontroversial assuming that materialism is correct (which I do). This claim is neutral with respect to the specific nature of the relation between the mental and the physical (i.e. reduction, supervenience, etc.)

Before proceeding, it is worth providing a rough outline of what follows. First, I establish a set of criteria that a good theory of concepts must meet. Many of these will be standard fare for researchers familiar with the field, but some associated with the aforementioned neuroscientific concerns are novel. Second, I briefly evaluate existing theories of concepts and demonstrate their inability to satisfactorily meet these criteria. Third, I outline and briefly defend the theory of content being employed. Fourth, I draw from both this theory of content and the list of criteria to establish grounds for adopting the view of concept structure that I endorse. Various objections to my arguments are considered throughout.

1.2 Criteria for a Theory of Concepts

While concepts are almost universally agreed to be the constituents of thought, they are also theoretical postulates that serve an explanatory function. Disputes over the nature of concepts, as Laurence and Margolis (1999) note, often reflect disagreement over what sorts of things a theory of concepts ought to accomplish. Psychologists typically focus on the use of concepts to explain the ability to categorize things into different classes, while philosophers are often more concerned with things like the matter of how concepts refer. In what follows, I argue for a list of six explanatory targets that I think a good theory of concepts ought to accommodate. These are: scope, publicity, categorization, content, compositionality, and neural plausibility. Many of these so-called 'desiderata' are commonplace in the concept literature, but I occasionally argue for a unique understanding of the conditions under which a particular criterion would be satisfied. Others, such as the neural plausibility criterion, are not common in the existing literature.

1.2.1 Scope

As Jesse Prinz (2002) notes, there is a great deal of variety amongst the different types of concepts that human beings are capable of possessing. To give a rudimentary catalogue: there are concepts for perceivable objects, such as TABLE; there are abstract concepts, such as VIRTUE; there are concepts for theoretical posits, such as ELECTRON (Prinz, 2002); there are mathematical concepts, such as SUM; and there are non-referring concepts, such as CENTAUR. Ideally, a theory of concepts should be able account for all of these different classes. I take this to be an uncontroversial criterion given its widespread acceptance amongst both psychologists (e.g. Barsalou, 1999; Smith & Medin, 1981; Murphy, 2002) and philosophers (e.g. Prinz, 2002; Laurence & Margolis, 1999).

1.2.2 Publicity

The fact that people are able to communicate effectively has led to widespread agreement that concepts must be shareable. If concepts are the constituents of thought, and language is used to express thought, then our words and sentences must connect in some way to our conceptual representations (Murphy, 2002). So, because we understand and use the same words and sentences, we must also understand and use the same underlying concepts.[3] A second reason that the publicity requirement is so widely accepted has to do with the

[3] This sidesteps to some extent the issue of the precise relation between natural language and concepts. Some, like Fodor (1998), suggest that the meaning of a word is the concept it refers to, while others, like Prinz (2002), disagree. There is a further debate as to whether or not language is necessary for the possession of concepts (Margolis & Laurence, 2005). My intention is to remain neutral on these matters, since nothing in what follows hinges on them.

predictive adequacy of generalized intentional explanations of behavior. Such explanations are of the sort 'If *X* is thirsty, then *X* will seek water.' As Fodor (1998) notes, generalizations of this kind would be explanatorily idle if everyone possessed a different WATER concept. Thus, concepts like WATER must be public. Psychologists also acknowledge the importance of this criterion via discussions of 'concept stability' (e.g. Smith & Medin, 1981, p. 10)

Meeting the publicity requirement can occur in two ways. On the one hand, a theory of concepts can specify a type-token relation such that everyone who possesses a given concept (e.g. CIRCLE) literally tokens the same concept type (Fodor, 1998). This version of the criterion can be construed as requiring shared concepts to be strictly identical. On the other hand, a theory of concepts can specify a similarity relation such that shared concepts need only be alike to some minimal degree. Fodor (1998) rejects this view because he thinks that the required notion of similarity can only be construed in terms of partial identity between sets of structural components. This, he claims, does nothing to explicate a notion similarity that could apply to relations between unstructured mental symbols. However, so long as one does not presuppose that concepts are atomistic, understanding similarity in terms of partial identity between sets of structural components is not problematic. Thus, I take concept similarity across individuals to be sufficient for achieving publicity.

1.2.3 Categorization

Amongst psychologists, concepts are often defined primarily as mechanisms for classifying different entities and drawing inferences about their properties (e.g. Medin & Murphy, 1985; Smith & Kosslyn, 2007; Smith & Medin, 1981). Accordingly, one commonly

held explanatory goal for postulating concepts is to account for the reams of empirical data describing such classificatory and inferential behaviors. For instance, so-called 'typicality effects' manifest themselves when subjects indentify certain category instances as more typical than others (Rosch, 1978). Typical instances of a class are categorized more quickly than atypical instances, and features of typical instances are more readily matched and generated in category production tasks (Rosch, 1978).[4] Exemplar effects have also been demonstrated wherein atypical category instances that are highly similar to previously experienced instances get classified as rapidly or more quickly than typical instances (Prinz, 2002). Moreover, subjects have been shown to override false general category judgments like 'all birds are eagles' by producing counterexamples derived from prior experiences (Smith & Medin, 1981, p. 144). Finally, knowledge effects have been demonstrated in experiments where subjects are asked to categorize entities from one class that are transformed to look like members of a different class. For example, Keil (1989) shows that young children will identify horses disguised to look like prototypical zebras *as* horses based on knowledge of the disguise procedure. This research also suggests that the categorization of natural kinds is more susceptible to knowledge effects than the categorization of artificial kinds (Keil, 1989). Altogether, though, given that categorization tasks are fundamental to the use of concepts, explaining this empirical data ought to be a goal of any respectable theory.

1.2.4 Content

[4] Category production tasks are tasks where subjects are asked to produce features that are characteristic of members of a given category, or apply features to the correct category. Tasks involving the generation of category instances are also sometimes referred to as category production tasks.

The fact that mental states and mental representations are *about* things means that they have content or meaning (Brown, 2002). This content, in turn, can be roughly defined as what any given representation "tells you about what it represents" (Eliasmith, 2000, p. 7). And since concepts are representations, the content of a concept is likewise what a concept tells you about the things it represents. Two aspects of conceptual content become salient here: first, there is the matter of determining what a given concept refers to, and second, there is the matter of determining what properties the concept ascribes to these referents (Eliasmith, 2000). In Fregean terms, this means that every concept has both a 'reference' and a 'sense.'

The idea of that concepts refer is essentially what supports the use of concepts for categorization. In other words, the fact that concepts are used to classify different entities relies on the assumption that concepts are applicable to some things and not others. Accordingly, one reason for upholding the reference requirement for content is to accommodate the idea that categorization errors are possible (Laurence & Margolis, 1999). If concepts did not have their reference determined independently of a given individual's categorization behaviors, there would be no meaningful sense in which a person could misclassify, say, a vegetable as a fruit. Clearly, categorization mistakes of this sort do occur, and the best way of explaining this fact is to stipulate that there is a class of things that a concept is *taken* to apply to on a given occasion, and that there is a class of things that a concept *actually* applies to on all occasions.[5] For example, the concept WINE might be taken

[5] I'm not committing to the claim here that all concepts have rigidly demarcated extensions. Fuzzy boundaries are very likely to occur in some instances; indeed, some scholars argue that such fuzziness is *necessary* (e.g. Murphy, 2002, p. 21).

to apply to grape juice in a wine glass, but it actually applies only to certain alcoholic beverages. Insisting on a reference criterion merely amounts to insisting that genuine misrepresentation is possible, which is in no way controversial (see e.g. Adams & Aizawa, 2010; Fodor, 1987).

There is, however, some debate over whether the categories concepts refer to are naturally bounded, or bounded as the result of human cognitive processes (Prinz, 2002). Rosch (1978), for instance, claims that the human cognitive system latches onto "natural discontinuities" and co-occurring features in the perceivable world to form categories (p. 192). Yet at the same time, Rosch (1978) also notes that the saliency of different features is likely to depend on human functional requirements that may vary across cultures. This suggests a more relativistic stance on category formation. Overall, I take the reference criterion to minimally require that categories be non-solipsistic, and I wish to remain neutral on deeper metaphysical debates about the reality of natural kinds.

The second, property-related aspect of conceptual content emerges from Frege's (1892) classic observation that co-referring terms may nonetheless communicate distinct information. In the context of concepts, this means that co-extensive concepts can be distinct if they ascribe different sets of properties to the same referents. For example, the folk concept BIRD and the taxonomical concept AVES both refer to all and only birds, but they ascribe different properties. BIRD ascribes the properties we ordinarily take to be characteristic of birds, while AVES ascribes properties highlighting the category relations between birds and other classes of life forms. Because BIRD and AVES are clearly distinct concepts that possess identical referents, it is apparent that reference is not sufficient for determining content.

However, a close relative of this conclusion is also true. Putnam's (1975) Twin Earth thought experiment illustrates that distinct concepts can ascribe the same properties to different referents. Twin Earth is a near-identical counterpart to Earth that differs only in the fact that the material that fills the role of water is actually composed of some material XYZ instead of H_2O (Putnam, 1975). Hence, WATER on Earth refers to the chemical compound H_2O, while WATER_T on Twin Earth refers to some other compound XYZ. But because XYZ and H_2O are literally indistinguishable, it is clear that both WATER and WATER_T play identical roles and hence ascribe identical properties to different substances. The point to draw here is that property ascription alone is also insufficient for determining the content of a concept. Rather, reference and property ascription determine content together. One can generalize this conclusion by saying that co-extensive concepts that ascribe identical properties to their referents are identical.

Overall, the motives for proposing a content criterion should be obvious. The notion of conceptual content enables categorization errors and misrepresentations to occur. If concepts did not have clear referents, there would be no meaningful sense in which a concept could be misapplied. Furthermore, the notion of conceptual content explains cases in which different concepts ascribe distinct properties to the same thing, and cases in which different concepts apply the same properties to distinct things. Without this aspect of content, we would have to concede both that all co-referential concepts are identical, and that all concepts that ascribe the same properties are identical. This would be an unfortunate result. If it is agreed that the pairs AVES/BIRD and $\text{WATER}/\text{WATER}_T$ are not identical, then a satisfactory theory of concepts ought to entail as much.

1.2.5 Compositionality

The claim that concepts are compositional is essentially the conjunction of the claims that complex concepts are created by combining simple concepts through certain rules, and that the contents of complex concepts are derived from the contents of their constituent parts. Based on the preceding discussion of content, this means that complex concepts ascribe properties as some function of the way their constituent parts ascribe properties, and that complex concepts derive their extensions from the extensions of their constituent parts. In philosophical circles, the compositionality criterion has been championed by Jerry Fodor (1998), amongst others, but analogues of it are common in the psychological literature as well. Smith and Medin (1981), for instance, note that concepts function to combine in novel ways, while Murphy (2002) discusses a number of proposed models of such combination.

There are three primary reasons for imposing the compositionality criterion on a theory of concepts. The first reason is that compositionality explains the *productivity* of thought, or the ability for an infinite number of thoughts to be produced from a finite number of concepts or representations (Fodor, 1998). In linguistic terms, this can be illustrated by the empirical fact that people can understand and produce infinite numbers of entirely novel sentences (Aydede, 1998). Through inference to the best explanation, one can presume that this is because the meaning of such novel expressions derives from the meaning of their well-comprehended constituent parts (Szabo, 2004). Likewise for concepts. The explanation of the reason I can produce the novel concept FURRY SPOON is because I already possess the concepts FURRY and SPOON and can combine them.

The second reason for the compositionality criterion is that it explains the *systematicity* of thought. Thoughts are systematic insofar as the ability to token one particular thought corresponds to the ability to token others with related content (Fodor, 1998). To use the classic example, being able to token the complex thought JOHN LOVES MARY empirically entails the ability to token the thought MARY LOVES JOHN in most if not all humans (Fodor, 1987, p. 151). The reason compositionality explains this capacity is because the possession of combination rules and the concepts MARY, JOHN, and LOVES clearly allows for the ability to construct thoughts involving any number of different syntactic structures. In short, inference to the best explanation of systematicity favors a compositionality criterion (Szabo, 2004).

Lastly, system constraints also favor the compositionality criterion. If every complex thought had to be stored and tokened as a unique unstructured representation, then the brain would have to be able to store and token unique representations for *every single thought* a given person is capable of entertaining. Given that this number is bound to be inordinately large, in seems implausible to suppose that the brain has the storage capacity to independently represent each and every possible thought. For example, storing separate representations for each of MARY LOVES JOHN AND PAUL, MARY LOVES JOHN, MARY LOVES PAUL, JOHN LOVES PAUL AND MARY and so forth is much more cognitively expensive that storing four representations for JOHN, MARY, PAUL, and LOVES, plus a set of combination rules that apply to all representations. Overall, then, a compositionality criterion is favored by considerations of productivity, systematicity, and cognitive economy. It is thus no surprise that so many theorists think concepts must compositional.

Some caveats are in order before proceeding. First, I do not wish to commit to the idea that a simple concept always contributes the same content to every complex concept of which it is a part (see Aydede, 1998). One reason to resist this understanding of compositionality emerges from a consideration of thoughts involving analogical relations. Consider the concepts DOG SHOW and DOG TIRED. The former roughly means an event where people enter their canines into various competitions, while the latter roughly means 'tired like a sleepy pet dog.' And clearly, the concept DOG contributes different contents in each case: there is nothing related to the relative wakefulness of dogs contained in the concept DOG SHOW, and there is nothing related to the various tricks dogs can perform in the concept DOG TIRED. A second reason can be extracted from Murphy's (2002) somewhat related discussion of noun-pairs like BOATHOUSE and HOUSEBOAT (p. 445). Put simply, if BOAT and HOUSE contribute the same content to both compounds, then the contents of BOATHOUSE and HOUSEBOAT would be identical, which is clearly false. Therefore, either HOUSE or BOAT (or both) contributes different contents to each compound.

The second caveat is that is that I do not wish to take a stand here on what constitutes a primitive concept (i.e. a concept that cannot be decomposed into any constituent sub-concepts). It might, for instance, be that lexical concepts are primitive (e.g. Fodor, 1998), or it might be that perceptual representations play the role of primitives and can be shared amongst numerous concepts (e.g. Barsalou, 1999). The reason for remaining neutral on this issue is that doing otherwise would in some sense presuppose an account of concept structure. This, in turn, would introduce an element of circularity into any evaluation of a particular theory of structure.

1.2.6 Neural Plausibility

Since I am arguing that concepts are a particular type of content-bearing neural vehicle, and since any account that assumes materialism has to relate concepts to brain activity, I take neural plausibility to be an important consideration. There are a number of different dimensions on which this criterion can be explicated. First, there are what one might call computational considerations. To give one example, if a theory requires an occurent perceptual representation to be compared to thousands or millions of stored concepts in order to be categorized, then it's likely that the theory is computationally infeasible given how quickly categorization occurs and how quickly neurons transmit information to one another. To give another example, if combing two concepts requires combining every single representation that could possibly be associated with either one, then the theory is again infeasible, because it is implausible to suppose that the brain carries out such an enormous number of unique binding operations for every simple case of conceptual combination.

Second, there are what might be called resource considerations. If, for example, one's theory entails that each concept requires millions of dedicated neurons, then there simply would not be enough neurons in the brain to support a normal range of human conceptual capacities. Likewise, if one's theory requires that the neural vehicle underlying a particular concept have direct links to a high number of other vehicles, then the theory may or may not be plausible depending on what is known about the interconnectivity of actual neurons.

Lastly, I take the presence of a highly biologically plausible example of implemented conceptual processing to count strongly in theory's favor, all else being equal. This is potentially relevant in the present context because the theory of content I employ very clearly ascribes meaning to groups of individually spiking neurons. To the best of my knowledge, aside from Eliasmith (forthcoming), there is no existing work that gives an account of conceptual structure specified in terms of the functional properties of individual neurons. Thus, any theory that can plausibly describe how conceptual processing takes place in the brain with such a high degree of biological detail is at a serious advantage over competitors that rely, for instance, on neural network models (e.g. Rogers & McLelland, 2004) or on models derived from fMRI data (e.g. Barsalou et al., 2003, 2008).

1.3 Conclusion

To conclude, I have articulated what I take to be the guiding criteria for a good theory of concepts. These criteria are based on satisfying explanatory targets derived from data on such things as the scope, use, and content of those things concepts are often pre-theoretically taken to be. In what follows, I primarily use an independently motivated account of representational content to formulate a theory of concept structure that I think satisfies these criteria. First, though, it is necessary to examine existing theories in greater detail.

Chapter Two

Evaluating Existing Theories of Concepts

2.1 Introduction

The widespread interest in the study of concepts that I mentioned at the start of the previous chapter is not the consequence of an absence of well-described theories. On the contrary, there are many existing accounts of concept structure, each of which can list a collection of explanatory successes in its favor. Though no single theory has achieved widespread acceptance, some scholars firmly advocate for one specific variant while others have suggested 'mixed' or 'dual' theories as a strategy to overcome some of the explanatory shortcomings that a favored account possesses when considered in isolation (see e.g. Murphy, 2002; Laurence & Margolis, 1999; Margolis & Laurence, 2005 for discussion). This latter possibility of theoretical pluralism, in particular, raises the question of whether or not there is even a need for a novel approach to the problem of concept structure.

In what follows, I argue that the main existing theories of concepts cannot satisfy the explanatory goals set out in the previous chapter, regardless of whether these theories are considered individually or in a mixed form. More specifically, I consider classical views that identify concepts with definitions, similarity-based views that identify concepts with prototypes or exemplars, theory-based views that identify concepts with their inferential role, and atomistic views that identify concepts with unstructured mental symbols that enter into nomic relations with their referents. Given the diversity with of the views just mentioned (and for the sake of brevity), I restrict my discussion to representative accounts and do not comprehensively survey the space of possible theories. Finally, while many of the criticisms

examined below have already been widely discussed in the existing literature, I offer novel evaluations in some cases.

2.2 Classical Views

Until roughly the middle of the twentieth century, the default view amongst philosophers and psychologists was that concepts are structured as definitions. Definitions, in turn, are sets of features that specify individually necessary and jointly sufficient conditions for the proper application of a concept (Laurence & Margolis, 1999). The concept BROTHER, for example, can be defined as MALE SIBLING, and therefore applies only to those things that are both males and siblings. It is important to note that such definitions are constituted by *other* concepts, which in turn have definitions of their own; SIBLING, for instance, might be defined as PERSON WITH SHARED PARENTS. Eventually, though, most theorists take this definitional regress to reveal a stock of primitive concepts that cannot be structurally decomposed any further (Fodor et al., 1980; Laurence & Margolis, 1999; Prinz, 2002). These primitive representations can alternately be characterized as perceptual (which yields an empiricist version of the classical view), or as innate (which yields a nativist version). So, while there are variations on the specific way in which a definitional view[6] of concept structure gets articulated, the common theme is this: concepts are composed of mental representations that jointly determine reference through the specification of necessary and sufficient conditions for category membership.

[6] "Definitional views" and "classical views" are used interchangeably throughout.

Given as much, the well-known explanatory advantages of classical views are apparent. First of all, the publicity criterion is easily accounted for by the fact that different people can learn the same definitions (Fodor, 1998; Prinz, 2002; Smith & Medin, 1981). Second, categorization can be explained as the matching of an entity with a particular concept's definition (Laurence & Margolis, 1999). For instance, if something is obviously male and obviously a sibling, then one can categorize it as a brother provided that one knows the relevant definition. The content criterion can be handled in a similar manner: the defining features of a concept specify what it refers to (Prinz, 2002; Smith & Medin, 1981), while distinct but co-extensive definitions may ascribe different properties to the same things. Finally, the compositionality criterion also seems to be satisfiable considering that the definitions of complex concepts are constituted by sub-concepts, each of which may or may not be constituted by further sub-concepts (Fodor, 1998). Accordingly, the content of a complex concept is a function of the contents of its defining sub-parts. The classical approach to structure, in all, is favored by a tidy explanation of categorization, compositionality, publicity and, most notably, reference.

Despite the neatness of this story, definitional views of concepts are now widely disregarded (Murphy, 2002; Laurence & Margolis, 1999). One common criticism stems from a well-documented lack of success in coming up with workable definitions for many simple concepts. Wittgenstein (1953) famously illustrated that GAME has nothing close to a set of defining features. Rather, the concept applies to a range of activities sharing what he calls a "family resemblance" or a "complicated network of similarities" (p. 171). Ping-pong and solitaire, for instance, are different enough to avoid being encapsulated by a single definition,

yet they also share certain attributes, such as the requirement of skill and the production of amusement. Given that proposed definitions for both abstract concepts like KNOWLEDGE (e.g., Gettier, 1963) and mundane concepts like FURNITURE (e.g., Laurence & Margolis, 1999) have been continually refuted with counter examples, the point to draw here is that classical views have had a tough time meeting the scope requirement (Prinz, 2002).[7] It might be that we simply have yet to find the right definitions, but induction from numerous examples of failed conceptual analysis favors the conclusion that there simply are no successful definitions to be found.

The main objection to classical views, however, stems from their failure to account for large swaths of empirical data regarding categorization. The typicality effects discovered by Rosch and her colleagues (1976, 1978) simply are not predicted when concepts are structured as definitions. Nor is the tendency for subjects to categorize objects at what Rosch (1978) calls the "basic level of abstraction" (p. 192), or an intermediate level in a taxonomical hierarchy. In the case of a kitchen chair, for instance, the category *chair* is more salient than *furniture* or *kitchen chair*, even though all three apply (Rosch, 1978). This effect is also manifested when subjects are engaged in production tasks. When asked to describe a kitchen chair, subjects list features of chairs in general, rather than features unique to the specific chair type (Rosch, 1978). The reason, more generally, why classical views fail to satisfy these varied empirical requirements is because a definitional structure treats all

[7] One response to this criticism is to suggest that concepts have disjunctive definitions. Wittgenstein (1953) rejects this proposal on the grounds that such specifications are unnecessary for the regulation of concept *use*, while Smith and Medin (1981) note that the characterization of definitions as sets of necessary and sufficient features excludes the possibility of disjunctive concepts, because no individual disjunct is ever *necessary* for reference determination.

category members equivalently (Laurence & Margolis, 1999; Murphy, 2002). For example, if both a robin and ostrich satisfy the definition of BIRD, then one would have no reason on a classical view to expect them to be treated differently in categorization tasks. Yet, subjects perform more quickly when answering questions like "Is a robin a bird" than when answering questions like "Is an ostrich a bird" and robins are rated as good examples of a bird, while ostriches are rated as poor examples (Smith, Shoben, & Rips, 1973). These results, along with others of a similar nature, indicate that concepts must encode more than just a set of defining conditions for correct application. Accordingly, definitional views just do not seem to be up to the task of meeting the categorization criterion.

There are two rebuttals to this conclusion that a classical theorist might invoke. First, if one assumes categorization involves a process of feature matching computations, then it is possible for some typicality effects to be explained through reference to definitional complexity (Smith & Medin, 1981). On this view, complex definitions yield longer processing times because more features have to be matched together in order to achieve a categorization judgment. The unfortunate consequence of this approach is that typical category members have to have simpler definitions than atypical ones. No evidence suggests that this is the case, as Laurence and Margolis (1999) note.

The second response involves postulating a "dual" model wherein concepts have a prototype structure for quick identification procedures, and a definitional core for more contemplative acts of categorization (Laurence & Margolis, 1999; Murphy, 2002; Smith & Medin, 1981). The problem here, however, is that one can think of this move as yielding the

worst of both worlds. In addition to the difficulties associated with definitions in general,[8] one inherits all of the problems associated with prototypes at the same time, such as those involving compositionality (see below). Accordingly, it is not clear that much is purchased with the proposal of a dual theory.

With respect to meeting the content criterion, definitions show promise. For example, the ability for multiple co-extensive definitions to ascribe distinct properties to the same things seems obvious. Regarding reference, though, problems crop up. As Fodor et al. point out (1980), saying that some entity satisfies the defining features of a particular concept deflects rather than explains the required notion of reference. If something satisfies BROTHER by satisfying MALE and SIBLING, then the classical theorist still owes an explanation of what it means to satisfy these sub-concepts (or of what it means to satisfy a primitive, if one follows the definitional regress back that far; Prinz, 2002). In the absence of such explanations, classical theorists have trouble meeting the content criterion.

Compositionality is another area in which definitional views are often thought to have a distinct advantage over their competitors. There are reasons to doubt this supposition too, though. First, it helps to observe a distinction offered by Laurence and Margolis (1999) regarding two general types of model for conceptual structure. On a so-called "containment model" (Laurence & Margolis, 1999, p. 5), a complex concept *C* might have concepts *X, Y,* and *Z* as component parts such that every tokening of *C* necessarily involves a tokening of *X,*

[8] Further problems not mentioned thus far include the problem of people not requiring the right definitions in order to possess a concept (consider, for instance, a health-related concept like DEPRESSION, which was once constituted by erroneous beliefs involving supernatural spirits), and the problem of ambiguous category membership, which cannot be accounted for with definitions (Laurence & Margolis, 1999).

Y, and *Z.* On an "inferential model" (Laurence & Margolis, 1999, p. 5), the component parts *X, Y,* and Z merely stand in some privileged causal relation to *C* such that the tokening of *C* often produces but does not require the tokening of *X, Y,* and *Z.*

If definitional views are understood to operate in accordance with a containment model, then compositionality problems immediately crop up. Put simply, if *C contains X, Y,* and *Z,* then *C* also contains whatever components make up *X, Y,* and *Z,* along with the sub-components that make up the components, and so on. Assuming even a modest definitional regress, it is plausible that on this view many ordinary lexical concepts would have to 'contain' hundreds or thousands of sub-concepts. Yet, when one combines, say, LOUD and GUITAR to form LOUD GUITAR, it is clear that one is *not* combining both the definitions of LOUD and GUITAR along with all of the further definitions of these definitions. Moreover, there are numerous cases in which compound concepts have meanings with no clear relation to the definitions of their constituent parts; consider DAILY GRIND or SUNNY DISPOSITION. A consideration of compositionality accordingly reveals the implausibility of the definitional view when it is construed as a containment model.

One might respond here by favoring an inferential model. However, Quine's (1953) well-known critique of analyticity reveals some problems with this move.[9] On an inferential model, a concept possesses certain *defining* inferential relations, and certain *non-defining* inferential relations that vary from person to person (Fodor, 1998). The most apparent way to

[9] Quine's attack on the analytic/synthetic distinction is often thought to present more general problems for classical views of concepts (for a summary, see Laurence & Margolis, 1999), but at present, it just worth noting that a critique of analyticity also undercuts the containment model just discussed, since the notion of containment being employed likely presupposes an analytic relation between a concept and its definition.

understand this distinction is to characterize the defining relations as analytic. But, as Quine argues, it has yet to be demonstrated that there is a non-circular way of explicating the notion of analyticity – arguments that employ ideas like definition, synonymy, meaning, truth preservation, or semantic rules either presuppose some notion of analyticity, or fail to achieve a robust explanation. For example, to say that 'bachelors are unmarried' is true by definition does nothing explicate an analytic relation between 'bachelor' and 'unmarried' because the notion of definition being employed either presupposes analyticity, or merely describes the common linguistic "usage" of the two terms (Quine, 1953, p. 156).

The reason the lack of a credible analytic/synthetic distinction poses a problem for the compositionality criterion is because if there is no way to determine which inferences are defining of a concept, then there is no way to determine how such inferences compose when a complex concept is formed. For example, if my concept TALL has, for whatever reason, inferential connections to ARROGANT, then my complex concept TALL TREE would also have inferential connections to ARROGANT via compositionality. Excluding this bizarre conclusion requires characterizing the inference from TALL to ARROGANT as non-defining, but this likely cannot be done without an account of analyticity. Moreover, reference becomes a problem if ARROGANT is a part of the compositional structure of TALL TREE, because the set of tall trees would be empty if arrogance were a necessary condition of membership. Overall, inferential models of definitions simply do not compose well in the absence of a clear analytic/synthetic distinction.

Lastly, on the matter of neural plausibility, it is worth mentioning some empirical evidence indicating that definitions are unlikely to be psychologically real. In experiments

involving phoneme-monitoring tasks, where subjects are asked to both comprehend a verbal statement and listen for the first instance of particular phoneme in it (see e.g. Fodor et al, 1980; Laurence & Margolis, 1999 for a summary), there is no observable effect of definitional complexity. When listening to sentence pairs like "John's bachelor friend has arrived" and "John's unmarried friend has arrived," subjects react to the underlined phoneme with identical speed, even though it should take longer to understand 'bachelor' than 'unmarried' if 'unmarried' is part of the definition of 'bachelor' (Fodor et al., 1980; Laurence & Margolis, 1999). So, while it is plausible that definitional structures *could* be represented in the brain,[10] there is no evidence that they actually are.

To summarize, then, classical views seem to fall short on almost all criteria necessary for a good theory of concepts. The scope criterion remains unsatisfied due to a widespread lack of definitions. The categorization criterion remains unsatisfied due to the absence of a plausible explanation of typicality effects. The content criterion remains unsatisfied due to the lack of a clear explanation of the reference relation between definitional representations and their intentional objects. Meeting the compositionality criterion is problematic regardless of whether one adopts a containment model or an inferential model. The neural plausibility criterion, finally, also remains unsatisfied given the available psychological evidence. With only the account of publicity safe from genuine criticism, it is no surprise that very few contemporary researchers advocate a classical view.

[10] See Eliasmith (forthcoming) for an example of a biologically plausible cognitive architecture that suggests a way in which some definitional concepts could be represented in the brain.

2.3 Prototypes and Exemplars

Amongst psychologists, the decisive blow against definitional accounts came primarily from Rosch's work on categorization in the 1970's (Murphy, 2002). In consequence, the notion that concepts have a 'prototype' structure, or a structure that lists typical but not definitive features of category members, became widespread. Prototypes can be alternately characterized as lists of variably weighted features, points in a multidimensional metric space, or as singular template representations (Smith & Medin, 1981). The template approach has not been particularly successful due to scope restrictions, while the dimensional and feature-based approaches can and often are combined into a single "component" approach (Smith & Medin, 1981, p. 164).[11]

To given an example of a concept structured as a prototype, consider CAT. CAT might be composed of both weighted features and dimensional attributes as follows: PURRS (1.0), CHASES MICE (1.0), HAS FUR (0.9), HAS TAIL (0.9), HAS CLAWS (0.9), HAS FOUR LEGS (0.8), DISLIKES WATER (0.7), SPEED: RANGE *X* (0.6), SIZE: RANGE *Y* (0.5), and so on. To determine whether a given thing is a cat, one simply runs a similarity measurement process that matches input representations with the feature representations. For every instance of a match, the corresponding feature weight is added to an accumulator, and once a critical numerical threshold is reached, a positive categorization judgment is made (e.g. any value over 7 might be sufficient for the application of CAT).[12] On a purely dimensional account, a positive

[11] Note too that feature-based and dimensional accounts are mutually translatable, given that a dimension can be represented as a graded set of features, and that a feature can be represented as a dimension with a binary value range (Smith & Medin, 1981).

[12] A more commonly used, and complex, formula for this similarity computation is as follows: $\mathrm{Sim}(P, I) = af(P \cap I) - bf(P - I) - cf(I - P)$ (Smith et al, 1988). The formula

categorization judgment is made when an input occupies a location in a metric space that is within a minimum threshold distance of the location of the prototype (Smith & Medin, 1981).

The explanatory successes of prototype theories are numerous. First of all, prototypes can be generated for a wide range of concepts, from natural kinds like BIRD to abstractions like ART (Prinz, 2002). This suggests an accommodation of the scope criterion. Second, because prototypes are thought to capture statistical regularities in the environment (Murphy, 2002; Rosch, 1978; Rosch et al, 1976; Smith & Medin, 1981), they are likely to be similar across individuals. Fur, for example, is a modal feature of dogs, so all DOG prototypes are likely to include FUR on their feature lists. Third, a wide range of categorization data is accounted for by prototypes. Fast categorization, for instance, is explained by the fact that fewer feature-matching computations are needed to reach the threshold value when a category instance is highly typical (Murphy, 2002; Smith & Medin, 1981). Fourth, content is explicable on the following grounds: similarity thresholds determine reference, while diverse property ascriptions can be accounted for by the possibility of distinct prototypes corresponding to the same category (Prinz, 2002). Finally, Eliasmith (forthcoming) proposes a general cognitive architecture that describes prototypes as high-dimensional vectors implemented in populations of neurons. This suggests that the neural plausibility criterion can also be satisfied. Overall, the prototype structure shows promise, particularly in satisfying the categorization criterion that was so devastating to the classical view.

---

incorporates measures of the features unique to each of the prototype and the input instance, while employing further weighting parameters (*a, b, c*) and a function (*f*) to evaluate the relative significance of the feature sets.

There are, however, three substantial problems with the prototype thesis. First of all, many concepts simply do not have them (Fodor, 1998; Fodor & Lepore, 1996; Laurence & Margolis, 1999; Osherson & Smith, 1981; Prinz, 2002). Boolean concepts[13] like NOT A CAT, for instance, lack prototypes because it is false that the less something is like a cat, the more likely it is to exhibit any particular set of features (Fodor, 1998; Fodor & Lepore, 1996). Mountains and popsicles, for instance, are highly unlike cats, yet they have essentially nothing in common. Moreover, both general and non-referring concepts also lack prototypes (Laurence & Margolis, 1999). There is, for instance, no prototypical OBJECT WITH FOUR COLORS. Likewise, a brief consideration of the concept INHABITANTS OF SATURN fails to identify a set of typical features for its (non-existent) category members. So, while prototypes may be better than definitions at meeting the scope criterion, they nonetheless fail to be good *enough*.

The second major problem relates to the content criterion. As many scholars have pointed out, prototypes do not adequately specify reference (e.g. Armstrong et al, 1983; Prinz, 2002). Something can be a diamond, for instance, without satisfying the feature-matching threshold necessary for the application of DIAMOND; consider a diamond drill bit. Likewise, something can possess many of the typical features of diamonds (and therefore meet the categorization threshold) without actually being one; consider jewelry made of cubic zirconia. The point is that prototypes often mischaracterize the extensions they purportedly represent. Note too that this problem can alternately be described as a failure to meet the categorization criterion: it is so-called 'knowledge effects' that manifest themselves

[13] Boolean concepts are concepts involving logical operators like IF, THEN, AND, and OR (Fodor & Lepore, 1996).

when people use their background beliefs about diamonds to dismiss a necklace made of cubic zirconia as 'fake.' These effects are an important set of data that prototype theories fail to explain (see Murphy, 2002). So, because prototypes fail to meet the reference criterion, they also encounter problems accurately describing how concepts are applied.

The most widely discussed problem for prototypes, however, has to do with compositionality. As Fodor and Lepore (1996) note, the prototype for PET FISH is not a simple combination of the prototypes for PET and FISH. A typical pet fish is something like a goldfish, but a goldfish is neither a typical pet, nor a typical fish (Fodor & Lepore, 1996). The more general point is that complex concepts tend to exhibit emergent features that are not at all typical of their constituent parts. There is accordingly no simple "composition function" that maps constituent prototype features onto a prototype for a complex concept (Fodor & Lepore, 1996, p. 254). One possible conclusion to draw here is that the prototypes for complex concepts are idioms, or constructions in which compositional semantics get overridden by conventions or learning experiences (Fodor & Lepore, 1996). But, idioms do not compose, so if prototypes are idioms, then they cannot be concepts.

Given the significance of this problem, a number of attempts have been made to create a workable model of prototype combination. One early approach was to construe prototypes as specifying graded extensions (i.e. typical members of class are *more* in the extension than atypical members). This allows for the characterization of complex concepts as representations of the intersection of a series of fuzzy sets (see Osherson & Smith, 1981). A BLUE BOOK prototype, for example, might take the minimum of the degrees to which something is blue and a book, and then produce a positive categorization judgment in the

event that this minimum is above a certain threshold. Put more specifically, if something is blue to degree 0.6 and a book to degree 0.2, then it will be judged to be a blue book to degree 0.2 (Laurence & Margolis, 1999).

However, as Osherson and Smith (1981) demonstrate, the fuzzy set approach yields erroneous predictions. If a pet guppy is deemed a fish to a degree of 0.7 and a pet to a degree of 0.6, then the natural conclusion to draw is that a guppie is a pet fish to a degree of 0.6. Guppies, though, are a prototypical pet fish, and should accordingly be included in the extension of PET FISH to a very high degree. Perhaps to a degree of 0.8 or 0.9. Moreover, depending on the categorization threshold for PET FISH, the fuzzy set approach might even result in one failing to identify the guppy correctly. It is accordingly clear that models appealing to the intersection of graded extensions cannot offer a robust explanation of prototype combination.

A more recent solution to the problem involves what is called a 'selective modification model' (Smith et al, 1988; see Murphy, 2002). Explicated only in terms of noun-adjective combinations, the model proposes that a complex concept is formed when features of an adjective prototype modify and reweight features of a noun prototype. Or, more specifically:

> Each attribute in the adjective concept selects the corresponding attribute in the noun concept; then for each selected attribute in the noun, there is an increase in the salience (or votes) of the value given in the adjective, as well as an increase of the diagnosticity of the attribute. (Smith et al, 1988, p. 360)

To cash this out with an example, consider the combination of SHRIVELED and APPLE. The suggestion is that APPLE contains an attribute for texture, and that SHRIVELLED boosts the diagnosticity of this attribute while shifting votes onto the 'bumpy' value from the 'smooth' value (Smith et al, 1988).

The problems with the selective modification model are threefold. First, there is lack of scope (Laurence & Margolis, 1999). Many complex concepts incorporate much more than simple noun-adjective pairings, and it is unclear how the model could generalize to cover such cases. Second, as Murphy (2002) points out, it is not the case that adjectives select dimensions[14] to modify in a clear or consistent manner. Consider the concepts CORPORATE LAWYER and CORPORATE DONATION. For the latter, the adjective likely modifies a dimension like 'received from __,' whereas for the former, the adjective likely modifies a dimension like 'works for __' (Murphy, 2002). A systematic mapping function between the features of adjectives and nouns is accordingly unlikely to be found. Third, even if there were some way of determining such a function, its existence would require nouns to actually possess *all* of the features that they could possibly have modified. This in turn would require every noun prototype to contain every feature that could be modified by any adjective. Representational economy accordingly goes out the window, perhaps along with neural plausibility.

Altogether, prototypes exhibit only marginal improvements on the shortcomings of the classical view. They lack the scope necessary to support the known range of human concept types, and they lack an adequate account of conceptual content. Most decisively,

[14] What Smith et al (1988) call 'attributes,' Murphy (2002) calls 'dimensions'.

they fail to accommodate the compositionality criterion in any persuasive manner. Prototypes, in short, cannot be successful candidates for a theory of concepts.

I'll only briefly discuss exemplar views. Because they share many of the same features as prototype theories, they also share most of the same problems.[15] The main difference is that instead of explaining categorization through matching inputs to summary representations, exemplar views suggest that categorization occurs through the matching of inputs to representations of previously perceived category instances (Smith & Medin, 1981). An advantage obtained with this approach is that one can explain things like the tendency for people to exhibit proficiency when categorizing familiar but atypical instances of a given category, along with a variety of other so-called 'exemplar effects' (see Murphy, 2002).[16] A further advantage is that exemplar views may also improve on prototype views with respect to scope. One's NOT A CAT concept, for instance, could be comprised of example representations of mountains and popsicles. Likewise, a set of example representations of things like four colored flags could constitute the structure of OBJECTS WITH FOUR COLORS.

These apparent improvements merely distract from familiar problems involving content, categorization, and compositionality. First, exemplars cannot adequately specify reference for the simple reason that any given set of example representations cannot encompass the entire extension of a concept. To illustrate, the average person's experiences with instances of birds is unlikely to yield an exemplar that matches with penguins and

[15] In fact, on some accounts, exemplars *just are* prototypes at a subordinate level in a taxonomical hierarchy. For example, a prototype representation of a robin can be understood as the same thing as an exemplar representation of bird (e.g. Rosch, 1978).

[16] Typicality effects are also accounted for by the fact that one is more likely to store exemplars of highly typical category members (see Smith & Medin, 1981 and Prinz, 2002)

ostriches. And even if a given individual's experiences of birds do include penguins and ostriches, then there is still the problem of cases of the sort described by Keil (1989), where, say, a fish is disguised to look like a swimming penguin, or a swimming penguin is disguised to look like a fish. The more general, content-related difficulty shared by prototypes and exemplars is that both can refer *only to what they are taken to refer to* (Prinz, 2002). Misrepresentation is impossible: anything that meets the similarity threshold for a given concept is, by necessity, a member of the concept's extension. I take this to be a decisive point against both types of theory.

Compositionality, if anything, is worse for exemplars than prototypes. It seems, for instance, impossible to generate a representation of a pet fish from exemplars of pets and exemplars of fish. Which pet exemplar gets paired with which fish exemplar? And, more specifically, why think that any features associated with a given pet exemplar (e.g. a dog, a cat, or a parrot) ought to be relevant to the generation of a representation of a pet fish?[17] The problem is that emergent features in PET FISH (e.g. 'lives in a tank') are inexplicable when a standard composition function is applied to exemplar representations of pets and fish. Moreover, even if one was willing to concede that exemplars or prototypes somehow contain such emergent content as part of their structure, neural plausibility likely goes out the window. Attaching all of the possible content that could emerge as a result of a given concept's inclusion in a complex construction to the structure of the concept would, quite simply, result in inordinately large concepts. Consider the variety of emergent features in

[17] Note that appealing to a pet fish as a pet exemplar is circular. Alternatively, claiming that pet fish prototypes or exemplars are generated through learning experiences with pet fish reduces PET FISH to an idiom, and idioms are not compositional (Fodor & Lepore, 1996)

such concepts as PET BIRD, PET ROCK, PET PROJECT, PET PEEVE, TEACHER'S PET, PET SITTER, PET STORE, PET GROOMING, PETTING ZOO, and so forth. If PET had to contain *all* of these potentially emergent features, PET would be massive.[18] Assuming an inference model rather than a containment model might mitigate the problem, but in this case, concerns regarding the selection of the right inferences by a composition function come to the fore again. Exemplars, overall, do little to succeed where prototypes fail.

2.4 Theory Theory

In tandem with the ongoing development of prototype and exemplar views, the idea that concepts are structured like miniature theories of the categories they denote has become an increasingly popular target of psychological research (see Keil, 1989; Murphy & Medin, 1985; Prinz, 2002). The basic insight prompting this development is that individuals possess beliefs about things like causal relations, essences, and ontological distinctions that seem to influence how they form and use concepts (Keil, 1989; Murphy & Medin, 1985; Prinz, 2002; Rogers & McLelland, 2004). For example, the reason why BIRD denotes a coherent category and groups entities in the way that it does is because many of the features shared by most birds (such as flight, wings, feathers, and hollow bones) are related to one another via a set of one or more explanations: birds can fly because they have wings, feathers, along with hollow bones; and birds fly because doing so helps them gather food and avoid predators (Murphy, 2002; Rogers & McLelland, 2004). So, the benefit of describing BIRD as a miniature theory of

[18] It is possible that this degree of representational complexity could be incorporated into a prototype or an exemplar model if the various features of a given prototype or exemplar are accessed through transformations of a single representation (see Chapter 3).

birds is that one gains an explanation of *why* certain features of birds are deemed relevant or salient in the formation of the category,[19] and of *how* the attributes characteristic of category members are related to one another (Murphy & Medin, 1985). Neither of these virtues are present in the explanations provided by prototype, exemplar, and classical theorists.

Some clarity is required regarding the exact nature of the account of concept structure being proposed by theory theorists. Given that concepts are alternately referred to *as theories* and as being *embedded in theories*, the nature of the relationship being proposed between the two is not always clear (Fodor, 1998; Laurence & Margolis, 1999; Prinz, 2002). The general consensus amongst philosophers, however, is to characterize the thesis as one suggesting that concepts are representations individuated by the role they play in a theory constituted by a set of concepts and the inferential relations that hold amongst them (Fodor, 1998; Laurence & Margolis, 1999; Prinz, 2002). Thus, to take an arbitrary example, JAR would be a singular representation whose content is defined by its place in inferential schema[20] involving other concepts like CONTAINER, SEALED LID, USED FOR STORAGE, PRESERVES PERISHABLES, and so on. The essential difference between this account and, say, a prototype view that lists some of the same concepts on a feature list is that the nature of the *relations* between the features is accounted for by the inferences that make up the explanatory schema – a jar is used for storage, for instance, because it is a container, and it has a sealed lid so as to be of use in

[19] The standard explanation of category formation is that categories latch onto kinds that exhibit clusters of similar properties while lacking properties that are highly typical of other such clusters. However, this is rather uninformative given that "any two entities can be arbitrarily similar or dissimilar by changing the criterion of what counts as a relevant attribute" (Murphy & Medin, 1985, p. 428).

[20] It is not generally supposed that a mental 'theory' is of comparable scope to a proper scientific theory. Rather, the term 'theory' refers to any minimally sophisticated explanatory framework or schema. Hence, 'explanatory schema' and 'theory' are used interchangeably.

preserving perishables. Accordingly, the inferences warranted by one's JAR concept serve to individuate JAR while also collectively forming a general explanatory account of what jars are.

There are a number of benefits afforded by indentifying the structure of a concept with inferential role in this manner. First of all, scope is handled fairly easily given that just about any concept one can think of can be matched with some sort of mental schema. For example, the fact that we can understand concepts like SUM and CENTAUR indicates that we must have some sort of explanatory account of what it takes for something *to be* a sum or a centaur. Second, theory-based views arguably provide the best account of categorization yet discussed. Knowledge effects are handled via the numerous inferences that a concept embedded in a theory licenses. For example, Murphy (2002) discusses an experiment conducted by Lance Rips in which subjects were asked to categorize transformed animals (e.g., a small bird exposed to toxic chemicals that cause it to lose its feathers and develop thin, transparent wings that make it look highly similar to an insect). The theory theory explains the tendency for subjects to still categorize the creature as bird because it allows BIRD to have inferential connections to some sort of notion of 'bird essence' that stipulates that things with bird innards born of bird parents always remain birds. The relation between parentage and essence, in other words, is an explanatory relation captured by a theory rather than a feature list or definition. Third, neural plausibility might also be handled on some models of conceptual processing. Rogers & McLelland (2004), for instance, describe a connectionist model called the "parallel distributed processing framework" that can arguably implement the type of conceptual system described by theory theories (p. 49). Likewise,

Eliasmith's (forthcoming) account of a biologically plausible cognitive architecture proposes a manner in which concepts that include inferential connections can be implemented in populations of spiking neurons. Overall, then, the theory theory is initially favored by a promising account of scope, categorization, and neural plausibility.

A common complaint against theory theories is that they cannot meet the publicity criterion. Fodor (1998), for instance, argues that when theory theorists claim that concepts are constituted by their place in the theories in which they are embedded, they are simply reiterating the central tenet of inferential role semantics. And inferential role semantics, Fodor contends, presents an unfavorable dilemma to the concept theorist. Either *all* of the inferential relations a given concept possesses are relevant to the determination of its content, or only *some* are. If all the relations are relevant (i.e. holism is true), then no one shares the same concepts because everyone's total set of mental representations is unique. Hence, concept publicity is impossible if one chooses this horn of the dilemma. And if only some of the inferences are content determining (i.e. molecularism is true), then the theory theorist needs say *which* ones. Doing so has been proven to be problematic given the aforementioned challenges faced by accounts that specify *defining* inferences (i.e. classical views) or *characteristic* inferences (i.e. prototype and exemplar views). So, opting for the other horn of the dilemma only benefits the theory theorist if he or she has some sort of account of which set of inferences or 'theory' one needs to have in order to possess a given concept. Fodor's (1998) complaint is that no such account has been given.

I think this evaluation is problematic. First of all, if one opts for holism, strict concept identity is likely just too high of a burden to impose. Stating that two people have distinct

CAT concepts simply because each representation possesses different inferential relations by virtue of being in a unique cognitive system ignores the fact that the concepts might nonetheless have highly *similar* inferential roles. Suppose one person infers USUALLY MEAN AND ALOOF from CAT, while a different person infers USUALLY FRIENDLY AND PLAYFUL. So long as many of the other inferences deriving from each person's cat concept are identical, it is very reasonable to suppose that the two individuals can effectively communicate when talking to one another about cats. To the extent that they do not, I would argue that a certain degree of misunderstanding is commonplace in human interactions, and that a theory of concepts (if tied to language use) ought to anticipate and accommodate this fact. Finally, to address the obvious concern about what constitutes sufficient similarity for mutual concept possession, I think that there is a simple pragmatic benchmark that can be invoked: one possesses a concept (such as CAT) insofar as one is able make oneself understood when using this concept in communicating with other individuals in the same discursive community. Communicative efficacy relativized to a specific context, in other words, is likely sufficient for concept possession.[21] More argumentation will offered in support of this claim in chapter three.

If one opts for molecularism instead of holism, the problem of specifying which inferences are the constitutive ones can be solved in a nearly identical way. Inferences that

[21] This probably entails some mild form of conceptual relativism between communities and temporal periods. But given that children can understand adults, adults can understand children, non-native speakers of language can understand native speakers, modern readers can understand medieval authors, and so on, I think that cases of incommensurable concepts are far less common than it might seem. Some natural objections to this account will be considered later.

are highly important for achieving communicative efficacy are constitutive of content, while ones that are not communicatively important are not constitutive of content. For example, if possessing the inference from GLASS to IS TRANSPARENT or USED FOR MAKING WINDOWS is important for one to understand what other people mean when they talk about glass,[22] then these inferences are likely content determining. In contrast, the inference from GLASS to WHAT UNCLE BARRY MAKES HIS CRAFTS OUT OF in unlikely to be important for understanding discourse involving glass-talk, and hence unlikely to be content determining. Overall, these (albeit brief) arguments suggest a natural solution to the problem of publicity for theory theories provided that one abandons the requirement of strict concept identity.

This solution, however, highlights a further complaint regarding the content criterion. Laurence and Margolis (1999), for example, suggest that reference is underspecified by theory theories because if we concede that the same concept can be embedded in slightly different explanatory schemas across individuals, then each individual's version of the concept will necessarily pick out a slightly different extension. It is also possible for an individual to possess a mistaken theory, in which case their concept picks out the *wrong* extension (Laurence & Margolis, 1999). Moreover, given that the theory associated with a given concept essentially constitutes the set of properties it ascribes to its intentional objects, difficulties also emerge for the aspect of content that deals with property ascription. Saving publicity, in short, might come at the cost of losing a clean account of content.

[22] The relevant notion of "understanding" being employed here could operationalized in terms the generation of predominantly successful behavioral predictions. For example, if you tell me you are going to buy glass at the hardware store, I understand your use of the concept GLASS to the extent that I am able to predict what you end up doing (in the absence of extenuating circumstances).

Again, though, a plausible solution to these problems can be sketched out. If one entertains the proposal that concept possession is a matter of communicative efficacy relativized to a context, then an important entailment follows. Namely, concept possession becomes a matter of degree, ranging from some sort of minimal competency (e.g., that of a child first learning the concept BACTERIA) to a height of maximal expertise (e.g., that of a renowned bacteriologist). The extension of BACTERIA, on this view, is the class of entities picked out by our best scientific account of what it takes for something to be a bacterium. It is possible, of course, that something could be a bacterium even if *no one* thought so, but any attempt to accommodate this fact into an account of reference would essentially make the proper extension of BACTERIA epistemically inaccessible. So, provided that one takes the best available account of a concept to specify the best approximation of its extension, then the matter of reference is largely cleared up. Non-experts and those with minimal competency may make mistakes and misrepresent regularly, but insofar as they are communicatively efficacious (and responsive to reasoning) then they can have their concepts revised and updated through discourse with the more competent. Again, these comments are merely suggestive, but I think they show promise, especially given that they are consistent with the notion that concepts are regularly modified and revised through debate and scholarly investigation.

In fact, I only identify three shortcomings with the theory theory. The first is that it is underspecified. Proponents of the view typically fail to offer a concrete model or set of mechanisms showing how it could implement some sort of basic conceptual processing (Rogers & McLelland, 2004); simply stating that a concept's structure is determined by its

inferential role is not particularly illuminating or explanatory. The second problem is that there is no natural explanation of typicality and exemplar effects. And the third problem is that in the absence of a more clear and precise account of how inferential role individuates a concept, the matter of how concepts compose is left unaddressed. Presumably, theories do not just sum together to produce a characterization of the inferential role of a complex concept – emergent attributes of the sort discussed in the section above on prototype combination have to be accounted for in some way. Explaining how the inferences that characterize sub-concepts are modified and combined when conjoined to form a complex concept is accordingly a goal that a fully fleshed account ought to strive for.

Altogether, while the theory theory in its standard guises fails to meet all six of the explanatory criteria, it does offer a number of important insights. First, any account of concept structure that can meet the categorization criterion is likely going to have to incorporate explanatory beliefs to some extent; otherwise, knowledge effects will remain a problem. Second, the accounts of content and publicity suggested above provide a sketch for how any theory that accommodates knowledge effects might be able meet these criteria. Overall, concepts are probably not embedded in theories as described by researchers like Keil (1989) and Murphy and Medin (1985), but it seems clear that this approach comes the closest of any account discussed thus far to hitting the explanatory mark.

2.5 Informational Atomism

One feature common to all of the theories that have been discussed up to this point is the assumption that concepts are structurally complex. Atomistic views, in contrast, start with

the assumption that most, if not all, lexical concepts are completely *un*structured (e.g. Fodor, 1998). This is a controversial position not widely held amongst psychologists or philosophers (Fodor, 1998; Laurence & Margolis, 2002), but it is worth briefly discussing given some of the issues it highlights.

The central motivations for informational atomism are its accounts of reference and compositionality. Compositionality, at first glance, is handled simply. Because concepts are atoms, complex concepts are simply structured groups of such atoms, and the contents of the structures are accordingly a function of the contents of their parts. Content, in turn, is determined by law-like mind/world relations that obtain between unstructured representations and their intentional objects (Fodor, 1998). The best way understand these 'nomic' relations is to view them as counterfactual supporting causal links between some property *P* and the tokening of some concept *C*: *P's* reliably cause *C's* to be tokened (Prinz, 2002). And as should be clear, nomic relations are also what enable Fodor to meet the reference requirement. Because *C's* are reliable indicators of *P's, C* refers to *P.*

The most interesting feature of this characterization of reference is that it includes an explanation of misrepresentation, something few if any of the previously examined theories touched on. Because concepts do not refer to only to those things that reliably cause their tokenings, some explanation must be given for why only *certain* nomic relations are the ones that determine reference (Eliasmith, 2000). Consider, for instance, that on dark nights, foxes might reliably cause DOG tokenings given the close resemblance between the two types of

animal (Adams & Aizawa, 2010).[23] Fodor's (1987) solution to this problem is to propose that the fox/DOG relation is *asymmetrically dependent* on the dog/DOG relation. The reason, in other words, that DOG does not refer to foxes is because the nomic relation between DOG and foxes would not exist if the nomic relation between DOG and dogs did not exist. Misrepresentation is accordingly the result of concepts having multiple, asymmetrically dependent nomic links to external entities.

Other benefits of informational atomism are evident too. Scope is easily handled by stipulating that every lexical concept has a dedicated symbol, and publicity follows trivially from the fact that if people share the same symbols, then they also share the same concepts. The property ascription aspect of content is accounted for by the possibility of the symbols of different formal types possessing the same intentional objects (Laurence & Margolis, 1999). Categorization, finally, is addressed by distinguishing between detection mechanisms and the actual concepts they activate (Fodor, 1998). Because such detection mechanisms vary widely between individuals (e.g some people are blind), and because they are not compositional, Fodor claims that they are not part of the structure of a concept. The way to deal with categorization effects, in short, is to shift the burden of explaining them off of the concept theorist.

Not surprisingly, many theorists have complained that this approach renders concepts explanatorily impotent (see Laurence & Margolis, 1999). If a theory of concepts does not account for categorization data, then an explanation of this data must be provided elsewhere.

[23] In the literature, this is referred to as the "disjunction problem," or the problem of ensuring that a concept like DOG doesn't refer to all of its potential causes such DOG means 'dog' or 'fox on dark night' or 'dog statue' or 'large cat' and so on (Fodor, 1987).

However, given that there are currently no theoretical candidates *aside* from concepts for explaining such data, informational atomism has the unfortunate consequence of reducing the overall explanatory power of any particular theory of mind into which it is incorporated. Another big problem is that no explanation is given of how the asymmetric dependency relations underlying Fodor's account of reference are actually determined (Eliasmith, 2000). It is possible to appeal to things like naming ceremonies, learning histories, and evolutionary functions, but Fodor makes no effort to elucidate matters on this front, so it remains unclear that he has actually provided an explanation as opposed to a description of the dependencies that exist amongst nomic relations (Eliasmith, 2000). A further concern is that asymmetric dependency relations create compositionality problems (Laurence & Margolis, 1999). If the reference of a concept is determined by its asymmetric dependency relations, then the content of a complex concept has to partly be a function of the asymmetric dependency relations of its constituent parts. But, clearly, the relations that explain how one can misrepresent a fox as a pet and a whale as a fish do not combine to create the relations that explain how one can misrepresent a minnow in laboratory tank as a pet fish. I take this to be counter-example to the idea that atomistic concepts compose easily.

Overall, though, the main reason atomistic accounts are held in such poor regard is because they entail the innateness of all lexical concepts. Fodor, for example, argues that because unstructured concepts are primitive concepts, and because primitive concepts must be innate, lexical concepts are therefore innate by virtue of their lack of structure (Fodor, 1998; for discussion, see Laurence & Margolis, 2002). The reason Fodor thinks lexical concepts lack structure is because he disputes that they can be decomposed into definitions,

prototypes, or anything else. The reason Fodor thinks that unstructured/primitive concepts are innate has to do with his account of concept acquisition. Assuming that hypothesis testing is the only available method for learning new concepts, primitive concepts must be unlearned because one cannot formulate the relevant hypothesis in any given case unless *one already possesses the concept*. (Fodor, 1998). For example, learning RED by formulating and testing the hypothesis "that things fall under RED by virtue of being red" requires that one already have RED; otherwise, how could one form the hypothesis (Fodor, 1998, p. 124)? The best conclusion to draw, then, is that such concepts are *not* learned, but rather innate. The unfortunate consequence of this argument is that it entails bizarre conclusions like the idea that the ancient Greeks possessed concepts such as ASTRONAUT and INTERNET, and the idea that we currently possess concepts for things that have yet to be invented and discovered.

Many scholars accordingly view this radical nativism as a reductio ad absurdum of informational atomism (Laurence & Margolis, 2002). I tend to agree with them, and even if I did not, it is quite clear that the theory fails to satisfy all six of the necessary criteria. Categorization is left unexplained entirely, while the lack of an account of the determining factors of asymmetric dependency relations undercuts the ability of the theory to meet the content criterion. The fact that asymmetric dependency relations do not compose also presents a problem. And, finally, given that scope and publicity follow trivially from assigning each lexical concept a unique mental symbol, it's hard to construe these explanatory successes as virtues of informational atomism. Concepts, in all, just cannot be structured in the manner Fodor proposes.

2.6 Pluralistic views?

As mentioned at the outset, one theoretical approach that has become increasingly popular involves the proposal of a pluralistic concept structure. Murphy (2002), for instance, favors that conclusion that "people attempt to form prototypes as part of a larger knowledge structure" while also retaining exemplar memories (p. 492). Similarly, Laurence and Margolis (1999) suggest a preference for combining conceptual atomism with prototype theory. The idea, it seems, is to selectively use different theories to satisfy whichever of the criteria they most readily accommodate. Dual theories, as mentioned, use cores to satisfy the reference and compositionality criteria, and prototypes to satisfy the categorization criterion. Adding exemplars and theories in place of cores accordingly might seem like a good way to hit most of necessary explanatory targets. Notwithstanding the question of whether such a hybrid can actually satisfy the compositionality criterion (theories, recall, do not compose well), I think that the point to notice here is that it is not at all clear what sort of *structure* such concepts would actually have. Is a single concept really a gigantic sum of typical feature representations, exemplar representations, and causal/explanatory knowledge representations? How do these different components connect and relate to one another? There are, in short, so many loose ends that it almost seems that hybrid models abandon the very question that they were designed to answer.

Take the model offered by Laurence and Margolis (1999) as an example. They propose a combination of informational atomism and prototype theory whereon conceptual cores (i.e. symbols) determine reference, compose, and explain publicity, while prototypes account for psychological effects like typicality. There are two problems with this approach.

First, none of the challenges associated with prototypes and informational atomism are resolved by simply merging the two theories together. For example, neither prototypes nor informationally grounded symbols compose well, so there is absolutely no reason to think that a concept structured as both a prototype *and* a symbol would either. Moreover, even if one disregards the arguments against the composition of asymmetric dependency relations and thereby supposes that symbols *do* compose well, there is still the problem of explaining how the prototypes for complex concepts are generated.[24]

Second, there is no explanation of how the prototype and symbol that make up a given concept are related to one another. Presumably, the features that characterize prototypes are made up of concepts, but if so, then prototypes would consist of subparts that are themselves both symbols and prototypes (and one could naturally think of a prototype as a set of symbols). This is a peculiar state of affairs for two reasons. First, it introduces a regress that atomistic views are designed to avoid. If a concept is made up of a set of symbols, each of which in turn is made up of a set of symbols, then it becomes very difficult to demarcate plausible boundaries to the concept's structure. Second, the view yields confusion regarding how symbols are tokened. Presumably, all of the symbols that make up a given concept all enter into nomic relations with their referents, yet it would be highly odd if the tokening of one conceptual 'core' always involved the tokening of other conceptual cores simultaneously. This would effectively prohibit the tokening of one concept independently of various others. Yet one can presumably token DOG without tokening FUR, even if FUR is a

[24] One option for getting around this problem is to suggest that a prototype is not actually part of the structure of a concept, but merely an associated representation that aids in categorization. In this case, though, one has abandoned pluralism in favor of informational atomism.

prominent feature of a dog prototype. I take these points to collectively indicate that a simple marriage of existing theories is unlikely to result in a substantially improved account of the structure of concepts.

2.7 Conclusion

In all, this chapter has served two functions. First, it has provided a brief overview of the existing philosophical and psychological literature on concepts while examining many of the standard arguments for and against particular theories. Second, it has attempted to demonstrate the need for a new approach by highlighting the various challenges that existing theories face. The foundation for such an approach, I contend, is to be sought in a correct characterization of the representation relation between neural populations and their intentional objects. Once this relation is specified, a candidate theory for the structure of concepts should follow.

Chapter Three

Dynamic Holism and the Structure of Concepts

3.1 Introduction

Up to this point, I have gestured repeatedly towards the idea that paying closer attention to how representational content is ascribed to neural populations is the key to developing a successful account of the structure of concepts. What I have not done is spell out the details of this strategy or give much in the way of a description of the proposal it entails. The purpose of this chapter, accordingly, is to make the argument from a theory of content to a theory of concepts in a way that satisfies the explanatory goals outlined in chapter one.

There are three steps involved. First, I describe the theory of representational content being employed and briefly discuss some reasons for preferring it to competing theories. Second, I examine a possible method for using this theory of content to scale up from the level of basic perceptual representations to the level highly complex conceptual representations. One purpose of this section is to situate my proposal more clearly in the context of the aforementioned accounts offered by Eliasmith (2000, forthcoming), Barsalou (1999, 2003, 2008), and Prinz (2002). Third, I make a detailed case for the view that concepts can be identified with representations that bear as their content a set of potential transformations that can be used to yield further contents. I refer to this account as "Dynamic Holism," and its essential claims are threefold: first, that concepts are structured dynamically

through sets of transformations of a single source representation; second, that the content of a concept is specified holistically; and third, that the content of any given concept changes over time.

3.2 A Two-Factor Theory of Representational Content

Debates over the nature of mental content are widespread in contemporary philosophy (see e.g. Adams & Aizawa, 2010; Brown, 2002), and it would accordingly be impossible to consider the full range of available theories. For simplicity, I consider three general types.[25] Causal theories, first of all, propose roughly that mental representations mean what they are caused by. Conceptual role theories, in contrast, propose that the meaning a mental representation is a consequence of the inferences it licenses. And two-factor theories, finally, suggest that *both* causes and conceptual roles are determinants of content. In what follows, I briefly summarize the difficulties faced by each family of theories, and then outline a set of reasons for adopting a specific two-factor theory of content developed by Eliasmith (2000). I also quickly outline some of the technical details of how this theory ascribes content to populations of neurons.

The problems with causal theories should be familiar from the discussion of informational atomism in the previous chapter. In order to avoid the so-called "disjunction problem" and provide an explanation of misrepresentation (Fodor, 1987), causal theorists have to establish a principled method by which to distinguish between causes that *are* content determining, and causes that *are not* content determining. I have already articulated some of

[25] This threefold division is borrowed from Eliasmith (2000).

the challenges involved with appealing to asymmetric dependency relations, and it is likely that the other options available to the causal theorist fare just as poorly.

One could, for instance, postulate a learning period in which a person learns to identify content determining causes (Adams & Aizawa, 2010; Eliasmith, 2000). The problem here is that becomes extremely difficult to specify when the learning period starts and ends, and how the learning actually takes place (Adams & Aizawa, 2010). Another option is to propose that evolutionarily determined representational functions pick out content determining causes (Eliasmith, 2000). Rabbits, for example, might evolve a dog-representing capacity in order to avoid predators. In this case, the problem is that evolving a representational capacity that is responsive to both dogs *and* dog-look-alikes would have equal or greater survival value than a capacity that represents dogs alone. If so, then one has no principled reason to say that the representational function picks out dogs rather than dog look-alikes (Adams & Aizawa, 2010). And since there is nothing unique about this particular case, it is fair to suppose that similar difficulties emerge with other examples of evolved representational functions. Overall, appropriately handling cases of misrepresentation is a challenge for a number of theories that identify content with cause.

Conceptual role theories face an unrelated but equally significant challenge: they cannot account for truth conditions (Eliasmith, 2000). Because the conceptual role of a representation is not stipulated by its causal relation to things in the world, there is no sense in which one can distinguish between right and wrong applications of representations with the same conceptual role (Eliasmith, 2000). For example, if one's WATER concept had its content determined by the inferences it licenses, there would be no possible way to

differentiate between WATER and $WATER_T$, and accordingly no way to decide whether one has correctly or incorrectly applied the concept WATER to a clear liquid substance that is composed of XYZ (Eliasmith, 2000). To put the complaint in more general terms, a conceptual role theory of content cannot distinguish between concepts that apply identical properties to distinct referents. This is a clear violation of the explanatory goals outlined in chapter one, and I accordingly take any account that identifies content with conceptual role alone to be insufficient for the purposes at hand.

Two factor theories are troubled by what is sometimes called the 'alignment problem,' or the problem of ensuring that the truth conditions and conceptual role of a representation are consistent with one another (Eliasmith, 2000). As Fodor (1987) articulates the challenge, nothing in a standard two-factor theory prevents a mental state from having the content "dogs are animals" by virtue of its causal connections to the world, and the content "grass is purple" by virtue of is conceptual role (p. 82). There is accordingly a clear need for a mechanism that ensures that the content assigned to a representation by its causal connections is compatible with the content assigned to it by its conceptual role.

Helpfully, recent work by Eliasmith (2000) proposes just such a mechanism. If one reasonably assumes that neural populations are the vehicles of mental representations, then one can conjoin the relevant external causes of a representation with its conceptual role by describing both in terms of the *same* functional properties of a single group of neurons. Put another way, the postulation of an underlying neurocomputational factor arguably unifies causal and conceptual role factors under a single description and thereby has the benefit of eliminating the alignment problem. A further benefit of this approach is that it has been

successfully employed in modeling neural systems (Eliasmith & Anderson, 2003), and accordingly exhibits some predictive and explanatory adequacy in practical contexts. I take this combination of theoretical and practical virtues to constitute a strong reason for adopting a two-theory of content for the remainder of this essay. In general, I adopt the theory of content described and used in a modeling approach known as the "Neural Engineering Framework" or NEF (Eliasmith & Anderson, 2003), but little aside from the discussion of the neural plausibility and content criteria depends upon this specific choice.

To get a bit more specific, the NEF defines mental representations in terms of both the nonlinear encoding of stimuli into patterns of neural spikes, and the weighted linear decoding of such spikes into the physical variables they represent (Eliasmith, 2003). The encoding procedure is characterized by the 'tuning curve' or response function of a given neuron in relation to a particular stimulus, and by the timing of the intervals between the neuron's spikes in response to changes in the stimulus. To give a simple example, two regions in the brainstem called the nuclei prepositus hypoglossi (NPH) and the vostral medial vestibular nucleus (VN) contain neurons with tuning curves that plot a relation between horizontal eye position and spiking activity (Eliasmith & Anderson, 2003). Accordingly, neurons in NPH and VN collectively 'encode' a measurement of eye position into a pattern of neural responses.

The decoding procedure, also necessary for a full characterization of neural representation, is specified by weighting the responses of each neuron in accordance with how useful it is for describing the stimulus, and by summing these responses over the relevant population and over time (Eliasmith, 2003). The result of this sum is an estimation

of the variable that has been encoded (i.e. the position of the eye).[26] One can accordingly think of the decoding procedure as a method for extracting the content of a representation, since what is being decoded is the set of properties the representation ascribes to its referent. Lastly, it is also helpful to point out that this characterization of encoding and decoding does not limit one to the representation of scalar variables alone. Arbitrarily complex representations involving mathematical objects ranging from scalars, vectors, and functions to vector fields can also be constructed using the NEF (Eliasmith & Anderson, 2003).

Just as important as defining representations, however, is accounting for how representations can be transformed. It must be possible, in other words, to take the content of a representation and change it into to something else in a way that is useful to the cognitive system as a whole. Put this way, transformations can be characterized using the notion of decoding introduced above (Eliasmith, 2003). Rather than using decoders that yield an estimate of the variable or mathematical object a given neural population is taken to represent, one can use alternate (or biased) decoders that yield some *function* of this original variable or mathematical object (Eliasmith, 2003; Eliasmith & Anderson, 2003). In the case of NPH and VN, what are called *representational decoders* would produce an estimate of eye position (e.g., 15 degrees from midline), while what are called *transformational decoders*

[26] Mathematically, the encoding and decoding relations necessary for defining a scalar representation can be characterized in basic form as follows (Eliasmith & Anderson, 2003):

(1) $a_i(x) = G_i[\alpha_i x + J^{bias}]$ (encoding)

(2) $x_{estimate} = \sum a_i(x)\phi_i$ (decoding)

where x is the variable represented, i is an index of the neurons in the population, a is the firing rate of a given neuron, G_i is the response function of a given neuron to an input current, α_i reflects input current as function of the variable x, J^{bias} is the background input current, and ϕ_i is the decoder applied to a given neuron.

would produce some function of this estimate (e.g., the square root of the eye position). Again, it is important to note that this account of transformation can be generalized to accommodate arbitrarily complex representations – a high dimensional vector, for instance, could be transformationally decoded so as to yield an estimate of its inverse (Eliasmith & Anderson, 2003). Moreover, this description of transformation provides a natural way to understand how the alignment between cause and conceptual role is established within the NEF: cause specifies encoding, while transformational decoding specifies conceptual role. So, assuming that a complex linguistic representation can be instantiated by a population of neurons, it follows that the linguistic inferences licensed by this representation can be also described in terms of transformational decoding. The NEF, in short, remains responsible to a two-factor theory of content at any level of representational complexity.

A final point worth mentioning about the NEF is that it also incorporates dynamics into its characterization of the representational and transformational capacities of neural systems (Eliasmith, 2003). This is achieved by adapting modern control theory to describe the representations instantiated by neural populations as state variables in a dynamical system. The reason this is important is that it allows for a theory that is responsive to temporal considerations and thereby avoids certain challenges faced by classicist accounts of cognition that generally ignore time. For details about dynamics in the NEF, see Eliasmith & Anderson (2003).

To conclude, a brief consideration of the main available accounts of mental content suggests that, in the absence of the alignment problem, a two-factor theory can outcompete both causal theories and conceptual role theories. The specific two-factor theory I adopt is

strongly motivated by the details of how neural systems function, and it can describe how content is ascribed to such systems in a mathematically precise and theoretically sound manner.

3.3 Scaling Up to the Level of Concepts in a Representational Hierarchy

One notable feature of the account of content described thus far is that it is has only been explicated in terms of very low-level representations (i.e. those describing proprioceptive states). But since the content of a concept is *prima facie* more complex than the content of a basic perceptual representation, we need some way to bridge the gap between the two. Helpfully, Eliasmith (forthcoming) has developed a description of this scaling process in the context of work on a general account of biological cognition known as the Semantic Pointer Architecture (SPA). I initially consider the technical details of this account in order to describe how it can be used to generate complex representations, and I next consider a couple of different approaches to the challenge of elaborating upon these details to fully account for the criteria from chapter one.

The first idea to highlight from the SPA is that complex perceptual representations can be compressed into simpler summary representations that capture statistical properties of the inputs they derive from. A retinal image of a table in one's visual field, for instance, will initially be encoded as a pattern of activity in a large population of neurons. Through transformations, however, further layers of neural populations can produce increasingly abstract statistical summaries of the original visual input. Eventually, a highly compressed version of the input can be produced. This representation will be instantiated as a high

dimensional vector represented by the activity of a population of neurons at the top of a hierarchical network, and is referred to as a *semantic pointer.* A compressed representation of a table in one's visual field would accordingly be referred to as a semantic pointer of a table image. Similar semantic pointers can also be generated using other modalities such as audition and sensation.

The reason these compressed perceptual representations are called semantic pointers is because they retain semantic information about the states they represent by virtue of being non-arbitrarily related to these states through the compression process. The reason why the representations are called pointers is because they can be used to "point to" or regenerate the representations at lower levels in the compression network. Moreover, any given semantic pointer can be manipulated independently of the network that was used to generate it. A pointer for a table image, for example, could be used in cognitive tasks related to tables without necessarily prompting a reactivation of the richer perceptual representations at the bottom of the relevant compression network. Eliasmith (forthcoming) puts the overall idea this way:

> Semantic pointers, generated by grounded perceptual processing, can be "stripped-off" of that processing and used to both carry shallow semantics and be treated like a symbol. If deep semantics are needed, the semantic pointer can be used to clamp the top layer of the perceptual network that gave rise to it, and the network can re-generate the deep semantics. (p. 108-109)

The point to take away from this discussion as a whole is that compressing perceptual representations into symbol-like entities provides an important first step in scaling up to the level of concepts in a representational hierarchy.

The final step in bridging the gap between basic perceptual representations and concepts involves combining multiple semantic pointers into a single representation. Eliasmith (forthcoming) describes this process in terms of the binding operation known as circular convolution. Leaving the mathematical details aside, circular convolution can essentially be thought of as a function that blends two input vectors into a single output vector. And because a semantic pointer *just is* a high dimensional vector represented by a population of neurons, convolution can easily be carried out using the principle of transformation discussed above. If two neural populations each instantiating a semantic pointer both project into a third population, and the synaptic weights between the populations reflect the transformational decoders necessary to compute a convolution function, then the resulting vector implemented in the third population will accordingly represent the combination of the two input vectors (Eliasmith, forthcoming; see also Thagard & Stewart, 2011). This process can be repeated indefinitely to build up arbitrary complex representations. Finally, such representations are also semantic pointers, because they can be unbound to regenerate the subordinate representations from which they are constructed.

To show how this binding process can produce something resembling a concept, consider a simple example like BOOK. Semantics pointers for visual and tactile images of books could be combined, along with pointers for an auditory image of the sound 'book' and a visual image of the letters 'b-o-o-k'. Additionally, linguistic inferences could be

incorporated by binding in representations like CREATED BY AN AUTHOR or MADE OF PAPER. Each of these further representations could also be decomposed into a constituent set of semantic pointers, perhaps of visual images of a known author, or of an auditory image of paper being folded. In all, by selectively binding together semantic pointers from multiple perceptual and lexical sources, highly complex representations can be created that contain many of the features that we ordinarily take concepts possess.[27] For example, a perceptual semantic pointer that has been bound into a concept can be thought of as an exemplar representation, while a lexical representation like CREATED BY AN AUTHOR can be thought of as partially specifying the theory in which the concept is embedded (see Eliasmith, forthcoming).

One might object here that it is a bit unfounded to claim that lexical representations like MADE OF PAPER can be presupposed to be available to bind into a concept. However, it should be clear that PAPER could be generated on a prior occasion in manner analogous to that described in the paragraph above (i.e. by binding together various paper-related semantic pointers). MADE OF, in turn, could be thought of as a role used for creating role-filler pairs in numerous different representations. Convolving this MADE OF role with the concept PAPER, in this case, would give the role-filler pair MADE OF PAPER. A table concept, in comparison, might utilize the same role representation to form a different role-filler pair, such as MADE OF

[27] This description of the process through which complex representations are generated is bit superficial, primarily due to a concern for brevity. Both circular convolution and vector addition are components of the construction process, as in the following example from Eliasmith (forthcoming): "robinPercept = visual⊗RobinVis + auditory⊗RobinAud…" (p. 166). The '⊗' symbol here indicates the circular convolution operation, while the '+' symbol indicates vector addition. 'robinPercept' is accordingly a semantic pointer incorporating various perceptual representations of robins.

WOOD. Importantly, this notion of binding concepts to roles can be generalized to account for the creation of sentences (Eliasmith, forthcoming). Binding various representations of grammatical roles to concepts can, for instance, create a sentence like "the boy chased the dog" (Eliasmith, forthcoming, p. 137). The concept CHASE would be bound to a 'verb' role representation, the concept BOY would be bound to an 'agent' role representation, and so on.[28]

Altogether, the point of spending all of this time giving an overview of the semantic pointer architecture is to illustrate its ability to provide a possible bridge between the theory of representational content underlying the NEF and an account of concepts. As mentioned, though, this description is primarily technical and does not offer a complete solution to the explanatory challenges posed in chapter one. I accordingly adopt the SPA largely as means by which to approach the dual challenges of explaining content and neural plausibility. So, while the SPA is highly suggestive of a theory of concepts, it nonetheless leaves a number of questions regarding things like publicity, reference, and compositionality somewhat open ended from a psychological and philosophical perspective.

To approach matters more from this perspective, one can think of the semantic pointer architecture as providing a neural implementation of something fairly similar to the accounts of concepts given by Lawrence Barsalou (1999) and Jesse Prinz (2002).[29] A core

[28] Important questions regarding the origin of such role representations could be raised here, but examining these in detail is outside the scope of this paper. Suffice it to say that I think it is likely that these role representations will turn out to be largely innate in some cases (e.g., grammatical roles), while largely learned in other cases (e.g., roles specifying the nature of some entity, as in the case MADE OF).

[29] Eliasmith (forthcoming) states as much, while highlighting some differences between these accounts and the characterization of concepts the SPA suggests (p. 361-363). One primary

tenet of these so-called neo-empiricist accounts is that conceptual processing always involves 'simulation' or the re-activation of multimodal representations that have become stored in long-term memory.[30] On Barsalou's view, concepts just are the 'simulators' that produce such simulations. A simulator, in turn, is constituted by an organized system or "frame" of perceptual symbols that correspond to a category, and by the simulations that this frame produces through the selective transfer of different symbols into working memory (Barsalou, 1999, p. 590). So, just as a semantic pointer unbinds to reproduce various low-level perceptual states it, so too does a simulator. The deep semantic information 'pointed to' by a pointer is likewise akin to information contained within a simulator's frame. One can accordingly think of a semantic pointer, at least in part, as a neurocomputational description of a simulator (Eliasmith, forthcoming).

Prinz (2002) develops Barsalou's account from a more philosophical perspective by adopting the notion of a "proxytype" (p. 148). Concepts, according to Prinz, are proxytypes, and proxytypes are "perceptually derived representations that can be recruited by working memory to represent a category" (p. 149). Put this way, a proxytype can be best understood as a specific simulation of a category, and the capacity to token proxytype can be explained

---

difference is that the SPA avoids any strict opposition to the postulation of amodal symbols in conceptual systems, contra Barsalou and Prinz.

[30] To provide a bit more detail about the nature of simulation, consider the following description offered by Barsalou et al. (2008): "On perceiving dogs, for example, the brain captures modal states in the visual, auditory, and somatosensory systems about how dogs look, sound, and feel, respectively…On later occasions, when representing knowledge about dogs, the brain attempts to reactivate these multimodal states, typically succeeding only partially. The resultant simulation of the brain states associated with experiencing dogs can then be used for a wide variety of purposes, including inference, recollection, language, and thought" (p. 247).

by appeal to a simulator.[31] One difference, then, between the proxytype view and the simulator view is that the proxytype view is consistent with the idea that one can have multiple concepts for single category. For example, when thinking about bicycles on different occasions, one can undoubtedly generate different bicycle simulations. But if this true, then each time one produces a novel bicycle simulation, and hence a novel bicycle proxytype, one also tokens a novel BICYCLE concept (since concepts are proxytypes). I think this result actually counts against Prinz's view to some extent – multiplying concepts to such a degree is not parsimonious – but one could adjust the theory slightly and still get something roughly equivalent to the account of perceptual symbol systems offered by Barsalou. So, it is perhaps best to think of the difference between simulators and proxytypes in terms of their respective explanatory targets – Prinz focuses primarily on giving an account of concepts that meets a number philosophical criteria akin to those set out it chapter one, while Barsalou's goal is in large part to provide a scientific explanation of various observed conceptual phenomena involving things like categorization, inference, and related fMRI data.

Given the commonalities between semantic pointers, proxytypes, and simulators, one might also be inclined to borrow from the theoretical aspects of the work Barsalou and Prinz to fill in any philosophical gaps left by the notion that concepts can be constructed by binding together various compressed perceptual representations. Marrying the three views, in other words, could give one an account of concepts that has the advantage of simultaneously possessing neural, psychological, and philosophical plausibility. I am sympathetic towards this approach, but propose instead to start at the neural level, build upwards, and consider and

[31] Prinz (2002) would likely agree with this assessment given his acknowledgement of the commonalities between his work and Barsalou's (p. 152).

borrow from theories described at the level of psychology as they become applicable. The result of employing this methodology is essentially a philosophical elaboration of the semantic pointer architecture, and as mentioned, one that presents concepts as dynamically structured entities that have their content determined holistically by the system of representations in which they are situated.

3.4 Dynamic Holism Defined and Defended

Dynamic holism is a thesis about the structure and content of concepts, and it is the result of employing the two-factor theory underlying the NEF to analyze the representations built using the semantic pointer architecture. Put broadly, the thesis is this: a concept is a representation that specifies a set of possible transformations,[32] and because these possible transformations are determined relative to other representations in a system, the content of a concept is determined holistically. The main upshot of this thesis is that it makes sense to characterize the structure of a concept *dynamically,* since gaining access to the content that results from the transformations a concept describes is an inherently temporal process. In what follows, I develop a detailed argument for dynamic holism that draws on further considerations of the nature of content and structure.

A few initial observations can help set the stage for the argument. First of all, if one takes seriously the idea that what a given neural population represents can be extracted using a representational (rather than transformational) decoder, then it likely follows that *no single population* can represent all of the content typically associated with the average lexical

[32] To give an example, the concept DOG might be instantiated as a vector that has the potential to be transformed into a vector that represents WAGS TAIL, amongst other things.

concept. Consider an example. If BOOK is hypothesized to be instantiated by a population of neurons that has encoded the result of some binding operation (or some transformation of a prior binding operation), then decoding this population will simply give a vector that is the convolution of two input vectors (or some transformation of a vector that is the convolution of two input vectors). Thus, decoding the population is unlikely to yield content like "made of paper" or "created by an author" unless various further transformations of what the population represents are decoded too.[33] Overall, it is simply unrealistic to think that a single population of neurons can be decoded to extract all of the content of even a simple concept.

Next, one might therefore be inclined to think that the content of a concept is distributed amongst several *different* neural populations that are activated simultaneously. In the case of BOOK, decoding each population would give a different portion of the content of the concept – one population might yield the content "made of paper" while another might yield "created by an author," and so on.[34] The problem with this view is that is that it implies an unrealistic containment model of concept structure. Consider that one can token the concept BOOK *with or without* being aware of the content "created by an author."[35] And if this is true, then it cannot be the case that tokening BOOK is equivalent to the activation of a set of

[33] For clarity, recall that I take the content of a representation to be the set of properties it ascribes to its intentional objects.

[34] It is, of course, unrealistic to suppose that a single population of neurons could be decoded to yield a content like "made of paper" given the preceding discussion of the SPA, but the present discussion is not impacted significantly by this oversimplification.

[35] To provide some empirical evidence for this claim, consider that in experiments involving word association and property generation tasks for various concepts, subjects consistently perform more quickly on the association tasks than the property generation tasks (see e.g. Barsalou et al., 2008). This indicates that people access different contents associated with a concept at different times. This evidence arguably undermines the plausibility of containment models in general.

neural populations that can be collectively decoded to yield all of the contents associated with BOOK. Some of these contents will simply not be apparent on most of the occasions that one thinks about books. So, while identifying a concept with a single neural population yields too little in the way of content, indentifying a concept with a whole *set* of populations likely yields too much.

A natural solution to this problem is to adapt a containment model so that only a variable subset of the populations instantiating a concept are activated at any one time. But given that the NEF describes representations in terms of encoding and decoding procedures, I think it makes more sense to describe such variable tokenings in terms of *transformations* that are licensed by a concept. To explain, suppose that one neural population instantiates a representation with the content "book" (i.e. the sound of the word), while another population instantiates a representation with the content "made of paper." In order to account for the fact that a tokening of BOOK can yield one, both, or neither of these contents, it makes the most sense to suppose that both populations instantiate *possible* transformations (or biased decodings) of a single representation that is the source of all book-related content. Depending on how this source representation is transformed (and on how *these* transformations are subsequently transformed), one can derive any number of what one might call different 'unpackings' of BOOK. Putting things this way, one is reminded of the semantic pointer architecture, since a high-level semantic pointer would be a natural candidate for the single representation that enables such unpacking.

Lastly, it is worth observing a distinction suggested by Mendelovici (2010) between the *source content* and the *derived content* of a concept.[36] Source content in this case would be the content of the representation that licenses the various transformations that 'unpack' BOOK. Derived content, in contrast, would be the content of the various representations that result from these transformations. Strictly speaking, I use the term 'content' hereafter to refer to source content, since I have already established that the representations we are interested in characterizing are the representations that are the source of the various 'unpacked' contents implicated in conceptual processing.

These initial observations about neural representation and content are suggestive of the following formulation: a concept, to a first approximation, is a representation that possesses as its content *a set of possible transformations* which can collectively yield all of the further content necessary for things like inference, categorization, and communication. This conclusion is, of course, entirely consistent with the SPA, but it illuminates something important about the content of those representations deemed concepts. Put simply, decoding the population of neurons that instantiates a given concept will not immediately yield any content that could be used to describe what the concept refers to. Rather, such decoding would simply give an estimate of a vector that has what we might call the *transformation potential* to yield further contents that do actually describe what the concept refers to. BOOK, for instance, would be represented only as a vector that has the potential to be transformed to

[36] I adapt Mendelovici's (2010) usage, since she defines source and derived content in the context of a comprehensive theory of mental representation that differs in important ways from the account being developed here.

yield other vectors that correspond to representations like CREATED BY AN AUTHOR. The content of concept, in short, is rather uninformative without further modification.

With the basics of content firmly established, it is now possible to explain why the account of concepts being developed here is inherently holistic. For one thing, in order for a cognitive system to realize the potential transformations specified by a concept, the system must also contain a variety of *other* representations that enable these transformations to take place. Recall from the description of the semantic pointer architecture that complex representations are created by binding together various other representations. If one assumes that a concept is constructed through such binding procedures, then the only way to realize the transformation potential of a concept is to carry out various *un*binding operations that yield those representations that were initially conjoined. And in order for these operations to take place, the system must counterintuitively *already possess* each representation that can be derived as a result of unbinding.

The reason why these derived representations have to be present in the system in order for the necessary unbinding to take place has to do with the technical details of reversing a convolution operation. Convolution, recall, involves combining two input vectors into a third, output vector. The purpose of reversing this operation is, accordingly, to reproduce one of the input vectors, and doing this is best achieved by convolving the output vector with the inverse of the other, unwanted input vector (see Eliasmith, forthcoming). More formally, one can characterize this process using three vectors **A**, **B**, and **C** as follows (Eliasmith, forthcoming):

(3) $\mathbf{C} = \mathbf{A} \otimes \mathbf{B}$ (Binding of vectors **A** and **B**)

(4) $\mathbf{A} \approx \mathbf{C} \otimes \mathbf{B}^{-1}$ (Unbinding of vector **C** to recover vector **A**)

The important point to draw from this description of binding and unbinding is that realizing the potential transformations from **C** back to **A** or **B** *requires* both **A** and **B** to already be represented and then transformed (to get either $\mathbf{A}^{-1}$ or $\mathbf{B}^{-1}$) somewhere else in the neural system. Put more generally, the transformation potential of a concept (e.g., **C**) only exists relative to the presence of other representations (e.g., **A** and **B)**. And finally, since this transformation potential just is the content of a concept, one can likewise say that the content of a concept only exists relative to the content of other representations in the neural system in which it is instantiated

Having established this much, one could now opt either for molecularism (i.e., that the content of a concept depends on *some* of the other representations in the system in which is instantiated) or holism (i.e., that the content of a concept depends on *all* of the other representations in the system in which it is instantiated). Some further consideration of the way in which the transformation potential of a concept has been specified, however, indicates that holism is the preferable choice. Consider an example. Suppose SPIDER is represented as a vector that can be transformed through various unbinding procedures of sort just described to yield further representations such as HAS EIGHT LEGS, OCCASIONALLY POISONOUS, EATS INSECTS, PRODUCES SILK, and so forth. In order for these representations to emerge out of transformations of a SPIDER vector, they all have to be present elsewhere in the system so that SPIDER can be convolved with their inverses in the necessary manner. More importantly, the transformation potential of, say, LEG could also be contained in the transformation potential of SPIDER, since SPIDER could be unpacked in a multitude of ways that incorporate

representations derived from LEG (e.g. LEG could be unbound to yield further representations like BODILY APPENDAGE USED FOR MOBILITY, so SPIDER could be unbound to form HAS EIGHT BODILY APPENDAGES USED FOR MOBILITY). This indicates that the representations necessary for unpacking LEG are also relevant to the determination of the content of SPIDER, since the *full* transformation potential of SPIDER can only exist relative to these further representations.

If one generalizes from this example, it follows that essentially *every* representation in the neural system in which a concept is instantiated is relevant to the determination of the content of that concept. SPIDER could be unpacked to yield EATS INSECTS, while INSECT could be unpacked to yield HOUSEHOLD PEST, while HOUSEHOLD could be unpacked to yield FAMILY RESIDENCE, and so on. Presumably, if one were to continue unpacking in this manner indefinitely, one could access essentially any representation that is present in the system. And if so, then holism follows trivially, since all representations that can be accessed via the unpacking of a concept will be relevant to the determination of the content of that concept. To a second approximation, then, a concept is a representation that possesses as its content *a set of possible transformations* that are specified relative to *all other representations* in the neural system in which the concept is instantiated.

Given this argument for holism, an argument for the dynamical aspects of concept structure follows in a fairly straightforward manner. One can think of concepts as being dynamic in two distinct senses. First, since the content of a concept is best understood as a set of potential transformations, one can think of a concept as being structured as a set of sequential processing states that realize some subset of these transformations. Second, since the set of possible transformations that a concept represents is bound to change over time

(e.g., either through binding the concept with some further representation as a result of learning, or through changes to the other representations in the system), the content of any given concept will never be, strictly speaking, static. Elaborating on these two claims is the purpose of the remainder of this section.

First, consider the idea that a concept is structured as a set of sequential processing states that realize a set of transformations. This is a fairly significant departure from the dominant accounts of structure in the concept literature.[37] Typically, concepts are thought to be structured as sets of representations rather that as sets of processes. The crucial advantage to characterizing structure dynamically, however, is that it is most consistent with the description of content outlined above. If one were to adopt the set-of-representations view of structure, one would be committed to the containment model wherein a concept is instantiated by a set of neural populations that can be collectively decoded to yield a wide range of content. This model, recall, was dismissed on the grounds that it is insufficiently flexible to account for the fact that a concept can be tokened *with or without* tokening any range of other related contents. So, if our account of content emphasizes sets of possible transformations, then identifying structure with process is favorable to identifying structure with a collection of representations.

In some ways, however, the separation between the set-of-representations view and the process view is a distinction without a difference. According to Eliasmith (forthcoming),

[37] Note, however, that this characterization of structure in terms of transformations is closely related to the account of structure offered by theory theories. The theory theorist, recall, identifies the structure of a concept with its inferential role, which is not very different at all from saying that a concept is structured as set of sequential transformations of a source representation.

the account of representation underlying the NEF and SPA naturally suggests that representations "are best thought of as often temporary processing states of activity, rather than as objects that reside in a specific location in the system" (p. 360). If so, then a representation *just is* a process of sorts, and one could accordingly translate between process-based and representation-based descriptions of structure. However, given that representations are commonly thought of as static symbols rather processing states, adopting a process-based description helps to highlight where this view departs from conventional accounts.

Next, consider the idea that the set of possible transformations a concept represents is continually being modified over time (i.e. that the content of any given concept is always changing). In one sense, this is a relatively uncontroversial claim given that the study of conceptual change is an active area of research in both developmental psychology and the philosophy of science (see e.g. Carey, 1985; Thagard, 1992). In another sense, though, it is a highly controversial claim given that so many theorists think that a high degree of concept stability is necessary to enable communication and to explain the predictive adequacy of propositional attitude ascriptions (e.g. Fodor, 1987; Fodor, 1998; Laurence & Margolis, 1999; Prinz, 2002; see Chapter 1). I argue below that a dynamic account of content can support a *sufficient* degree of concept stability, but for the time being, an example can be used to illustrate why conceptual content varies temporally.

Suppose one learns the concept CANCER from public health advertisements or as a result of the diagnosis of an ill family member. Initially, CANCER might have as its content a set of possible transformations that would yield representations like ILLNESS THAT GIVES ONE A TUMOR or ILLNESS THAT IS TREATED WITH CHEMOTHERAPY. Later, however, one might learn

that cancer is caused by the uncontrolled division of cells, in which case the vector representing CANCER would be convolved with other vectors such that the transformation potential of CANCER would come to include CAUSED BY THE UNCONTROLLED DIVISION OF CELLS. Learning, in short, involves conceptual change. And as mentioned, if holism is true, then *any* change that takes place in a cognitive system will subtly affect the content of representations throughout the system. So overall, one cannot have either holism or concept-related learning unless one is willing to accept that the content of a concept varies over time.

At this point, I conclude by offering a brief summary of dynamic holism. Concepts, as a final approximation, are representations (i.e., vectors) that specify a transformation potential relative to the other representations in a neural system. Accordingly, the content of a concept is a set of potential transformations, and the structure of a concept is the sequence of processing states that realizes these potential transformations. Because the potential transformations a concept represents only exist relative to other representations in the system, the content of a concept is specified holistically. And, because holism entails that a concepts' content changes whenever the system changes, the content of a concept is inherently dynamic. A dynamic and holistic account of content combined with a dynamic account of structure yields dynamic holism.

3.5 The Criteria Revisted

Having described dynamic holism and offered a few key arguments in its favor, it is now necessary to evaluate how well it fares on the criteria from chapter one. Doing so serves two functions: first, the evaluation can hopefully demonstrate that dynamic holism is a viable

option alongside classical views, prototypes views, exemplar views, theory theories, and informational atomism; second, the evaluation will help to elaborate on a number of aspects of the account that have not yet been given due attention. The six relevant criteria, recall, are scope, publicity, categorization, content, compositionality, and neural plausibility.

3.5.1 Scope

Scope is handled relatively easily. Since the content of a concept could be any number of sets of potential transformations, it is conceivable that a single high dimensional vector could represent even the most abstract concepts. Take DEMOCRACY as an example. Presumably, the vector representing DEMOCRACY would specify a set of potential transformations that could result from various unbinding operations to yield numerous other representations like SYSTEM OF GOVERNMENT, VOTING, ELECTED REPRESENTATIVES, ORIGINATED IN ANCIENT GREECE, EXAMPLES: CANADA, BRITAIN, NORWAY, and so forth. Moreover, because any number of adjustments can be made to the vector representing DEMOCRACY, the potential transformations specified by the vector can be made as complex or as simple as one's understanding of democracy warrants. Similar examples could be constructed to account for non-referring concepts, mathematical concepts, and concepts for theoretical posits.

3.5.2 Publicity

Publicity presumably presents one of the most significant challenges to an account of concepts that is holistic and treats structure and content dynamically. As Fodor (1998) and

others argue, if everyone has slightly different concepts, and everyone's concepts change from one point in time to the next, then communication and generalized intentional explanations of behavior become impossible. Dynamic holism is surely open to this criticism, but I have already sketched a response in the discussion of theory theories from the previous chapter.

First of all, establishing strict concept identity across individuals and across time is simply an unrealistic benchmark. Non-identical concepts can nonetheless be very similar (by specifying a very similar set of potential transformations), and as mentioned earlier, the conditions for concept possession can be set through reference to communicative efficacy. To elaborate, consider BACTERIA once again. A child's version of this concept would likely specify a relatively small number of possible transformations relative to an adult's version or a bacteriologist's version. Accordingly, the child will likely misidentify bacteria if probed with various descriptions of microscopic life forms, while the bacteriologist will likely not. Nonetheless, the bacteriologist will surely understand what the child means when he says "Wash your hands to get rid of bacteria" and the child will surely understand the bacteriologist when she says "There are bacteria in this Petri dish."

The question, then, is how many (or which) of the potential transformations characterizing BACTERIA must be shared in order for these two people to mutually possess the concept.[38] At minimum, I think it would be possible for someone possess BACTERIA even if the representation instantiating the concept only specified a few potential transformations. Primary amongst these would likely be a simulation of the sound of the word 'bacteria.' But

[38] Imposing necessary transformations amounts to definitionism, so I ignore this variant of the question.

even if the only other transformation yielded something like KILL WITH LYSOL IN THE KITCHEN, one could still be communicatively efficacious because most people would immediately infer from a conversation that the things being killed with Lysol in the kitchen are bacteria. Moreover, even granting that various misunderstandings might nonetheless arise, it is also likely that failures in communication function to prompt learning. For instance, one might ask, "What do you mean?" or "Are you talking about bacteria?" and thereby align his or her concept more closely with that of the rest of the linguistic community. So, in cases where communicative efficacy fails, it is most likely quickly re-established through concept revision on the part of the less competent individual.[39]

And finally, even if none of the preceding argument is persuasive, consider that non-identity amongst shared concepts is likely *necessary* for development in domains such as science. If everyone had strictly identical concepts, then it would be impossible for, say, a theoretical physicist to develop a revised understanding of BOSON and subsequently inform his colleagues. To give another example, it is not even clear that a concept like WATER could change throughout history (as it undoubtedly did) unless various people developed slightly different water concepts that subsequently became widespread. Overall, so long as communicative efficacy is achieved, concept possession follows, and the publicity criterion can accordingly be satisfied.

3.5.3 Categorization

[39] This glosses over the fact that people might still disagree after questioning, or be unable to achieve mutual understanding, but I think such cases would either be rare or restricted to abstract or theoretical concepts that are the focus of intense debate.

At first glance, it is not all clear what sort of account of categorization the dynamic holist can offer. If concepts are simply vectors with transformation potential, then it remains an open question how perceptual inputs could be matched to these vectors, or how one could explain things like typicality and exemplar effects. Here, however, one can happily defer to the SPA. Perceptual inputs, recall, can be compressed into abstract representations that carry shallow semantics. These representations, in turn, can then be bound together to form a high-level semantic pointer. So, in the case of an encounter with a novel entity in need of categorization, one could form a semantic pointer on the fly that combines all of the available perceptual information describing the entity (along with descriptive lexical representations too) and then match this representation to the concept is it most (or sufficiently) similar to.

To explain, since the newly constructed semantic pointer would be represented as a multidimensional vector, and concepts are also represented as multidimensional vectors, one can think of categorization as an operation that finds the preexisting vector that is smallest distance from the input vector in a vector space (see Eliasmith, forthcoming). In many ways, this is reminiscent of the categorization method employed by dimensional prototype models, but the key difference is that incorporating various linguistic representations into the relevant vectors can result in similarity calculations that measure something akin to *theoretical* similarity (as opposed to superficial perceptual similarity).

Moreover, typicality effects can be accounted for by the possibility of a vector being generated and bound into a concept that is the mean of a range of vectors representing perceptual instances of the concept's category (Eliasmith, forthcoming). Put another way, one of the possible transformations a concept specifies could be the generation of a

prototype. Exemplar effects are also easily handled given the previous discussion of compressed perceptual representations. Since a semantic pointer generated from the perception of a category instance can be bound into a concept, it would also be natural for some of possible transformations encoded by a concept to specify the simulation of an exemplar. Knowledge effects, as mentioned, can be accounted for by the wide range of linguistic representations that can result from the set of potential transformations a concept represents. Such representations might describe the 'essences' of natural kinds, to give a simple example.

Overall, since dynamic holism is an elaboration of the SPA, and the SPA has all of the requisite tools needed to explain categorization, the categorization criterion can be met satisfactorily. It might be objected here that I have simply adopted a pluralistic view that tacks prototypes and exemplars onto semantic pointers, but the important point to notice is that there is a big difference between an account that identifies the *structure* of a concept with a prototype or an exemplar, and an account that identifies a concept with a representation that can be transformed to *generate* a prototype or an exemplar. The account described here is more consistent with latter of these two options, and is therefore not best described as a pluralistic amalgamation of existing theories.

3.5.4 Content

Almost all of the initial motivation for developing dynamical holism stemmed from an elaboration of the account of content underlying the NEF and the SPA. But it should be clear that there are a number of unresolved issues regarding the ability of a concept to refer to

a set class of entities, and the ability of a concept to ascribe certain properties to these entities. For example, it seems odd to think of a concept as *ascribing* a set of potential transformations to its referents. Moreover, it is not clear how these potential transformations could even pick out the right referents.

The first point to notice here is that the potential transformations represented by a concept can easily be thought of as specifying a miniature theory or explanatory schema of a category.[40] And I have already argued that theories do a good job of specifying reference for two reasons. First, it makes no sense to say that a concept like BACTERIA refers to things that are not captured by even the most up-to-date theories of bacteria because the boundaries of the concept's extension would become epistemically inaccessible. Second, even if concept possession exists on a spectrum ranging from minimal to maximal competency, the fact that people can revise and adapt their concepts through learning and communication suggests that, over time, most people's concepts will converge on a fairly close approximation of the proper extension.

Next, to account for cases in which concepts ascribe distinct properties to the same referents, one can appeal to the possibility of representations specifying different sets of potential transformations that nonetheless pick out the same things. Take BIRD and AVES. Both apply to all and only birds, but they could each specify potential transformations that yield representations that characterize different features of birds. For instance, transforming BIRD would likely yield representations of the dominant characteristics of birds, such as wings, feathers, and beaks. Transforming AVES, on the other hand, would likely yield

[40] This amounts to saying that the content which ascribes the properties that describe a given category is actually derived content, and not the content of the concept proper.

representations that relate birds to a set of taxonomical categories (e.g. MEMBER OF THE PHYLUM CHORDATA).

One might object here that I have not provided an explanation that can cover Twin Earth style cases. Specifying reference through appeal to theories shies away from externalism to some extent, so this is an important objection, especially considering the commitment to a two-factor theory of content expressed earlier. There are, fortunately, two available responses. First, it may be the case that one is not aware of the difference between H_2O and XYZ due to an ignorance of current theories about the clear, colorless liquids that fill lakes and streams. If so, then one is merely in possession of a degenerate water concept (i.e., one is minimally competent), and any situation in which he or she misapplies WATER to XYZ is an understandable case of misrepresentation. Second, if *no one* is in possession of an explanatory schema that can discriminate between H_2O and XYZ, then we can simply say that there is a metaphysical distinction between WATER and WATER_T (since tokens of each concept are caused by different things), but no epistemological distinction (since no one can tell the difference between these causes under any present conditions). So overall, Twin Earth cases pose no particular problem.

3.5.5 Compositionality

As a slight caveat for what follows, it should be noted that none of the accounts considered in chapter two had a fully adequate explanation of compositionality. The reason, it seems, is finding such an explanation is a difficult and complex task. So, while I think that

some unresolved issues are bound to remain after the discussion offered here, I also think that dynamic holism offers some interesting insights into how concepts combine.

In one sense, the SPA has already given an account of compositionality in terms of the convolution of vectors. For example, in the case of BOOK, it was argued that binding together various compressed perceptual and lexical representations of books could produce a representation that functions like a concept. One might have to concede that this is a demonstration of representational compositionality rather than conceptual compositionality (since BOOK is not a compound concept), but I think that this approach can be extended to account for paradigmatic combinations like PET FISH and DOG SHOW.

To start, consider that when concepts are combined on the present account, all that is being combined are two vectors that represent sets of *potential* transformations. If so, then the sort of emergent features that posed problems for prototypes, exemplars, theories, and definitions only manifest themselves when a compound concept is *unpacked*. So one can say that it is the *derived* content of compound concept that yields the features of that have traditionally been difficult to explain. Moreover, the manner in which any given concept is unpacked is largely dependent on the system in which it resides – recall that the potential transformations a concept represents are only realizable in conjunction with other representations. This suggests a novel approach to characterizing compositionality: concepts combine in a straightforward manner through convolution, but the dynamic structure of a compound concept emerges in novel ways due to the interaction of the compound representation with the system of which it is a part.

The crucial advantage of this approach is that essentially shifts the burden of explaining compositionality off of individual pairs of concepts and onto the system as a whole. To illustrate, one can think of a compound concept as a vector waiting to have the system in which is situated impose various transformations on it. Depending on the nature of the system, the potential transformations specified by a single vector can change in numerous ways. In one sense, this means that the other representations in the system embody all of the background knowledge that determines how concepts compose. If these other representations collectively specify contents that fail to cohere with the initial unpacking of a complex concept, then they can be used to extract a different unpacking that *is* coherent.

To illustrate this idea with an example, consider the combination of PET and FISH. Individually, these two concepts would likely specify a set of fairly predictable potential transformations. PET could unpack to yield DOMESTICATED ANIMAL, LIVES IN HOUSE, EXAMPLES: GOLDEN LAB, SIAMESE CAT, CANARY, and so on, while FISH could unpack to yield LAYS EGGS, LIVES IN WATER, BREATHES THROUGH GILLS, HAS FINS AND SCALES, and so forth. Now assume that the two vectors specifying all of these potential transformations are convolved to form a third vector. The problem, it seems, is that any unbinding operations applied to this third vector will likely only re-access those representations initially yielded by the transformations encoded in PET and FISH. Unpacking PET FISH, in other words, might yield contents that describe pets and fish, but it will not easily yield contents that describe emergent features like the fact that pet fish live in tanks.

The way around this problem is to note that unpackings of PET FISH that are conducted in this manner will eventually yield incoherent groupings of contents. To give a simple

example, if one were to unpack PET to yield LIVES IN HOUSE, and then unpack HOUSE to yield STRUCTURE SHELTERED FROM WATER, one would produce a conflict given that FISH can be unpacked to yield LIVES IN WATER. Creatures that live in water cannot live in structures empty of water, so if PET FISH is compositional, then some additional potential transformation must be available that eliminates this conflict. One candidate for this additional transformation would be a movement from WATER to CONTAINED IN TANKS, BOTTLES, GLASSES, ETC. Unpacking TANK, in turn, would yield a variety of contents indicating that tanks are exactly the right size, shape, and appearance to house a domesticated fish. Accordingly, one could go on to derive a representation like LIVES IN A TANK WHICH CONTAINS WATER to modify or supplement the original representation of LIVES IN WATER, and thereby reestablish coherence amongst the various representations that have been unpacked. Lastly, rather than go through a similar derivation every time the concept is tokened, one could simply bind this representation of an emergent feature back into the concept through convolution and thereby make it more easy to access in the future.

The point, then, is that the system of representations in which PET and FISH are situated will prompt the derivation of new representations that describe the emergent features of PET FISH. A composition function, in other words, is specified by the coherence relations that hold between *all* of the representations that can be accessed through transformations of a vector representing a combined concept. And given that this means that all of the representations in the system are relevant to understanding the composition function, concept combination is better described as a phenomenon involving an entire system of representations rather a simple pair.

Finally, given that people are likely to share many of the same representations and possess concepts that specify many of the same potential transformations, they are also likely to derive the same emergent features when combining concepts independently. Although the question of how a given individual will combine any two of concepts is ultimately an empirical one (since the answer depends on the what representations the individual already possesses), I think the notion that similar representational systems can lead to similar derivations of emergent features provides a natural way to preserve concept publicity in the context of compositionality. More could undoubtedly be said here, but suffice it to say that if compositionality is phenomenon that can only be specified in relation to an entire system of representations, it is possibly even more complex of a phenomenon than is currently acknowledged.

3.5.6 Neural Plausibility

Neural plausibility needs little in the way of discussion. Since dynamic holism is essentially a philosophical elaboration of the semantic pointer architecture, and the semantic pointer architecture relates symbol-like representations to the functional properties of neurons, dynamic holism is fully consistent with the relevant neuroscientific details in a way that few other accounts are. Moreover, Eliasmith (forthcoming) demonstrates that the SPA is able to account for the resource considerations and computational considerations highlighted in chapter one. For example, binding together two vectors sufficiently complex for encoding adult-sized vocabularies only requires about 70,000 neurons, and empirical data indicates that interconnectivity in the cortex supports the projection of a given neuron into any of

approximately 85,000 others. Considering that there are approximately one hundred billion neurons in the brain, these numbers leave ample room to support the full range of human conceptual capacities.

3.6 Conclusion

Altogether, I think dynamic holism offers a promising characterization of the structure of concepts. It is highly plausible from a neuroscientific perspective due to its roots in the NEF and the SPA, and it is equal or superior to existing theories when it comes to explaining scope, publicity, categorization, content, and compositionality. Consider each theory in turn. Definitional views offer a decent account of publicity, but they lack sufficient scope, they fail to specify reference properly, and they do not provide an adequate explanation of compositionality. They also fail to explain categorization effects.

Prototypes have the advantage of explaining typicality, but like definitions, they lack scope, they do not specify reference, and they do not compose well. Further categorization phenomena like knowledge effects are also a problem for most models of prototypes. Exemplar models, in contrast, offer scope and a slightly improved explanation of categorization, but content and compositionality remain difficulties. In comparison, Dynamic Holism can account for the available categorization data, explain content and compositionality, and meet the scope and publicity criteria. It is accordingly preferable to views that suggest that concepts are structured as definitions, prototypes, or exemplars.

A similar conclusion emerges from a comparison with theory-based views and informational atomism. Theory theories can explain scope, publicity, and content, but they

cannot easily account for typicality effects or compositionality. Dynamic Holism does an equal job of matching these successes (given that it offers many of the same explanations), and it can improve on the failures. Lastly, with informational atomism, only scope and publicity are easily explained. Categorization is ignored completely and the postulation of asymmetric dependency relations creates a problematic account of both content and compositionality. Again, Dynamic Holism overcomes these explanatory shortcomings. I thereby conclude that it is preferable to both theory-based views and informational atomism.

Finally, even if one is not entirely persuaded by the philosophical arguments I have offered in support of dynamic holism, I still think that the commitment it displays to the available neuroscientific evidence makes it highly worthy of attention. In other words, even if one disagrees with the exact route that I have taken from an account of neural content to an account of concepts, my arguments should nonetheless make it clear that *any* account that emerges from methodology that starts with neural considerations is likely going to look quite different from the dominant alternatives. So, while much work remains to be done, and many questions regarding matters like concept acquisition and language use remain unresolved, I take it that I have shown that the notion that concepts are structured dynamically is plausible entailment of a theory of content that defines mental representations in terms of both cause and computational role.

References

Adams, F. & Aizawa, K. (2010). Causal theories of mental content. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* Retrieved on 5 June 2011 from http://plato.stanford.edu/archives/spr2010/entries/content-causal/

Armstrong, S., Gleitman, L. & Gleitman, H. (1983). What some concepts might not be. In E. Margolis & S. Laurence (Eds.), *Concepts: Core readings* (p. 491-513). Cambridge, MA: MIT Press.

Aydede, M. (1998). The language of thought hypothesis. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* Retrieved on 6 June 2011 from http://plato.stanford.edu/archives/fall2010/entries/language-thought/

Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 577-660.

Barsalou, L., Simmons, W., Barbey, A., & Wilson, C. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Science*, 7.2, 84-89.

Barsalou, L., Santos, A., Simmons, W., & Wilson, C. (2008). Language and simulation in conceptual processing. In M. De Vega, A.M. Glenberg, & A.C. Graesser (Eds.) *Symbols, embodiment and meaning* (pp. 245-283). Oxford: Oxford University Press.

Brown, C. (2002). Narrow mental content. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* Retrieved on 5 June 2011 from http://plato.stanford.edu/archives/win2008/entries/content-narrow/

Carey, S. (1985). *Conceptual change in childhood.* Cambridge, MA: MIT Press.

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

Eliasmith, C. (2003). Moving beyond metaphors: Understanding the mind for what it is. *Journal of Philosophy*, C(10), 493-520.

Eliasmith, C. & Anderson, C.H. (2003) *Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems.* Cambridge, MA: MIT Press.

Eliasmith, C. (forthcoming). *How to build a brain: An architecture for neurobiological cognition.* Oxford: Oxford University Press.

Fodor, J. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind.* Cambridge, MA: MIT Press.

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

Fodor, J. & Lepore, E. (1996). The red herring and the pet fish: Why concepts can't be prototypes. *Cognition,* 58, 253-270.

Fodor, J., Garrett, M., Walker, E., & Parkes, C. (1980). Against definitions. In E. Margolis & S. Laurence (Eds.), *Concepts: Core readings* (p. 491-513). Cambridge, MA: MIT Press.

Frege, G. (1892). On sense and reference. In A.W. Moore (ed), *Meaning and reference* (pp 23-42). Oxford: Oxford University Press.

Gettier, E. (1963). "Is knowledge justified true belief?" *Analysis*, 23, 121-123.

Keil, F. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.

Laurence, S. & Margolis, E. (2002). Radical concept nativism. *Cognition,* 86, 25-55.

Laurence, S. & Margolis, E. (1999). "Concepts and cognitive science." In E. Margolis & S. Laurence (Eds.), *Concepts: Core readings* (p. 3-81). Cambridge, MA: MIT Press.

Margolis, E. & Laurence, S. (2005). Concepts. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* Retrieved on 22 April 2011 from http://plato.stanford.edu/archives/fall2008/entries/concepts/

Medin, D. & Murphy, G. (1985). "The role of theories in conceptual coherence." In E. Margolis & S. Laurence (Eds.), *Concepts: Core readings* (p. 425-458). Cambridge, MA: MIT Press.

Murphy, G. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Osherson, D. & Smith, E. (1981). "On the adequacy of prototype theory as a theory of concepts." In E. Margolis & S. Laurence (Eds.), *Concepts: Core readings* (p. 261-278). Cambridge, MA: MIT Press.

Peacocke, C. (1992). "Précis of *A Study of Concepts."* In E. Margolis & S. Laurence (Eds.), *Concepts: Core readings* (p. 335-338). Cambridge, MA: MIT Press.

Prinz, J. (2002). *Furnishing the mind: Concepts and their perceptual bases.* Cambridge: MIT Press.

Putnam, Hilary (1975). "The Meaning of Meaning." *Philosophical Papers, Vol. II: Mind, Language, and Reality*, Cambridge: Cambridge University Press.

Quine, W.V.O. (1953). Two dogmas of empricism. In E. Margolis & S. Laurence (Eds.), *Concepts: Core readings* (p. 153-170). Cambridge, MA: MIT Press.

Rogers, T. & McClelland, J. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.

Rosch, E. (1978). Principles of categorization. In E. Margolis & S. Laurence (Eds.), *Concepts: Core readings* (p. 189-207). Cambridge, MA: MIT Press.

Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). "Basic objects in natural categories." *Cognitive Psychology,* 8, 382-439.

Smith, E. & Kosslyn, S. (2007). *Cognitive psychology: Mind and brain.* Upper Saddle River, NJ: Pearson Education.

Smith, E. & Medin, D. (1981). *Concepts and categories.* Cambridge, MA: Harvard University Press.

Smith, E., Shoben, E. & Rips, L. (1973). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review,* 81, 214-241.

Smith, E., Osherson, D., Rips, L., & Keane, M. (1988). "Combining prototypes: A selective modification model." In E. Margolis & S. Laurence (Eds.), *Concepts: Core readings* (p. 355-390). Cambridge, MA: MIT Press.

Szabo, Z. (2004). "Compositionality." In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* Retrieved on 4 July 2011 from http://plato.stanford.edu/archives/win2008/entries/compositionality/

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

Thagard, P. & Stewart, T. (2011). "The AHA! Experience: Creativity through emergent binding in neural networks." *Cognitive Science,* 35, 1-33.

Wittgenstein, L. (1953). *Philosophical investigations,* sections 76-78. In E. Margolis & S. Laurence (Eds.), *Concepts: Core readings* (p. 171-175). Cambridge, MA: MIT Press.