BAPTIZING MEANINGS FOR CONCEPTS

by

IRIS OVED

A Dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Philosophy

Written under the direction of

Jason Stanley

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2009

ABSTRACT OF THE DISSERTATION

Baptizing Meanings for Concepts

By IRIS OVED

Dissertation Director:

Jason Stanley

Most people find it *obvious* that concepts like APPLE, DOG, WATER, CACTUS, SWIM, CHIRP, FURRY, and SMOOTH (i.e., lexical concepts) are acquired from perceptual experiences along with some kind of inferential procedure. Models of how these concepts are inferentially acquired, however, force the acquired concepts to be representationally complex, built from, and composed by, the more primitive representations (e.g., GOLD is built from perceptual representations of yellowness, shininess, malleability, and so on). Since at least the time of Plato, philosophers and psychologists have struggled to find complex sets of representations that have the same meanings, definitionally or probabilistically, as these concepts. For example, to think about the property-kind *being gold* is not the same as to think about the complex property-kind *being (probably) yellowish & (probably) shiny & (probably) malleable*.... I call this Fodor's Challenge: Find an acquisition process that is genuinely inferential and yields a concept that genuinely is one of these lexical concepts. Rather than continue the pursuit of a complex representation that has the same meaning as our concept GOLD. I offer a model on which many lexical concepts are acquired from perception and inference, without being built up from, and composed by, the representations involved. The model, Baptizing Meanings for Concepts (BMC), is inspired in part by Saul Kripke's (1970) baptism process for assigning meanings to linguistic terms. Many lexical concepts, according to the BMC, are acquired by inferring the presence a new property-kind, picking out the property-kind in terms of those perceptible features, and then assigning a simple mental term, a concept, for that (purported) property-kind.

DEDICATION AND ACKNOWLEDGEMENTS

For my mom, Anne Oved

My most special thanks goes to John L. Pollock, my life's inspiration and mentor. Working with you has been my life's greatest honor. I thank Jason Stanley for turning my half-baked ideas into a dissertation, and for his persistent and faithful support. I also thank Matthew Stone for showing me how to crystalize my philosophical questions about representation in terms of Artificial Intelligence. Barry Loewer and Brian McLaughlin have also played key roles in supporting my philosophical ambitions. Special thanks to Jerry Fodor and Brian Loar, whose works have profoundly influenced the development of my views. Thank you to my brother, Robert G. Oved, for always being there, and giving me a 'home base'. Finally, thank you to my intellectual and personal friends, for blurring that distinction: Saba Bazargan, Chris Brown, Mike Bruno, LeeSun Choi, Joshua Cowley, David DeVault, Kate Devitt, Mercedes Diaz, Richard Dub, Kenneth Ellison, Spencer Frohwirth, Jeff and Kristy Glick, Adam Goldstein, Angie Harris, Joshua Howard, Joshua Knobe, Keith Lehrer, Ed and Susan Levinson, Alex Perlis, Andrew Sepielli, Ken Shan, Martin Sodomsky, Pär Sundström, Susan Viola, and Tom Walsh.

Table of Contents

List of Figures

Part I: Introduction

This is a dissertation about lexical concepts, with the primary question of how lexical concepts come to have their meanings. This first part (Part I) is the introduction to the dissertation. There are three chapters in Part I. I begin, in chapter 1, by describing the topic of lexical concepts and situating my primary question in the context of other questions in the vicinity. In chapter 2, I highlight the major contributions of this dissertation to the concept debate, and then give an outline of the dissertation. Then, with chapter 3, I carve out the primary question and discuss the puzzle of *intentionality/aboutness/meaningfulness* that is at the heart of the question. In that chapter I also list key terms and distinctions that I use throughout the dissertation. Much of my argument relies on unconventional distinctions, so the reader should not assume that the last section of chapter 3 can be skipped or merely glossed.

1 The Topic of Lexical Concepts

In this first chapter I give an introduction to the topic of lexical concepts as it will be discussed in the dissertation. I also canvas several fields of research to which the issue matters, and then state the questions that arise along the way in the study of how lexical concepts come to have their meanings.

1.1 What are Lexical Concepts?

Reflect on the visual and auditory sensations that strike you now. Almost immediately you form beliefs about the world. Perhaps, you form the belief that there is a cup of coffee and some papers on a table before you, or that there are cars and birds outside of the building. Something similar happens when you are driving to work. You see the road, some cars, trees, and maybe a cactus or a moose. Not only do you see these things, but you see them *as* cars and trees and cacti. When you experience the world perceptually, you usually experience it as containing various entities, which you automatically classify into categories or kinds. When you classify the things you experience into kinds, like *cactus*, *paper*, *moose*, etc., there is something inside of you, some mental entity, through which you think about that class or kind.

It is these mental entities that I am calling 'concepts'. When you think of something as being a cactus, or a car, or a tree, there are differences in the mental entities that allow them to be about the one category rather than another. That is, one mental entity is the one for *cactus* and some other mental entity is the one for *paper*. We can say that it is by invoking your CACTUS concept that you think about the kind *cactus* and it is

by invoking your PAPER concept that you think about the kind *paper*. Moreover, when you think about a particular object as being a cactus, you use a different mental entity than when you think about that very same object as being a plant, or as being alive.

The majority of this dissertation treats concepts as *representations*, but only in a minimal, rather uncontrovertible sense. I take it for granted that thoughts exist as actual entities of some sort (either concrete or abstract, objects, or graspings, or dispositions, or abilities, or properties, or relations, or processes, or events…)[1]. Whatever their nature, they are the entities that human beings have *in here* and are about kinds.[2]

There are several kinds of mental entities that I do not include in my notion of a 'concept'. This is not to say that these things shouldn't be treated as concepts; they are just not among the entities I wish to include in the domain of this dissertation. Among these entities are (a) the things an agent invokes when thinking about particulars, like when you think about *Aristotle*, or *your car*, or the object that you are looking at now, (b) entities that express the agent's logical relations and quantifiers, OR, AND, IF-THEN, SOME, (c) the entities that might be said to encode the rules for thought processing, such as the rules for deduction, induction, and inference to the best explanation, (d) whatever may be said to encode the agent's design goals, (e) emotions, such as excitement and anger, (f) sensations, like the qualitative feel of smells, pains, tiredness, (g) entities involved in perceptual-processing before the perceptual output, (h) perceptual demonstrative representations, such as seeing one object as being on top of another

1 For arguments against the existence of concepts, see Dennett (1987), Stich (1983), Matthews (2007)

2 Andy Clark (2008) seems to depart from this inside/outside distinction, but it isn't really a departure –he just takes thoughts and contents as overlapping in some ways.

object, (i) full propositional thoughts, such as the thought *that* there is a puppy on a table, (j) pre-motor/proprioceptive representations, like the demonstrative motor command to lift an object, and (k) events involved in motor-processing after taking a motor command. Again, I leave it open whether these other entities are significantly different in their natures from what I call 'concepts'; they are simply not among the primary targets of this dissertation.

Any of the mental entities listed above, as well as concepts, can be divided further in two dimensions: conscious versus unconscious (i.e., reflected upon or not reflected upon) and occurrent versus non-occurrent (i.e., *on*, in the sense of being invoked, currently used, or *tokenized*, in thought processing).[3] A 'concept' for me is a *stored* mental entity that is usually unconscious and non-occurent. It can be brought into consciousness and reflected upon, and it is occurrent when and only when the agent is using it in thought processing. [4]

The convention in philosophy and cognitive science is to use all-capital letters when talking about concepts with a given meaning, as in ROAD, CAR, CACTUS, and MOOSE. A common kind of expression in the literature is 'the concept MOOSE' or 'the concept APPLE', to pick out concept tokens of some concept *type*, where the tokens are typed by their meanings. Just like using quotes around a word, as in "road" and "apple", to mention the word, rather than use it, the convention of all-caps helps to distinguish

---

3 These are orthogonal to each other –Not all mental entities being used at a given moment are introspected and perhaps a mental entitiy can be introspected without being invoked.

4 See the distinction I make between a token concept and the *tokenizing* of a concept in section 3.3.

concepts from the categories they are used to pick out, like being a car and being an apple.[5] Extreme cares is needed when using this convention. Do not confuse CACTUS for example, a mental entity that means *cactus*, with the meaning it spells out. Most likely, 'CACTUS' is not the actual shape of any mental entity.

So far, we can say that any mental entity that is about a *kind* is a concept. The focus of this dissertation, however, is *lexical concepts* in particular. These are the mental entities that represent categories that are typically named with a word or morpheme. Examples of lexical concepts are APPLE, DOG, CACTUS, PAPER, CYLINDER, LIFT, EAT, WALK, BLUE, FURRY, SLOPED, BEHIND, BETWEEN, and NORTH-OF. Contrast these with *phrasal* concepts, like FURRY DOG, and SLOPED PAPER CYLINDER. Also contrast them with sensory representations, like a visual experience of an object's color, which usually don't correspond to words in a spoken language. Note that this way of picking out lexical concepts does not commit either way to whether knowing words is involved in having the concepts. It is left open whether, for example, a horse might have a mental entity, APPLE, that is about the kind *apple*.

So the central question about lexical concepts for this dissertation is, *How do lexical concepts come to mean what they mean?* This question breaks into very many further questions that seem relevant. Are lexical concepts innate or acquired after birth? If they are acquired after birth, are they acquired as a result of learning about the world? Do lexical concepts have meaningful parts from which they derive their meanings? What is the relationship between perception and concept acquisition? What about perception

5 Ultimately I reject the practice of typing concepts by their meanings. I also reject typing them by their syntax or causal role, or anything else.

and concept application? What role does action have in making a concept mean what it means? What is the relationship between beliefs about a category or kind and the concept for that category or kind?

1.2 Who Cares?

The question of how lexical concepts come to mean what they mean is of relevance to any field of study that aims to find regularities between kinds. All areas of philosophy and the sciences aim to do so. More than mere relevance, the questions about lexical concepts are of *central* interest to a variety of fields of research. These include the philosophy of mind, the philosophy of language, psychology, cognitive science, and artificial intelligence.

The analytic philosophy of mind has focused its attention on two general phenomena that seem to be peculiar to the mental. One of these is the phenomenon of consciousness and the other is intentionality/aboutness. These are two features that mental entities seem to have and that non-mental entities seem to lack. While it is plausible that the two phenomena interact with one another, most contemporary research tries to focus primarily on one or the other.[6] The puzzles about concepts are mostly in the domain of intentionality, and they are of central focus of that domain, as concepts are the mental entities that are the most uncontroversial examples of entities with intentionality.

Another domain in which lexical concepts are a central issue is what is known as Folk Psychology –the theory, or set of beliefs, that people ordinarily seem to have in ordinary contexts when attributing concepts to themselves and others. When I see a

6 For work on their interactions, see Horgan and Tienson (2002), Chalmers (2004), Crane (1992), and Ludwig (2001).

woman waiting at a train station, for example, and then I see her leaving her seat and walking to the tracks, I attribute to her the belief that a train is approaching, and that the train will stop for her. In other words, I attribute to her beliefs that involve the lexical concepts TRAIN, APPROACH, and STOP. Likewise, when I introspect my own thoughts, while driving to a movie, for example, I might notice myself having thoughts about tickets, popcorn, and soda, and I might then realize that I'm not thinking enough about cars around me. By noticing my thoughts in this way, I ascribe to myself the tokenizing of my stored concepts TICKET, POPCORN, and CAR.

The folk themselves may not care about their attributions of lexical concepts, but philosophers and psychologists take very deep interest in such attributions. Concept researchers notice that human beings seem to be quite successful in predicting and explaining the behaviors of themselves and others on the basis of their concept ascriptions. Just as with the sciences, the success of folk psychology is an indication that folk theories are right. If our folk theories are right, then it should be possible to make folk psychology into a genuine science that uses very similar categories and laws of thought and thought-processing. Folk attributions of concepts then become data for the more rigorous science of concepts.

Cognitive science is the contemporary approach to the science of thought and thought-processing. The approach treats thoughts as analogous to symbols being processed in a computer program, and it treats thinking as analogous to symbol processing by computational programs that are sensitive to syntactic features of the thoughts. Within this framework, questions arise about how mental symbols come to mean what they mean, whether meanings can be part of a cognitive science, and the

relationships between meanings and the syntactic features of the symbols. Cognitive science may or may not turn out to be a rigorous version of folk psychology, but folk psychology is one source for its data. Further data come from experiments that indirectly reveal relationships between external stimuli, internal mental entities, and external behavior. These data then guide the computational theories of the internal mental events.

With the computational approach of cognitive science, lexical concepts become central to the field of artificial intelligence (AI). One of the biggest objections to AI is that its systems symbols are meaningful only to the human programmers and users (see, for example, Searle, 1980 and Dreyfus, 1981). Most of the systems in AI during the 1980s process symbols that are almost entirely detached from their referents. Steven Harnad (1990) calls this the *symbol-grounding problem*. Much AI research has more recently abandoned the goal of building an agent that has symbols that are genuine *thoughts*, focusing instead on intelligent symbol *processing*. It has even become a slogan in AI to "take care of the syntax, and the semantics will take care of itself" (Haugeland, 1979). This perspective brings to the foreground in cognitive science the question of what role, if any, meanings have in thought processing.

2 The Contributions and Plan of the Dissertation

2.1 Three Primary Contributions of the Dissertation

The major controversy about concepts in philosophy and cognitive science has to do with the acquisition and structure of lexical concepts. It seems obvious to most of us that we do not have many lexical concepts innately; we seem to acquire them somehow from our experiences with the world. Moreover, it seems obvious that we acquire them through *inferential* processes –the world and its properties do not seem to just *brute*-causally impose on us our wealth of categories. Of course, we may have been designed with *some* innate representations, like the ability to experience redness and sourness (or representations that immediately follow the experiences), and maybe even the concepts MOTION, FACE, and WATER. We may even be designed to get some concepts from brute-causal (i.e., non-inferential) interactions with the world, like normal brain development, or in the way that the perception of movement causes a newborn duck to imprint on the moving object. It seems highly implausible, though, for most our lexical concepts (concepts that tend in most natural languages to correspond to morphemes), that they are given to us without any inferential procedures. We seem to develop new concepts as we discover new categories in the world through experience. There seems to be some kind of inferential procedure leading us from patterns in our perceptual experiences to the formation of concepts.

How does the inferential acquisition of concepts work? Almost all models of inferential concept acquisition involve combining simple representations into complex ones. Usually, the more simple representations are representations for properties that the

agent believes to be the essential, defining properties for the concept and/or the representations that are used in the recognition of things as falling under the concept. For example, APPLE is usually taken to have a complex meaning that is represented in terms of more simple concepts, like RED, ROUND, SWEET, and maybe FRUIT, which may themselves be reduced even further, ultimately to a set of innate (mostly perceptual) representations.

The trouble is that there does not seem to be any way to define the concept APPLE, or the concept WALRUS, or most any other lexical concept. None of the properties associated with apple-hood seem to be necessary (it is conceivable for an apple to be blue, cubical, or bitter), and there does not seem to be any set of properties that together are sufficient for apple-hood (a thing could be red and round and sweet without being an apple).

Contemporary accounts of such complexity in lexical concepts abandon the idea that the more simple representations make up strict definitions (necessary and sufficient conditions) and instead consider them as representations of prototypical or exemplar instances, with weights and probabilities over properties (see, for example, Rosch, 1978; Barsalou, 1987,1999; Smith, Osherson, Rips, and Keane, 1988). An agent might notice, for example, some objects as being similarly reddish, roundish, and sweetish. The agent may then trivially infer from the objects having each of these properties to their having the conjunction of properties –infer from '(x is Reddish) & (x is Roundish) & (x is Sweetish)' to 'x is (Reddish & Roundish & Sweetish)'. As long as (Reddish & Roundish & Sweetish) is the complex meaning of APPLE, the agent has thereby acquired the concept APPLE.

Again, it turns out to be horribly difficult to defend the thesis, for many lexical concepts, that they have such complexity. The meaning of the conjunction (REDDISH & ROUNDISH & SWEETISH) is not the meaning of the concept APPLE. Being an apple is not the same as (even probably) having that conjunction of perceptible properties. They may be the properties involved in *detecting* that something is an apple, or the degree to which we detect a thing as being an apple, or the properties we believe apples tend to have. But they are not connected to apple-hood by virtue of meaning. Perhaps more properties need to be added to the set, or perhaps the truly essential properties still need to be discovered. Theories that take this approach have become more and more convoluted to accommodate these difficulties. For example, APPLE would be acquired by *being* a composition of more primitive representations. That is, APPLE would come to represent the property of appleness by virtue of being a complex representation that picks out that property.

The trouble is, ever since at least the time of Plato, nobody has found a complex representation that has the same meaning as APPLE, or the meaning of most any other lexical concept. It cannot be part of the *meaning* of APPLE, for example, that apples are usually reddish, roundish, sweetish, fist-sized, and have a core. Apples might have a different distribution over features elsewhere in the world or in other possible worlds, and that distribution of features might be symptomatic of some other fruit elsewhere in this world or in other worlds. This is puzzling because it seems to lead to the implausible conclusion that APPLE is innate. In spite of this, it has to be possible to entertain the concept APPLE without thereby entertaining the concepts CORE or RED or SWEET.

Try as they might, for most lexical concepts, nobody has ever found a clear complex set of properties that plausibly make up the meaning of the concept. Simply put, the set of properties that are believed to be present in things that fall under the concept do not seem to be the properties that make up the meaning of the concept itself. For more arguments against the representational complexity of lexical concepts, see the work of Jerry Fodor (for example, his 1975, 1981, 1998).

Unfortunately, this problem is where Fodor's infamous Lexical Concept Nativism arises. Fodor argues that we have to give up on the idea that most of our lexical concepts have complex meanings. Consequently, he argues, these concepts cannot be inferentially acquired from other representations. As counter-intuitive as it seems, it follows from the logic that most lexical concepts must either be innate or else acquired by merely brute-causal (i.e., non-inferential) processes. To most of us, that conclusion is just unpalatable. As inferential agents, it seems it is among our design goals to arrive at a conceptual scheme that makes the most sense of the course of our experiences. The intuition that we acquire most concepts by reasoning from experience is just too strong to give up so quickly. I call this tension *Fodor's Challenge* –we must find a way to explain how an agent can use other representations to inferentially acquire a concept that means *apple* (or any other lexicalized kind) by either (a) finding a complex concept that means *apple* from which to acquire the concept or (b) showing how an inferential process can yield a representationally simple concept that genuinely means *apple*.

One major contribution of this dissertation is a framework, which I call *Baptizing Meanings for Concepts* (BMC), that answers to Fodor's Challenge by taking approach (b). A version of the baptism idea, coming from Saul Kripke, is generally endorsed in the

philosophy of language. In spite of the striking analogy between the question of how linguistic terms get their meanings and the question of how mental terms do, the analogy is rarely made explicit. Philosophers of mind, even those who endorse the baptism account for the linguistic case, have overlooked the mental version as a solution to the tensions in the debates about concepts. Moreover, the description of the baptism account for assigning meanings to linguistic terms often implicitly *appeals* to the possession of concepts for picking out the meanings to be baptized with linguistic terms.

A second major contribution is to show how the baptism process in the philosophy of language is not merely *analogous* to the baptism process for concepts. The linguistic process *relies* on the baptist's possession of the concept. That is, the baptism of the meaning for a linguistic term cannot work unless the baptizer has in possession a mental way of representing that meaning. The fleshed out account given here for the baptism of meanings for concepts becomes a proper part of a non-nativist version of the linguistic baptism account.

A third contribution of this dissertation is to direct the study of aboutness/intentionality to the key components from which concepts derive their meanings. The property of intentionality is the property that mental representations (but no other representations) seem to have, of being *outwardly directed* and *about* things in the world, in a way that is independent of any further interpretation. That is, non-mental representations, such as words, footprints, paintings, a fuel gauge pointing at 'empty', the ringing of an alarm clock, are representations that are meaningful, but only in a sense that is derivative on someone, some *agent*, interpreting them as meaning or indicating that something is the case. In contrast, mental representations seem to be somehow

intrinsically meaningful, allowing for the meaningfulness of non-mental representations to reduce, with a terminus, to the meaningfulness of mental representations. One implication of the BMC framework is that the intentionality of concepts is derivative, in much the same way as linguistic terms. The foci for the study of intentionality then become the perceptual and motor demonstrative representations for particulars and the agent's using kinds to explain patterns in its demonstrative representations.

2.2 Chapters and Sections

This dissertation divides into five parts. Instead of tracking the order in which the project is developed, or the order in which the project is best understood, the parts divide the chunks of the project itself. To avoid repetition, the material references forward and backward across the parts, chapters, and sections.

Part I: Here I introduce the topic of lexical concepts as will be addressed in the dissertation. I highlight the minimal, mostly uncontrovertible, version of the Computational-Representational Theory of Thought (CRTT) that is assumed in the background. In chapter 1, I distinguish the topic of lexical concepts with as few theoretical assumptions as possible. I canvas several fields of research to which the issue matters. In chapter 2, I highlight the two major contributions of this dissertation to the concept debate. I then outline the parts of the dissertation. Chapter 3 goes deeper into the topic of lexical concepts, carving out the issue of intentionality/aboutness that is central to the issue. I then set up and motivate the (CRTT) background. At the end of chapter 3 I have listed key terms and distinctions that will be used throughout the

dissertation. Since much of my discussion accuses theorists of conflating importantly different notions, and many of my uses of common terms are unconventional, *the section on terms and distinctions should not be skipped*.

Part II: This part of the dissertation steps through the Baptizing Meanings for Concepts (BMC) framework that I offer. In chapter 4 I give an initial sketch of the framework and then I step through some observations that make the general view plausible. Then in chapter 5 I fully spell out the BMC framework, motivating in detail the novel and controversial features of the theory. In the last section I spell out more precisely the major claims of the theory.

Part III: In this part of the dissertation, I step through the other major theories in the study of concepts. Chapter 6 describes what I call *the Building-Blocks framework*, which sets up the main geography for the space of theories. In chapters 7 and 8 I detail the variations on the views within that framework. Then, in chapter 9 I discuss a mental version of a recent theory, Two-Dimensional Semantics, for the meanings of linguistic terms and then show why the approach cannot be fleshed out in a mental version. I also consider some non-representational theories of lexical concepts and show how to fit them within the Building-Blocks framework for the purposes of this dissertation.

Part IV: This part of the dissertation is the defense of the Baptizing Meanings for Concepts (BMC) framework over its opponents. Chapter 10 is a defense of the proposal that lexical concepts are simple in representational structure, chapter 11 is a defense of

the rational acquisition of lexical concepts, in spite of their resulting representational simplicity, and chapter 12 defends the baptism process for rationally acquiring simple lexical concepts. Each of these three chapters is divided into two kinds of defense, first showing how the BMC better explains observations than its opponents and then responding to actual or anticipated objections.

Part V: This is the final part of the dissertation. In chapter 13 I summarize the observations and arguments that support the Baptizing Meanings for Concepts (BMC) framework and project its implications for the fields of research interested in a theory of lexical concepts. I consider some ways in which the BMC fits into the geography of models in the fields of Artificial Intelligence and Robotics, setting up computational experiments to further compare this framework with other theories of lexical concepts.

3 The Set-up

In this chapter I highlight the minimal, mostly uncontroversial, version of the Computational-Representational Theory of Thought (CRTT) that is assumed in the background of this dissertation.

2.2 Representation/Aboutness/Intentionality

As will be shown later in this chapter, it is surprisingly easy to conflate several aspects of representation when articulating theories about it. Let us begin by noticing that token cases of representation involve at least the following four distinct elements:

(1) An entity, *x*

(2) An entity, *y*

(3) A representation-making relation, *Rxy*, by virtue of which *x* is a symbol/representation and *y* is the meaning/content of *x*.

(4) A process, *P*, through which the state *Rxy* gets established.

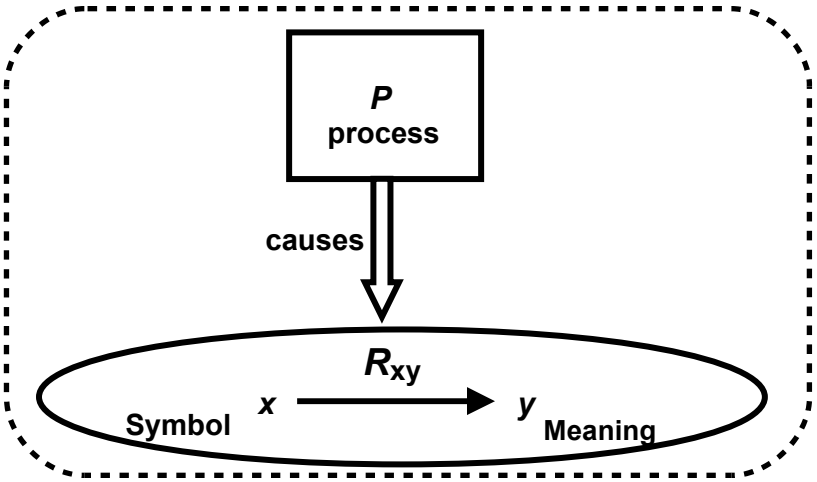


Figure 3.1: the elements of representation

The majority of this dissertation focuses on (3), the process that brings about the state of affairs of a case of concept possession. That state of affairs, element (4) above, and the nature of the relation R, is discussed in Chapter 13, from the perspecitve of the Artificial Inetelligence goal of building a thinking machine. For now, I will assume a commonsense notion for that relation, the notion that we use ordinarily when we treat one entity as representing another entity.

Notice that on this understanding of representation, there need not be something intrinsic to either entity, *x* or *y*, that alone makes one a symbol and the other its meaning. Some relationship between the entities is always involved in making the one a representation for the other. Also, notice that just about anything (at least anything actual) can be a symbol for just about anything (perhaps anything possible), as long as the two can enter into a representation-making relation. Moreover, the four elements do not map 1:1:1:1. Indeed, there conceivably can be any (real) number of symbols, with any (real) number of meanings, in any (real) number of representation-making relations, and any (real) number of processes that result in making it the case that one given entity means/represents another given entity.

Next notice that some symbols, like the no-smoking sign in Figure 3.2, have parts that are themselves symbols, organized in such a way that the meaning of the complex symbol is inherited, largely if not fully, from the meanings of its symbolic parts. This meaning inheritance is part of the representation-making relation between the complex symbol and its complex meaning.

Figure 3.2 Example of a Complex Representation

Let us characterize Simple vs. Complex Symbols as follows:

> ***Simple Symbol***: a symbol that is simple in representational structure, in the sense that it directly represents a simple meaning, rather than having the complex meanings of its representational parts. Simples have no analytic (by virtue of meaning) connections to other concepts.

> ***Complex Symbol***: a symbol that is complex in representational structure, in the sense that it has a complex meaning that is composed from the meanings of its representational parts.

Full theories of representation need to give accounts of the nature of all four of these elements of representation. Plausibly, there are certain kinds of entities that can only represent certain kinds of entities, only by certain kinds of representation-making relations, and can only be acquired by certain processes. For example, a theory might

hold that works of art have to be *objects* that represent *physical entities* (rather than, say, experiences), and can only do so by virtue of an *audience interpreting* the object as having that meaning, where this representation relation can be established by the process of the *artist explaining* that meaning to the audience. It may be different for other kinds of representation. Some argue that iconic representations (photographs, works of art, retinal images) are importantly different from linguistic representations. On the other hand, it is plausible that the iconic/linguistic distinction has to more to do with the representation-making relation, rather than with the nature of the symbols themselves. Indeed, linguistic symbols can just as well be regarded as iconic. They don't *look like* the things they represent, but again that has more to do with the representation-making relation than the intrinsic nature of the symbols themselves. Dretske's (1981) analog/digital distinction and Harnad's (1990) sub-symbolic/symbolic distinction between representations may likewise have to do largely with the representation-making relation, rather than a difference in the formats of the symbols themselves.

This dissertation assumes an important distinction between mental and non-mental representations. There is an intuitive difference in that mental representations seem to have the feature of *intentionality*/*aboutness* –when I think '*Alford is a puppy*', there is something in my mind is *about* a particular object in the world, and something in my mind is *about* the property of being a puppy. Some non-mental representations have aboutness, but it seems they only do so in a way that is *derivative* on the mental ones. In any case, the distinction will lie in some combination of the four elements. It may have to do only with the representation-making relation, or it may have to do with nature of the symbols themselves being *mental* entities. Perhaps for non-mental representations, like

footprints, states of a fuel gauge, or linguistic representations, the representation-makers are statistical/causal/mechanical relationships between the symbol and its meaning (this is plausible for footprints and fuel gauges) or that human beings *use* the entities to stand for certain meanings (as seems to be the case with linguistic symbols). There may be further distinctions to draw between different kinds of mental representation, having to do perhaps with their meanings being *kinds* versus *particulars*, or *properties* versus *objects*. My hope in this dissertation is to give an account of a certain class of mental representation –the lexical concepts. Toward this end, let's start with some distinctions that have already been made by philosophers.

H.P. Grice (1957) makes a useful distinction between two senses of 'meaning' and 'representation'. There is a notion of 'representation' that is used for (merely) causal and probabilistic relationships between naturally occurring phenomena. The sense in which spots *mean/represent* measles (or, rather that measles are present), smoke *means/represents* fire, and footprints *mean/represent* (the recent presence of) feet, is of this (merely) causal/probabilistic sort. Certain states of affairs are *indicators* of, or *raise the likelihood* of, other states of affairs. There is a somewhat different sense of 'representation' in which the no-smoking sign above *means/represents* that smoking is not permitted, the utterance "it's red" means that some object has the property of being red, and the English word 'foot' means/represents the property of being a foot. Grice calls the former "Natural Meaning" and the latter "Non-Natural Meaning". In the spirit of our naturalist pursuit, let us call the former "Indicator Meaning" and the latter "Semantic Meaning", highlighting the semantic/aboutness feature that seems more central to distinguishing them.

Fred Dretske (1988) makes a related distinction between three kinds of representation system. By "representation system", Dretske seems to be talking in part about systems involving all three of the elements of representation mentioned above. Symbols in his Type I system represent their meanings completely by virtue of a human agent *deciding* that a particular entity will be a symbol that represents a particular meaning. Maps, musical notation, the yin-yang symbol, and words in a natural language are of this type. There is nothing intrinsic to these entities that makes them represent, even in the sense of *indicating*, these meanings. However, because human beings use them as symbols, they come to be meaningful in both of Grice's senses of 'meaning'. Representations in Dretske's Type II systems include states of a fuel gauge and the ringing of an alarm clock. For these states to mean what they do, human beings are somewhat involved; humans created and set up these mechanisms to track certain aspects of world. As a result, they come to be meaningful in the sense of being indicators. I do not consider this kind of representation a *semantic* one. Representations in his Type III systems, which he, like Grice, calls "Natural" representation systems, do not require any human involvement for their relationship to their contents. These are naturally occurring indicator relationships. With Dretske's way of distinguishing representation systems, he uses Type III systems to characterize *mental* representations. But when we discuss the BMC theory offered later, we will see that this might not be right at all.

Notice that representations that are meaningful in Grice's Natural sense are not meaningful by virtue of the kind of human intervention that Dretske mentions for Types I and II. Since aspects of the world naturally track other aspects of the world, there are, in effect, naturally occurring *gauges*. Note, however, that this particular theory would be

false on the Gricean assumption that all representations are meaningful only in virtue of some interpreter with meaningful mental representations. Consider tree rings as a reliable tracker of the age of tree. Lack of human intervention on the tracking system, therefore, is not sufficient for characterizing intentional representations. Notice also that as a result of human involvement, the representations in Type II systems mean, in Grice's Natural sense, their contents. And notice that while representations from Type I triples don't on their own mean the contents that humans assign to them, they *do* mean/indicate *some things*, in Grice's Natural sense, simply by virtue of being material objects taking part in the causal world. This all suggests that finding the right way to cut up representation systems is not a trivial task. Nonetheless, let us consider some things that have been said about the three factors of representations, simply in order to establish a better grip on the distinctions.

In effect, *anything* (or, at the very least, anything that exists and is *accessible* to us) can be a representation. At first, the varieties seem extremely broad, since anything that enters in a causal/indicator-relationship gets to be a representation (in Grice's Natural sense). That's not quite right, however, because causation only relates *events*. The sense in which 'smoke means fire' is that *the presence of smoke* is a causal-indicator of *the recent presence of fire*. Still, just about anything (again, that exists and is accessible to us) can be a symbol, because it can be *used by an agent* to mean or stand for something else (representations in Dretske's Type I systems).

3.2 Concepts as Computational Representations

Recall that one of the goals of a theory of concepts is to characterize the elements to be quantified in a science of thought that is sensitive to thought constituents. It is usually assumed that in order to do so we must have some way of distinguishing thought-types, and doing so in part by concept-types. It is also usually assumed that thought- and concept- types are to be determined by what they are about, by their meanings. We will see later that these two assumptions prevent the success of a science of thought and concepts. In order to raise questions about these assumptions, I add to the elements of token cases of representation:

(5) a concept-typing factor, *T*.

When dealing with types and tokens, we must establish a type-determining factor. That is, we must decide by virtue of what two token entities are of the same type. The concepts literature usually types concepts by their *meanings*. I will do the same here. Notice, however, that there are other possible type-determining factors for concepts when understood as symbols. In computer science, symbols are often typed by their syntax –by the shape of the entity and the rules for processing entities of that shape. The literature often uses all caps, as in APPLE, for the concept of apple, where it is unclear whether the claims are about tokens, types, or *tokens of the type*. I may have a concept that means apple-ness and you may have a concept that means apple-ness, so we might say that we both have tokens of the concept type APPLE. So you and I may each have a token concept that is of the concept-type APPLE, in which case we each have a token mental entity that means apple. In this paper I am making claims about concept *tokens* and how they come to mean what they mean.

When theoriests talk about how to 'individuate' concepts, they are talking about the factor that makes a given token concept be a token of the concept type that it is.

Since I am only discussing mental entites with aboutness, I adopt a Representational Theory of Thought (RTT), which might be just part of a more general Representational Theory of Mind (RTM). On the version of RTT that I assume for this dissertation, thoughts are mental entities that are symbols with a lot of structure, and they come to represent, or be about, full propositions at least in part by virtue of their representational structure (see Fodor, 1975, 1979, and Pinker, 1997). A thought, x, is about that puppy out there being on that car, for example, by virtue, at least in part, of x having parts that are representations for parts of the proposition –such as the property of being a puppy, the property of being a car, and the relation on-top-of. This sort of (RTT) is explicitly assumed in much of the contemporary concepts debate and I focus on such theories for the dissertation.

Much of the contemporary debates on concepts treat them as *computational* symbols, where thinking involves computational operations on those symbols. For example, we can consider some entity in a cognitive agent's mind, call the entity ***P***, as a symbol for something else, say, the property of *being a tiger*, such that the agent's entertaining ***P*** is identical with the agent's thinking about that property. If the agent has another mental entity, call it ***N***, that is a symbol for, say, the property of *being friendly*, then the agent's using ***N*** is identical with the agent's thinking about that property. And so on. The agent may then form a complex conjunctive concept, ***P&N***, in order to think about the conjunctive property of *being friendly and a tiger*. Let's call this specific form

of RTT the Computational-RTT (CRTT). This dissertation adopts the CRTT, but the majority of what is said applies to RTTs more generally.

To focus on intentionality/aboutness, I intend to discuss primarily mental phenomena that fit within the perceptual representations, thoughts, and pre-motor representations. I treat these three kinds of mental phenomena as including symbols for *particulars* and symbols for *kinds*, distinguished by syntactic markers available to thought processing. Such representations are important parts of propositional thoughts, as in 'Entity e is of the kind k' and 'Entities of the kind k tend to be entities of the kind j'. Importantly, the distinction between particular-representations and kind-representations is not the same as the distinction between *entity-representations* and *property-representations*. My ontology includes representations for *entity particulars* and *entity kinds*, as well as *property particulars* and *property kinds*. I add to these representations for *relation particulars* and *relation kinds*. Again, all of these are syntactically distinguished by the thought processor. In this dissertation I collapse object-kind and property-kind into simply kind-representations. (Note that the distinction between particular-representations and kind-representations is also orthogonal to the distinction between *representation particulars* and *representation kinds*, which is simply the type/token distinction).

This representation ontology allows us to speak of an agent as having a representation involving an object-particular-representation (*that*) as predicated with some property-particular-representation (*such*), in contrast with a representation involving an object-particular-representation (*that*) as predicated with a kind-representation (APPLE), in contrast with a representation involving a kind-representation

(APPLE) as predicated with a property-particular-representation (*such*), and in contrast with a kind-representation (APPLE) as predicated with another kind-representation (RED).

A concept on this picture is a mental entity that is a *symbol*, and it represents some *content*, where that content is a *kind*. So concepts are kind-representations. Having such a symbol allows the agent to think about the kind of object or property by tokenizing the symbol. According to RTT, complex concepts are complex symbols that inherit their meanings from their parts, in much the same way as in the no-smoking sign (see Fodor 1985 for details on RTM).

There are plenty of good reasons to adopt the CRTT framework. First is the observation that we only entertain a finite number of thoughts in our lifetimes, yet the number of possible thoughts that any human being can have is in(de)finite. The idea that we store some basic set of thought constituents and then compose the constituents together as needed explains this observation. Another thing we notice is that full propositional representations, like FURRY DOGS EAT DOG FOOD and WHITE CATS EAT CAT FOOD, seem clearly to be complex symbols, and acquired by composing their more basic phrasal representations, like FURRY DOGS and CAT FOOD. The phrasal representations clearly break down also, into FURRY, DOG, CAT and FOOD.

The intuitions behind the RTT have been made more precise in the literature, and are known as the systematicity and productivity of thoughts (Fodor, 1975).

The Systematicity of Thoughts: Many mental representations have meaningful parts, and the parts contribute the same meaning to the complex representations in

which they appear. For example, the concept DOG contributes the same meaning to the complex representations FURRY DOG and DOG FOOD.

The Productivity of Thoughts: From a finite set of concepts, we come to be able to have in(de)finitely many complex concepts. For example, from TREE, BIRD, and FOOD, we can have BIRD FOOD, BIRD TREE, TREE BIRD FOOD, and so on.

The computational component in CRTT accounts for the further observation that a lot of our thought processing appeals to rules involving structured parts of thoughts. Our deductive reasoning, most clearly, seems to be sensitive to the constituents of thoughts that the Building-Blocks model posits.[7]

The Logic of Thought Processing: Much of the cognitive processing that operates over thoughts is a lot like formal logic, which is sensitive to sentence parts. We infer, for example, from thoughts of the forms 'P' and 'if P then Q' to thoughts of the form 'Q', and from thoughts of the forms 'All-x, if Px then Qx' and 'Pa' to thoughts of the form 'Qa'.

3.3 How I'm Using Some Terms

In this section I make explicit my unconventional uses of terms from the literature. Important distinctions are drawn and used throughout the exposition and argumentation in the dissertation. First I step through 14 variations on the expression 'Subject S has

7 See Fodor, 1998.

concept C', most of which are importantly different from one another. After that I go through an alphabetical list of key terms that appear in the dissertation.

In various contexts, there are different requirements on knowing what something is. Consider the question, 'Do you know who my father is?'. It can be used to ask if the hearer knows which of the 12 men in view is the speaker's father, or to ask if the hearer has met the speaker's father as such, or to ask if the hearer knows that the speaker's father is the mafia don. It can be used, that is, to ask if the hearer knows that the referent of 'my father' is the same as the referent of 'that man' (pointing him out), or if for some man the hearer has met it has the same referent as 'this man', or the same as the referent of 'the mafia don'.

But now consider, 'Do you know what an apple is?'. It can be used to ask if the hearer knows that apples are a kind of fruit, or if the hearer knows how to pick out apples visually, or if the hearer knows the structure of an apple's core. But more interestingly, we can ask, 'Does a horse know what an apple is?', where we are asking if a horse has a concept of apple-ness.

Notice that 'Do you know what an apple is?' can also be used to ask if the hearer knows what the word 'apple' means in the community. This question is confusing. It can be used to ask if the hearer knows which of her concepts has the same meaning at the word 'apple' (this is a likely interpretation if the hearer has English as a second language). But that seems to be asking whether the hearer knows that the word 'apple' means apple-ness, and that question opens a bucket of worms. What is it to know that the word 'apple' means apple-ness? In all of these interpretations, it seems clear that knowing the meaning of a word requires one's having a concept that means apple-ness.

On the understandings we have here, it should be clear that having a concept of apple is not the same as knowing what makes something an apple. What makes something an apple is an empirical (or perhaps metaphysical) issue about the property of apple-ness. Again, we are distinguishing the concept of apple from the property, apple-ness, that is the meaning of that concept. It may be a theoretical claim that in fact having a concept of apple-ness, i.e., a mental symbol that is about apple-ness, requires one's knowing which empirical or metaphysical features are the ones that make it true of a given thing in the world that it is an apple.

Sometimes 'knowing what makes something an apple' is used to talk about another issue –the issue of what considerations are made psychologically when someone *recognizes* an object as an apple. Perhaps the object's having certain perceptible features, like red-ness or round-ness, are among these considerations. This must be distinguished from what empirically or metaphysically *makes* something an apple. Of course, it may be a theoretical claim that in fact the perceptible features used in recognizing something as an apple is the concept of apple, but it should not be taken for granted.

With the distinctions made here, it should be clear that it almost never makes sense to ask about one's knowing the meanings of one's own concepts. All it can be to know the meaning of one's concept of apple is to know that one symbol in her mind has the same meaning as another symbol in her mind, where in fact the meaning is apple-ness. Representing that meaning, apple-ness, is exactly what is to have a concept of apple. The way we are using 'symbol' here, leaves no room for symbol-independent access to apple-ness. This is not a feature only of the RTM; it is a feature of *any* theory

that is taken to be a theory of the things that we have in our minds when thinking about apple-ness.

3.3.1 Variations on 'Subject S has the concept APPLE'

(1) S has <u>the</u> concept APPLE: I use this for any concept that means the kind *apple*. The danger in the choice of term is that it assumes there is only one way to have a mental term for apple.

(2) S has <u>a</u> concept APPLE: I prefer to use this for any concept that means the kind *apple*. It is less likely than (1) to implicate that there is only one way to have a concept that means apple.

(3) S has a concept <u>of an</u> apple: This strikes me as being interpretable as being the same as (1) and (2) as well as being interpretable as *S's (core) set of beliefs about things that are of the kind apple*. The latter may be an interesting issue but it must be kept distinct from issues (1) and (2).

(4) S has a concept for appleness: I treat this as more-or-less synonymous with (1) and (2).

(5) S has a conception of appleness: This strongly strikes me as being about S's (core) set of beliefs about apples. Again, it must be pre-theoretically distinguished from issue (1) and (2).

(6) S knows what an apple is: This seems to be about the *nature* of the kind apple, in the same sense in which H2O is the nature of water. Again this has to do with beliefs *about* apples. [See Token vs. Type vs. Tokenize below]

(7) S knows the word 'apple': This strikes me as saying that S knows which of her concepts expresses the same kind as the English word 'apple' does.

(8) S knows the meaning of the word 'apple': This is like (7).

(9) S understands the word 'apple': This is the same as (7) and (8).

(10) S understands the concept APPLE: This is sometimes used to talk about the *nature* of appleness (6), but it easily slips into the notion of understanding the *word*, as in (7), (8), and (9), and simply *having a concept* that means apple, as in (1) and (2). Strictly speaking, it doesn't make sense on a representational theory of concepts for an agent to *understand* their concepts. Concepts *comprise* the agent's understanding; an understanding of the *world* is done *via* beliefs that sometimes contain concepts.

(11) S is able to think about apples: This is often used for the same notion as (1) and (2), but should really be replaced with (12), in case S thinks about apples *as pears*!

(12) S is able to think about apples as such: This says that S is able to think about the kind apple. It also slides into saying that S is able to categorize apples correctly, which is different.

(13) S is able to think about appleness: This is a way of saying (1) and (2).

(14) S has a mental entity x that means/represents the property-kind appleness: However unwieldy, this is my favored way of talking about concept possession.

3.3.2 Other Terms and Distinctions

Acquired Symbol*:* a symbol that comes to have its meaning some time after birth. Symbols may be *Rationally Acquired* or *Brute-Causally Acquired*.

Brute-Causally Acquired Symbol: a symbol that comes to have its meaning after birth, but by a non-inferential/non-rational process, for example, by brain development or mere causal contact with entities. Contrast with *Rationally Acquired Symbol*.

Complex Symbol: a symbol that is complex in representational structure, in the sense that it has a complex meaning that is composed from the meanings of its representational parts.

Concept: a stored mental entity that is used by the agent to think about a property/kind. It is the mental thing you bring into occurrent processing when you are thinking about the property of being an apple. I mostly focus on *lexical concepts* when I say 'concept'.

Conception: a particular agent's (most salient) beliefs about the entity in question.

Concept Individuation: the *typing* or generalizing of concept tokens. Concept tokens may be individuated by syntax or by meaning or by any feature of the token. I use single quotes around syntactic items, as in 'Ms are furry'. Words in all caps are for concepts that mean what the word means, so a concept APPLE is any mental entity that means the kind apple.

Content: what a symbol is about or represents. Used interchangeably with *Meaning*.

Designate/Designation/Designatum: To pick out; the picking out of; the thing picked out. Used interchangeably with *Refer/Reference/Referent*.

Innate Symbol*:* a symbol that has its meaning at birth. Distinguish it from *Brute-Causally Acquired Symbol*.

Lexical Concept: a mental representation for a kind that tends, in most languages, to be expressed by a single word when and if expressed. The notion of a 'lexical concept' involves no commitment to the agent's having a corresponding natural language term.

Lexical Property/Kind: a kind that is typically given a simple name in natural human languages.

Meaning, Indicator: An Indicator Meaning of some entity, *x*, is an entity, *y*, where *y* (causally or otherwise informationally) indicates *x*. For example, smoke means fire, in this indicator sense. Rather, states of affairs in which smoke is present mean/indicate states of affairs in which fire is present. Contrast with *Semantic Meaning.*

Meaning, Semantic: Semantic Meaning is used interchangeably with *Content*. Contrast with *Indicator Meaning*. The meaning of a symbol is what the symbol is about/represents. It is surprisingly easy to confuse concepts with their meanings. I suspect this is because most theories individuate concepts (types, that is) by their meanings. You and I can each have the concept APPLE, in that we each have token concepts that mean apple-ness. Whatever your theory of concepts, it is crucial to keep as distinct entities the thing that is the symbol and the meaning (although your theory might make the claim that these are identical).

Rationally Acquired Symbol: a symbol that comes to have its meaning through a rational/inferential process that is sensitive to the meanings of symbols used in the inference. Contrast with *Brute-Causally Acquired*.

Refer/Reference/Referent: To pick out; the picking out of; the thing picked out. Used interchangeably with Designate/Designation/Designatum.

Representation: an entity that is about/represents/means another entity. Used interchangeably with "symbol". It is sometimes used to pick out the whole set of entities involved in cases of representation.

Semantics: Features of a symbol having to do with truth and/or meaning and/or interpretation.

Simple Symbol: a symbol that is simple in representational structure, in the sense that it directly represents a simple meaning, rather than having the complex meanings of its representational parts. Simples have no analytic (by virtue of meaning) connections to other concepts.

Symbol: an entity that represents/is about/means an entity (usually a different entity). It is used interchangeably with *representation*.

Syntax: Intrinsic features of an entity (when treated as a symbol). Processing operations are defined in terms of syntactic features, so the syntax of an entity determines how it interacts with other entities.

Thinking: the mental processing of thoughts.

Thought: a mental entity that represents a proposition (is truth-evaluable).

Token vs. Type vs. Tokenize a symbol: a token symbol is a stored symbol (either stored in the mind of a particular agent or in a community of language users). Token symbols can be typed in many different ways; if typed by meaning the English word 'cat' and the French word 'chat' are of the same type. A symbol is tokenized when it is used. (We are all familiar with this distinction in language – the word-type 'apple' is an

abstraction on the actual, physical instantiations of the word, each token word “apple” that is uttered or written. Likewise, we can say that there is an *object-type* apple, where each individual existing apple is a token. These distinctions are clear. It is easy to slip, however, when we make a distinction between tokens vs. types of *concepts*. We can say that you and I both have the concept APPLE. The convention of all caps is used in the literature when talking about concept-types, what we share when we have the same concept. The trouble is, when we treat concepts as entities that are stored in the mind of the agent, and which the agent uses, or instantiates, during thought processing, we need to introduce a third notion. Let us say that such instantiations by an individual agent of her own token concept are *tokenizing* of her concept token.)

Part II: Baptizing Meanings for Concepts (BMC)

4 The BMC Framework and its Motivations

In this chapter, I sketch a minimal illustration of the framework of Baptizing Meanings for Concepts, and then I step through some observations that give it a lot of intuitive plausibility. The details of the framework, as well as its more precise formulation, are in chapter 5.

4.1 Sketch of the BMC Process

Let's start this initial sketch by imagining an agent that has a built-in perceptual system that takes inputs from the world and presents objects as having some color value between red, yellow, and blue, and some shape value between rectangular and round. In other words, we can say that the agent has a 2-dimensional perceptual space, with one dimension for color and one dimension for shape, and every perceived object falls somewhere in that perceptual space. See Figure 4.1.

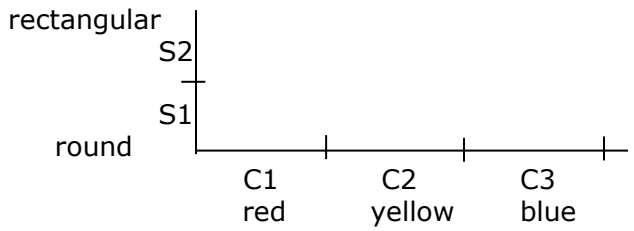


Figure 4.1: An agent's innate perceptual space

This agent can be said to have representations making up the two dimensions of perceptual space, corresponding to these colors and shapes.[8] A representation in this perceptual space can be said to be a *Simple Symbol* that is either an *Innate Symbol* or a *Brute-Causally Acquired Symbol*. They are also known as representational Primitives, as discussed near the beginning of Part III of the dissertation.

> ***Simple Symbol***: a symbol that is simple in representational structure, in the sense that it directly represents a simple meaning.
>
> ***Innate Symbol***: a symbol that has its meaning prior to any experiences.
>
> ***Brute-Causally Acquired Symbol***: a symbol that comes to have its meaning sometime after birth/creation, but by a non-inferential/non rational process, for example, by brain development or mere causal contact with entities in the world.

Imagine further the agent is designed to learn from its experiences how best to carve up its perceptible environment for its own purposes. The agent randomly generates syntactic strings to be used as representations. It might help to think of the agent as having two mental 'buckets' of arbitrary strings of syntax; in one bucket there are strings to be used as names for objects and in the other bucket there are strings to be used as names for kinds and properties. Suppose there is also a syntactic marker that stays with the symbol so that the agent can distinguish the object names from the property names (for example, object names are in lower-case and property names are in capitals).

---

8 No claims are being made presently about the actual perceptual space of human beings. The example serves merely to illustrate the *kind* of process that allows for the acquisition of simple concepts.

This machinery allows the agent's perceptual system to present to the agent objects as having properties in terms of positions of the objects on its perceptual space. This is a straight-forward sense in which the agent is able to compose its shape and color symbols together to entertain thoughts, like, OBJECT b HAS PROPERTIES C1 AND S1. Let each of the lower-case letters in Figure 4.2 be a randomly selected symbol that is assigned to the objects the agent has perceived.

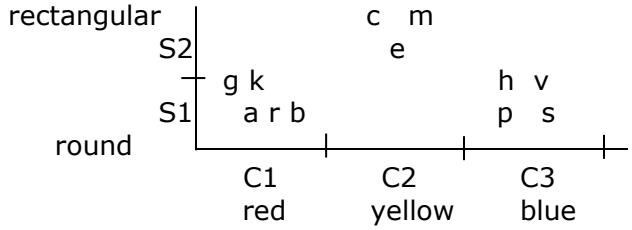


Figure 4.2: Objects perceived with color and shape properties

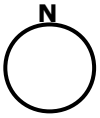
Now, imagine that the agent is able to detect patterns in its property space, as is done by commonplace clustering algorithms used in the Machine Learning branch of Artificial Intelligence (see Mitchell, 1997 for an overview). Clustering algorithms find sets of similar objects by measuring distances between them in a feature space. The set of perceptual representations that describe the cluster, as being around C1 and S1, can be composed into a new, complex representation. The new representation may be a conjunction that is defined by the perceptual representations, C1&S1, or alternatively the new representation may be a fuzzy, probabilistic, weighted conjunction of those representations. Unlike the perceptual representations, which are given innately, the agent plays a role in the acquisition of these complex representations. The complex representation is constructed as the result of a simple inference from objects have each

conjunctive property to their having the conjunction. Such a representation can be said to be a *Complex Symbol* that is also a *Rationally Acquired Symbol*.

> ***Complex Symbol***: a symbol that is complex in representational structure, in the sense that it has a complex meaning that is composed from the meanings of its representational parts.
>
> ***Rationally Acquired Symbol***: a symbol that comes to have its meaning through a rational/inferential process that is sensitive to the meanings of symbols used in the inference.

Now suppose that the agent is designed to go further, and treat some such clusters as an indication that the objects in the cluster have a similar *underlying* property that explains their perceptible similarities. The agent would be, in effect, designed to discover properties that are not directly perceptible. Suppose the agent is designed to then form a mental description to pick out the purported property that would explain the perceived similarity. Once the purported property is picked out by a mental description, the agent could then assign a new simple mental term, from to the property picked out by that mental description. This way, a new simple name comes to represent that newly discovered property, the property of being an *apple*, as the case might be.

This baptism is what initially determines the meaning of the newly acquired concept, and the information making up the description is then stored as contingent information about things that fall under the concept. Figure 4.3 shows property names in capital letters that are associated with each cluster.



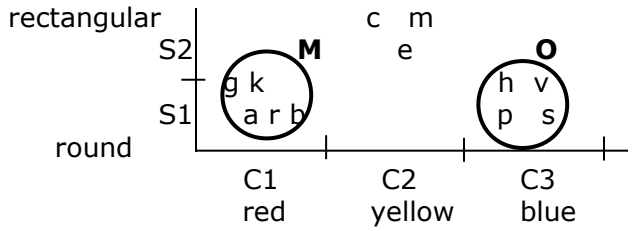


Figure 4.3: Objects perceived with color and shape properties

According to the BMC, many lexical concepts are acquired via the detection of such clusters of representations already in possession, but are simple in structure rather than being the composed conjunctive representation itself. This mental name can be said to be a *Simple Symbol*, as characterized for the perceptual representations, but is also a *Rationally Acquired Symbol* rather than being innate (or acquired by some brute-causal process –see Part III).

> ***Simple Symbol***: a symbol that is simple in representational structure, in the sense that it directly represents a simple meaning.
>
> ***Rationally Acquired Symbol***: a symbol that comes to have its meaning through a rational/inferential process that is sensitive to the meanings of symbols used in the inference.

Once the clusters are detected and names are given to the properties responsible for them, the agent can represent the objects used to discover the property *as* having the property. For example, the agent can entertain the representations ‘g is an M’ and ‘c is an N’.

This machinery also allows new objects to be recognized as having the newly discovered property. If a new object is perceived as falling near enough to the objects that are taken to be Ms, and far enough from the other objects (where enough is decided, however arbitrarily, by the clustering algorithm), then the new object will be judged by the agent to be an M.

The new property representations are acquired and the acquisition process is a rational/inferential one. These new property representations are also simple in their representational structure even though they are acquired from a process that is sensitive to the meanings of representations already in possession. That is, the property name is neither syntactically nor semantically identical with the description that picks it out. Notice first that 'M' is different in syntax from 'C1 & S1'. It is a simple name that does not have representations as parts. Notice that 'M' is being used to name a newly detected property, the *underlying property* that is the inferred explanation of *that observed pattern in the world*. M was not acquired by the agent's mere composition of its representations. Finally, this model allows the agent to think about blue apples and yellow apples, by simply composing M with other representations. This is because the meaning of the newly acquired concept is the kind that is contingently picked out by the mental description, instead of being the description itself.

So the proposal is as follows. An agent assigns an arbitrary mental symbol to a kind, when presented with a cluster of objects that belong to that kind. The mental symbol refers directly to the thereby introduced property. In the vernacular of the philosophy of language, the agent 'baptizes' the purported kind with the use of the mental symbol.

According to the Baptism Meanings for Concepts framework, most such concepts are neither Primitive nor Composite in the traditional sense (see chapter 6). Instead, they are the result of a process that is a mental version of the process in the Kripke/Putnam/Burge/Soames account of the baptism of kind-terms in language (see section 4.6). Mental terms for natural kinds are acquired by forming a mental description in terms of representations already in possession to pick out a natural-kind property, and then baptizing a new simple mental symbol to be used as a representation for the property that is picked out by the description. The mental description is formulated in terms of representations already in possession by the agent. Once a new property is picked out by the description, a new simple mental name is baptized as the mental name for the property. This resulting mental name becomes a representation for that property.

With more experiences of objects as falling on the feature space, the pattern of clusters may change. As more concepts are acquired, new dimensions for similarity may be used, making a larger similarity space in which to find clusters. With more and more dimensions, truly similar kinds will tend to form tighter clusters, and merely superficially similar ones will begin to spread apart. Under some such circumstances, the agent may be designed to re-baptize with terms for the kinds that explain its newly revised patterns of clusters.

The minimal sketch given here is just an example of a BMC process. Versions of the process may involve non-perceptual representations (for example, other concepts or representations of linguistic terms) and the process need not involve the detection of *clusters*. In discussing the details of the BMC framework in the next chapter, I highlight an important difference between concepts, which are representations for *kinds*, and

demonstrative perceptual representations, which are representations for object, property, and relation *particulars*. The later representations, the ones making up the *appearances* that are directly triggered by the world, can themselves be abstracted upon by the agent. I suspect that the acquisition of concepts for perceptible property *kinds*, such as RED, can be acquired from perceptible property *particulars*, such as reddish appearances, by using the latter in a reference-fixing description. The description might be something like 'the kind of property that tends to cause experiences like this (in me)'.

In terms of the elements involved in cases of representation, we can summarize the BMC framework as follows: lexical concepts are (1) simple in structure, with (2) simple meanings, (3) by virtue of being the unique property that is picked out by a reference-fixing description in terms of perceptual and motor demonstratives. The representation-making relation is (4) set up by the agent's noticing perceptual-motor similarities, inferring a property-kind as common cause, and naming that property-kind. (5) Concepts can't really be typed by syntax alone or by semantics alone, nor by a combination of the two. The carvings made by Folk Psychology cannot be the carvings made in a representationalist cognitive science.

4.2 Discovery of Diseases from Symptoms

The rest of this chapter further characterizes the BMC framework by showing how it intuitively explains many ordinary phenomena having to do with concepts.

Concept acquisition by the Baptism of Meanings for Concepts (BMC) framework is a lot like the discovery and naming of diseases on the basis of similarities in observed

symptoms. Moreover, the revisions in carvings and the concept re-baptisms that are proposed by the BMC are mirrored in the medical domain.

First consider Parkinson's disease, and the following passage from a website on the naming of diseases.

> Parkinson's disease was first described by British physician James Parkinson in 1817, and later amended by Jean-Martin Charcot 40 years later. It was only then that the affliction Parkinson originally described as "Shaky Palsy" became known as Parkinson's disease. [http://www.whonamedit.com/doctor.cfm/392.html]

The clustering of symptoms found by Parkinson became tighter as more data became available. This further confirmed that Parkinson had found a real *kind* of disease, rather than merely a superficial clustering of similar symptoms. With enough confirmation, with tight enough clusters, it became appropriate for the field of medicine to give a name to the kind of disease that Parkinson discovered. This next passage suggests further that the kind was named before its deeper underlying features were discovered.

> Although this neurological disorder has been long recognized in modern history, its biochemical origin was described almost 100 years later by Swedish Physician Arvid Carlsson, who would share the 2000 Nobel Prize in Physiology and Medicine for his discovery [http://nobelprize.org/nobel_prizes/medicine/laureates/2000/press.html].

Moreover still, the field of medicine explicitly distinguishes sets of symptoms from diseases. For the purposes of treatment, of course, it is extremely important in medicine to distinguish observable symptoms from the disease itself.

Consider the following passage.

> One must distinguish "parkinsonism" from Parkinson's disease. Parkinsonism is a syndrome (a complex of symptoms; in this context, a complex of various movement symptoms) that may be caused by Parkinson's disease, but which may also be caused by infectious, vascular, pharmacological, toxic, metabolic,

> structural, and various degenerative disorders. In other words, not every individual with parkinsonism has Parkinson's disease. [http://scienceweek.com/2003/sw030214.htm][9]

Other diseases follow similar patterns of discovery and naming. Here are some passages on Lyme disease from an article called "At the drop of a tick: a corps of Lyme-disease fighters meets its match in an army of arthropods"[10]:

There was a cluster of people in a similar location with similar symptoms.

> …It usually begins with a red dot on the skin, encircled by increasingly faint rings. Weeks to years later, the saga continues with episodes of chronic or acute arthritis, neurological problems ranging from a stiff neck to meningitis, and/or cardiac malfunctions.
>
> … but it was not until November 1975 that Yale rheumitologist Allen C. Steere launched the first survey for the disease and found an unusually high incidence (39 children and 12 adults) of what looked like juvenile rheumatoid arthritis in the towns of Lyme, Old Lyme and East Haddam, Conn. in 1976, steere and his colleagues named the disease.

At some point they noticed the cluster in observed properties and inferred that there was a single shared cause (or causal kind).

> In 1977, they published the first report on it in ARTHRITIS AND RHEUMATISM. The disease has plagued Europeans for nearly 100 years, but before the U.S. epidemic, no one had linked the seemingly unrelated array of symptoms to a single cause. Not until 1982 did scientists identify the bacterial perpetrator propelled by the bite of a tick.

They didn't know *what* the cause was, however, until later. That is, at first they didn't know anything *about* the cause, other than that there was one. Only later, after setting out

---

9 This article can also be found in ScienceWeek, February 14, 2003, Vol. 7 Number 7 (An Online Digest of Research in the Sciences).

10 http://www.thefreelibrary.com/At+the+drop+of+a+tick:+a+corps+of+Lyme-disease+fighters+meets+its...-a07502165

to discover more features of the cause, did the researchers find that it was a bacteria spread through a tick bite.

Now, it might be tempting to object to this BMC interpretation of the discovery of diseases by saying that once the deeper features of a disorder are discovered, a description that heavily weighs the deeper features becomes the *new* complex concept for the disorder; the name is just shorthand for those deeper features. The problem with this objection is that it eliminates the notion of there being something that is *the* concept for the condition, it makes it impossible for the doctors to have discovered new features of a given disorder, and it makes it impossible for a discovery of a cure to be a discovery of a cure *for that disease*.

Another aspect of the BMC framework that is mirrored in medicine is the revision of carvings of superficially diseases. Consider the distinction between the cold and influenza (the flu), two diseases that are notoriously similar in initially observable symptoms, but which have importantly different underlying properties and treatments. The field of medicine seeks out the underlying *mechanism* that explains the observed symptoms, as stated in this medical article passage.

> The common cold and influenza (flu) are the most common syndromes of infection in human beings. … New knowledge of the effects of cytokines in human beings now helps to explain some of the symptoms of colds and flu that were previously in the realm of folklore rather than medicine—eg, fever, anorexia, malaise, chilliness, headache, and muscle aches and pains. The mechanisms of symptoms of sore throat, rhinorrhoea, sneezing, nasal congestion, cough, watery eyes, and sinus pain...[11]

11 Ron Eccles (2005)

The flu and the cold have similar superficially observable symptoms, in spite of being different diseases. What makes them different diseases is that they have different underlying causes.

4.3 Inferences about the World are Both Rational and Informative

It seems intuitive that human beings make rational inferences about the world on the basis of perception. More specifically, it seems like there is some kind of thought processing that we as agents do that leads us to conceptual beliefs as the result of perception. The belief that there is a *cup* on a *table* seems to come by reasoning from the way that things perceptually appear. The belief doesn't seem to come as a result of direct causation from the cupness and tableness of the entities we encounter.

At the same time, the inferences are rational in a way that not merely a matter of trivial deductive logic. Reasoning with deductive logic is rational only because the information in the premises literally contains the information in the conclusions. The beliefs we infer about the world on the basis of perception are far from trivial. There isn't a *logical* link between the way an object appears and its belonging to one category or another. Still, there is some kind of inferential link, one that seems to be to some extent within our control, and it seems we can learn to be better at making such inferences (see Appendix B for more on this).

Moreover, we seem to learn from perception not only what kinds of objects are in the world, and patterns between them. We also seem to learn *the kinds themselves*, concepts for the kinds, that is. Perhaps 'learn' isn't the right word. But there is a sense in which we seem to form and try to confirm hypotheses about kinds that we couldn't represent before perception.

These intuitions are not enough on their own to guarantee that there is a combination of rationality and informativeness in our conceptual inferences from non-conceptual[12] perceptual experience. It will become clear in Part III of the dissertation that other views have to give up one or the other. The Baptizing Meanings for Concepts framework very naturally makes sense of the compatibility of the pair of intuitions. The details of how the BMC framework does this are in the next chapter (chapter 5) in the section on contingent a priori inferences.

4.4 Naïve Essentialism/Psychological Essentialism

Young children (as young as 12-months) seem to assume that objects have essences, necessary properties underlying and explaining their observable properties. During classification, deeper similarities and differences seem to take priority over superficial ones. As psychologists Susan Gelman and Ellen Markman (1986) put it:

> Natural kinds are categories of objects and substances that are found in nature (e.g., tiger, water, cactus)... natural kind terms capture regularities in nature that go beyond intuitive similarity... Natural kinds have a deep, nonobvious basis; perceptual features, though useful for identifying members of a category, do not always serve to define the category. For example, "fool's gold" looks just like gold to most people, yet we accept the statement of an expert that it is not gold... Because natural kinds capture theory-based properties rather than superficial features, some of the properties that were originally used to pick out category members can be violated, but we still agree the object is a member of the kind if there is reason to believe that "deeper," more explanatory properties still hold. (p. 1532)

12 Recall my notion of ‘concept’ as distinguished in Part I.

In one of the famous experiments from Gelman and Markman (1986), children were shown pictures of various animals, including a rhinoceros, a triceratops, and a brontosaurus. They were told that the rhinoceros has warm blood and the brontosaurus has cold blood, and they had to guess what kind of blood the triceratops has. With only information about the appearance of the animals, the children grouped the triceratops with the rhinoceros, and inferred that the triceratops has warm blood. However, when children were also told that the triceratops and brontosaurus are both dinosaurs, the children inferred that the triceratops has cold blood.

It has further been shown that people tend to classify things by their internal properties, rather than perceptible ones (for example, a skunk given the body of a squirrel) by Susan Carey and Frank Keil (1986).[13] And, Medin and Ortony (1989) suggest that people may use an 'essence placeholder' when they think there is some quality, some essence that is shared by, for example, bears, even if they are unable to identify any feature or trait as that essence. It has also been shown that young people tend to make inferences about the future states of kinds based on their internal similarities, for example, that an apple seed will become an apple tree (Gelman, S. A., & Wellman, H. M., 1991).

In a similar spirit, it has been shown that infants (6-12-month olds) posit goals in an object in order to understand and predict observed patterns in an object's movements. The representation of an unobservable goal is revealed by the infants' rationality-based predictions of future movements.

---

13 See also Frank Keil and Rips, L. J. (1989)

For example, in experiment by György Gergely and his collaborators (1995), infants were habituated either to a circle going over a wall to hit another circle (experimental group) or to a ball going along that same trajectory even though the two circles were on the same side of the wall (control group). During testing, the barriers were removed and the infants were shown (a) one circle making a straight line to hit the other circle and (b) a circle taking the trajectory used in the habituation phase (as if going over a wall). The barriers were removed for the test conditions and both groups were shown two conditions: one in which a ball moved in a straight line to contact the other ball (direct), and another in which the ball jumped along the same trajectory as in the habituation phase (indirect). The experimental group looked longer (indicating confusion) at the indirect cases than the direct cases, while the control group looked for a short while at both cases.

4.5 We Revise our Carvings

When we classify things by their perceptible similarities and differences, we do so defeasibly. We revise our carvings of the world when we find them to be inconsistent with deeper or more important dimensions of comparison. With the concepts JADE and ARYAN, for example, the superficial similarities (similar greenish appearance, or blue-eye/blond-hair combinations) in the samples are overridden by deeper, more seemingly important, properties in the samples. We may continue to use the categories for convenience, but we remove them from our general understanding of how the world is carved.

The defeasibility of our conceptual schemes, revising them as we discover patterns in deeper properties, is further supported by classical studies in cognitive psychology. Consider the revision that is made by children in the experiment discussed the previous sub-section, 4.4, on Naïve Essentialism. These revisions are discussed further in section 11.4.1, on the rationality of concept acquisition and in section 5.4, on existence and uniqueness conditions on the ability of a description to refer.

4.6 How Linguistic Terms Get Their Meanings

Much of the focus of 20th Century analytic philosophy was centered on the following issue. By virtue of what are natural-language terms meaningful? What makes the English word 'apple', the name 'Plato', and the sentence 'Plato ate an apple' mean what they mean? How do things of this sort come to be *about* other things?

Philosophers have appealed to a baptism process for the introduction of natural kind terms. For example, Hilary Putnam, when discussing how one can learn the meaning of the linguistic term 'water', seems to suggest a mental process very much like the Baptism Meanings for Concepts framework via reference fixing via similarities in a perceived sample.

> Suppose I point to a glass of water and say, 'this liquid is called water'… My 'ostensive definition' of water has the following empirical presupposition: that the body of liquid I am pointing to bears a certain sameness relation (say, *x is the same liquid as y*, or *x is the same$_L$ as y*) to most of the stuff I and other speakers in my linguistic community have on other occasions called 'water' [pg. 225 in *Mind, Language and Reality*].

Putnam's 'ostensive definition' here seems to require the same kind of mental work as is being proposed in the BMC. In picking out the kind, water, there is a presupposition that there is exactly one underlying kind that is common to a set of similar-looking samples.[14]

Similarly, Scott Soames (2001, pp. 266-7), when extending Kripke's account for the acquisition of linguistic natural kind terms writes:

> According to [Kripke's] account, the predicate is first associated by speakers with a kind –either ostensively of via a description. In the ostensive case speakers directly associate the predicate with a certain sample of individuals, which they presume to be instances of a single natural kind of a given type (e.g., a single substance or a single species). In the descriptive case, speakers employ a description that picks out a unique kind, often by appeal to contingent properties of the kind, or its instances.

Soames draws a somewhat artificial distinction between what he calls 'reference-fixing by ostension' and 'reference-fixing by description'. I take both processes to be species of Composite reference-fixing, albeit ones that differ in that ostensive reference-fixing involves descriptions that contain perceptual demonstrative reference to some sample of individuals. Soames' description of natural-kind term introduction by ostensive reference-fixing is analogous to the process of mental baptism of properties of the sort I have been discussing.

The consensus in the case of language is that complex terms, like sentences and phrases, come to be about other things by their combinations (in accordance with some grammar rules) of more basic meaningful parts. In language, it is clear that words (or rather, morphemes) are the most basic units of meaning. These include proper names

14 Notice that Putnam is careful to include 'liquid' in the presupposition, to ensure that the term 'water' gets hooked onto the property of being water rather than any of the other properties in the sample.

(like 'Plato', 'Scruffy', and 'Arizona') as well as single words for kinds of properties and events (like 'water', 'snail', 'ladder', and 'eat'). As for these most basic linguistic terms, the growing consensus is that they have their meanings because *we*, human agents, have assigned to them those meanings. Linguistic terms are meaningful (i.e., have intentionality/aboutness) in a way that is *derivative* on the meaningfulness of thoughts.

Many philosophers and cognitive scientists have come to adopt the Representational Theory of Mind (RTM) or Language of Thought (LOT), on which thoughts, full propositional ideas, are structured representations, just like complex linguistic terms (see Fodor 1975). The most basic meaningful parts of thought are composed together (in accordance with some 'mental grammar') into more and more complex units of meaning. Not everyone in the field accepts the RTM, but it is fair to say that almost all concept theorists do.[15] Indeed, it is natural even to interpret many of the 17th Century concept theorists (Locke, Hume, Descartes) as endorsing something like the RTM.

The growing consensus in the philosophy of language is that linguistic terms get their meanings by a kind of dubbing that is a lot like a baptism ceremony (Kripke, 1972; Putnam, 1975; Burge, 1979; Soames, 2002). For example, consider the following expression.

(1) I dub ‘apple’ as the name for apple.

---

15 Opponents to the RTM include Dennett (1977), Churchland (1981), and Matthews (2007). The alternative views take concepts to be *abilities* (Brandom 1994, Dummett 1993, Millikan 2000, and Peacocke 1992) or as *abstract objects* (Frege, 1892).

This dubbing statement is a way, according to these philosophers, for a linguistic term to be assigned a meaning. Of course, the dubbing requires that the agent is able to represent apple before having the word 'apple'. Usually the representation of apple that allows for the dubbing is thought to be achieved by pointing at samples, as in (2).

(2) I dub 'apple' as the name for *that* property.

Philosophers of language quickly notice, however, that merely pointing at some samples is not enough to do the work. Devitt and Sterelny (1999) call this the Qua-problem. There are *many* kinds instantiated in any sample of apples, so it is underdetermined whether the expression picks out 'that property' qua *apple*, versus qua *fruit*, or qua *organic*, qua *edible*, qua *apple peel*, or qua the disjunction *apple or wrench*.

To deal with the Qua-problem, philosophers of language usually assume that the agents involved in the baptism ceremony share a salient concept, a mental term, with which to pick out the property *apple* from among the other properties shared by the instances. Some psychologists (Rosch, 1978; Gleitman et al. 2005; Bloom, 2000) suggest further that children rely on such concept salience during word-learning, to determine the referent of a novel word. Using APPLE as a mental term that means apple, the agent can entertain (3).

(3) I dub 'apple' as the name for APPLE.

This dubbing assigns a linguistic term to the meaning of the mental term APPLE. Philosophers of language usually stop at (3), leaving it to the psychologists and philosophers of mind to worry about how the mental term APPLE comes to mean apple.

The BMC further appeals to a baptism process for assigning meanings to the mental terms. At first, this suggestion seems to give rise to a regress. That is, the agent would have to think a thought like (4).

(4) I dub M as the mental name for apple.

In order to entertain (4), the agent would have to be able to think about the property of being an apple before even having a *mental* term for apple.

We already saw that we ca not use mental demonstratives to point at the property, not on their own at least, because of the Qua-problem. But now it seems to suggest a circularity. We ca not solve the mental qua-problem of assigning meaning to the mental term for apple *with* a mental term for apple.

Once again, there seems to be only two options if the acquisition of concepts is to be an inferential process:

(a) find a conjunction of simple properties that necessarily applies to all and only the apples, so that APPLE can be assigned to the conjunction.

(b) find a way for a concept that directly refers to the kind apple to be acquired not by brute-causation, but by an inferential process.

Again, since (a) has been such a struggle, I pursue option (b). The description (D) and the baptism (B) in the next section, 4.8, serve to complete the linguistic baptism.

4.7 Concept Acquisition for Robots

The Baptizing Meanings for Concepts framework comes naturally when considering the philosophical puzzles about concepts from the perspective of Artificial Intelligence. Moreover, doing so makes many of the questions more precise. In asking how we come to have concepts, we can ask what it would take for *any* physical entity to have concepts. Machines, computational ones, are the best candidates we have.

One of the biggest philosophical objections to Artificial Intelligences is that the agents' symbols are meaningful only to us human users and engineers (Searle, 1980). The *agent* does not understand them. According to the model presently on offer, The Baptizing Meanings for Concepts framework (BMC), one thing that is missing is the agent itself making the symbol mean what it means.

Most of us believe that computers, even our best AI systems, lack concepts. What are they missing? Steven Harnad has pointed out that the agent's symbols need to be *grounded* in the world in order to genuinely have meaning (Harnad, 1990), to play a causal role mediating perception and behavior. Notice that not just *any* causal mediation will do; thermostats have sensors and behavioral motors, yet they lack mentality. What kinds of connections to their meanings make mental representations have genuine aboutness? Briefly consider the following three agents.

COREF (for COllaborative REFerence) is a natural-language dialog agent that works with its user to descriptively pick out objects in a shared visual scene (DeVault, Oved, and Stone, 2006). With Computer Scientists and Linguists at Rutgers University, I worked on a dialogue agent, COREF (for COlaborative REFerence, along the lines of Clark and Wilkes-Gibbs, 1986). COREF plays a matching game with a human user, either as the Director or the Matcher. The Director is to select a target object from a

shared visual scene, and then both players are to use natural language through a text box, to collaboratively find descriptions for the objects so that the Matcher identifies the target. Early versions of COREF have been evaluated on speed of matcher-identification in comparison with human-human performance with the same visual scenes and interfaces. COREF has an infrastructure that allows the exploration of various parameters in performing this task, including lexical entries, ability to clarify and request clarification, the ability to learn new words, and concept-acquisition algorithms. The current incarnation of COREF has words like 'square' connected, by the hand of the programmer, to a symbol SQUARE, which is connected, also by the programmer, to procedures for the agent's recognition of the property of being a square. When the conditions are as expected by the programmer, COREF is good at recognizing squares. Does COREF have a concept for square? Of course, COREF's symbol SQUARE means square to us human programmers and users, but does it mean square *to COREF*? Is it a genuine concept? If it does not have a concept, what is missing? What would it take for COREF to have a concept that means square, for it actually understand what 'square' means? A common suggestion is that the agent needs to *learn* the concept as a result of interacting with the world, either through language or through perception, or both. Perhaps the problem with COREF's symbol for square is that it is connected by the programmer to its meaning, rather than by the agent itself. Turning to learning, it is left mysterious what it would be for an AI system, and thus for *any* physical system, to have an innate concept.

Let us suppose that acquisition by learning from experience is at least one way to come to have genuine concepts. Again, the only clear way to do this is by noticing

patterns in the world in terms of representations already in possession. Consider Stanley, the winner of the DARPA Grand Challenge (2006). This robotic car was designed to learn how to distinguish between terrains as being 'drivable', 'occupied', or 'unknown'. Although the machine learned the appearances that were useful for the categorization of the three terrain types, it did not acquire the symbols themselves. Again, the *meaning* of, for example, 'drivable' is not the same as the properties involved in an agent's classification of things as falling under the label. What was learned is contingent (i.e., non-essential) information *about* drivability, not what drivability *is*. Stanley is not able to entertain the idea of a drivable surface that lacks those categorization properties.

Towards an agent that acquires concepts, a third agent to consider is the following robot built by Bethany Leffler (2007). This agent uses a classifier to partition its environments based on the appearances of various terrains which it will traverse. The sagent detected two different classes, one corresponding to a cloth terrain and one corresponding to wooden surfaces. After finding the classes, the agent learns how to move on each terrain type. Does the agent genuinely have a concept for wood? Certainly it comes to be sensitive to two classes of terrain based on perceptual experiences, and uses those learned classes to guide its behavior. This kind of classifying agent is the most compelling candidate for a concept-acquisition agent. Still, the agent cannot so clearly be said to have acquired the concepts inferentially. If the acquisition is merely brute-causal, it is not so clear that the agent genuinely has a concept that means wooden *to the agent*.

How do concepts, these mental entities, come to be meaningful? Let us consider the analogous question of how linguistic entities come to be meaningful. This question

has been explored in depth in the philosophy of language. The Baptizing Meanings for Concepts framework is not only an analog of the baptism of linguistic terms, because concepts have to already be in possession for the linguistic case to be carried out.

Let us imagine building a concept acquisition agent that uses perception and inference in the process. Suppose we try to build an agent, call it Pinoch-i/o, with the design goal of carving its world into categories that are useful for the agent. Suppose we also give Pinoch-i/o some 'innate' sensory detectors as well as some innate composition and inference mechanisms. Can Pinoch-i/o acquire a concept for apple through its innate detectors and reasoning mechanisms?

Let us suppose we give him three visual detectors: Color, Shape, Size. Let us also build in a 3-dimensional perceptual space so that every object he perceives is presented to him as a point in the space, as being a particular color, shape, and size.

We can add an innate clustering algorithm. These are commonplace in the unsupervised-learning and semi-supervised-learning branches of Artificial Intelligence (for an overview, see Mitchell, 1997). These algorithms measure similarities in a set of data points and group the points together. These algorithms use variations on the Euclidean distance formula familiar from high-school geometry. Variations on the formula allow for weighted properties and the consolidation of covariant properties.

After Pinoch-i/o is released into the world, he may find that some of his perceived objects form clusters. Suppose Pinoch-i/o finds a cluster of objects around the red, round, and 3Inch-diameter area of his feature space. Suppose that in fact those objects are apples.

We could build Pinoch-i/o to compose (by conjunction in this case) the properties that make up the cluster. We could also give him an inference rule to conclude that each of these objects has the complex feature of being reddish and roundish and 3-inches. This would amount to approach (a) for Fodor's Challenge. But that would commit us to saying that the meaning of APPLE is the same as the meaning of [REDDISH & ROUNDISH & 3INCHISH]. Again, these representations do not have the same meaning; they pick out different sets of objects in different worlds. Appleness, like water, seems to have an essence that cannot be known a priori. Approach (a) has a long history of problems with analyticity. Approach (b) has not been explored much. Moreover, naïve essentialism suggests that lexical concepts might directly denote essences, rather than their symptomatic properties.

My proposal is that we build Pinoch-i/o to infer from a perceptible similarity to the presence of a deeper, essential, natural property that explains the superficial similarity.

Suppose that Pinoch-i/o uses a reference-fixing description with demonstratives to pick out that natural kind. Again, demonstratives alone will not work. But the following description (D) would denote appleness relative to this context of use.

(D) the natural-kind property that these objects share that explains this perceptible similarity

Suppose this description does in fact pick out apple in Pinoch-i/o's context.

Now suppose we build him to assign a simple name to that property, to baptize it with a name, so that he can use the simple name for further cognition. We can think of the baptism process as implicitly involving a thought we can characterize with (B).

(B) I dub M as a mental name for the natural-kind property that these objects have that explains this perceptible similarity.

I claim that M would be a mental term that means apple.

For later inferential processes involving the concept and perception, like recognition and prediction, we can give Pinoch-i/o a way of storing the perceptible properties involved in the acquisition of M. The properties can be stored as contingent information about apples.

After the fixation of its reference, M is a logically proper name for apple. M was acquired by a process that relies on perception and inference. Pinoch-i/o inferred from a set of objects sharing certain properties to those objects sharing a property that explains their perceptible similarity. The description denotes a property. Once the property is given a name, the agent can think about the property in abstraction from the properties he used to pick it out. Pinoch-i/o can entertain the idea of *a blue one of those*.

4.8 The BMC Easily Resolves Tensions between Other Views

Much of the concepts controversy results from what I call *the Building-Blocks assumption*. This is the assumption that there are only two classes of mental symbols – Composite Symbols, which are Rationally Acquired by composing more primitive representations together into a Complex Symbol, and Primitive Symbols which are Innate

(or Brute-Causally, non-inferentially, Acquired) and Simple in representational structure. However, most of the arguments that treat lexical concepts as Composite Symbols rest on their being Rationally Acquired, while most of the arguments for their being Primitive Symbols rest on their being Simple.

The BMC is an account of concept acquisition that meets the challenge, offering a way for concepts to be Rationally Acquired even though they are Simple Symbols. Resolving the tension between the simplicity intuition and the acquisition intuition is what I am here calling *Fodor's Challenge.* These tensions are detailed in Part III of the dissertation.

5 Details of the BMC

5.1 Non-Conceptual Perceptual and Motor Demonstratives

What we are seeking, ultimately, is an account of what makes something a mental representation, so that we can determine whether and how it is possible to design an artificial thinking machine. One observation is that mental representations, unlike non-mental ones, seem to be *outwardly directed*, to have *aboutness*, or *intentionality*. As, we just saw, however, the mark of intentionality is *not*, as many have thought, that intentional representations can misrepresent and can have non-existent objects of their aboutness, for non-intentional representations, such as states of mechanical gauges, have these properties. What seems to be the mark of intentionality, rather, is that the meaningfulness of the representation is basic, in the sense that it is not reducible to the meaningfulness of anything else. Now we set out to discover what makes something meaningful in this basic sense. The present section offers and defends one suggestion to be explored further through an implementation of this kind of system. The suggestion is that the intentionality-making feature that mental representations have, and that non-mental representations lack, is *demonstrative identification* through their connection with *perceptual* (and perhaps in some cases, *motor* or *proprioceptive*) representations. In other words, mental representations are meaningful in a basic, non-derivative way by virtue of their involvement with representations that are meaningful in a *directly*, by perceptual demonstrative 'pointing', that does not require any further interpretation.

In the next few sub-sections we first consider some observations about visual representations and the acquisition of non-perceptual representations from vision, guided

by research in vision science. Then we state more precisely the potential account for the meaningfulness of perceptual and perceptually acquired (conceptual) representations that gives perceptual demonstratives a crucial role. Then we gesture towards the possibility of generalizing to a unified account of mental representations in general, extrapolating from perceptual representations and perceptually acquired representations to motor or proprioceptive representations as well as to innate and framework representations. The view at this early pass will have problems and be incomplete, but, again, implementation will be the method for revealing such problems.

For notation, we will use * after a term to indicate that it is a perceptual representation (for example, R* might be used to talk about the visual representation, or *look*, of typical red things for a particular agent); we will use all caps (as in RED) for concepts; and we will use > < around terms to indicate the *triggering* or *activation* of these representations (as in >RED< and >R*<).

5.1.1 Two vision-concept inferential links

In this subsection, we consider features of perceptual representations and concepts that (plausibly) are acquired through perception. We also consider a plausible story, based on earlier work (Pollock and Oved, 2005) on how the former cause the possession as well as the online triggering of the latter. From these links between perception and concepts we will see in the next subsection how their meaningfulness can be understood as being of the basic sense that is characteristic of intentionality.

The first thing to notice is that the concept RED is not logically or *by definition* linked to a perceptual representation or *look*. To isolate the difference between perceptual

color representations and conceptual ones, it may help to think about the logical possibility of so-called 'qualia inverts'—two people that have opposite experiences of the same pair of properties in the world. For example, there seems be a logical possibility that there are two people, Amy and Emma, such that Amy experiences the color of ripe tomatoes and spinach the way you experience them (call the experiences R* and G*, respectively), but Emma experiences tomatoes the way you experience spinach (G*) and spinach the way you experience tomatoes (R*). Now, both Amy and Emma when speaking English would *call* the color of tomatoes 'red' and the color of spinach 'green', as they've been taught. And, intuitively, neither Amy nor Emma would be *mistaken* when they think of the tomatoes as satisfying the concept RED and the spinach as satisfying the concept GREEN. They both refer correctly to *the* property in the world that is causing their respective experiences. It is the tomato's property of being red (that is, reflecting light at wavelengths approximately 700nm, among other reflectance factors) that causes both Amy's R* experience and Emma's G* experience, and it is the spinach's property of being green (reflecting light at wavelengths approximately 510nm, among other reflectance factors) that causes Amy's G* experience and Emma's R* experience. Now, in an important sense Amy and Emma both represent the same properties; they both represent *red* as the result of looking at a tomato and *green* as the result of looking at spinach. Amy and Emma are the same in their conceptual representations of the properties but different in their perceptual representations.

What this observation suggests is that most universal properties (even *red*) are plausibly not represented by symbols in Visionese, the language of vision.[16] It is clear

16 I do not claim that vision has a different *format* of representation from conceptual representations, other than their simply being a different set of symbols, such that cognitive processing is sensitive to the difference.

that there cannot be a logical link between a visual representation (i.e., a way of looking, or a *range* of ways of looking) and the universal property *red*. Thought experiments like the one involving qualia-inverts, like Amy and Emma, have been used to suggest this, but there has been some dispute as to whether inverted spectra are really possible. On the other hand, there are some very real phenomena that can serve to make the same point. One such phenomenon is known as *Photoxic lens Brunescence* (Lindsay and Brown, 2002), the natural yellowing of the lens with age, which results in the gradual shifting in the visual appearance of things. The lens, in effect, gradually becomes more and more of a yellow filter. Despite the resulting shift in the way things look, there is no shift in what we *recognize* as red, since the shift is gradual enough for us to adapt our mapping from looks to the concept. This suggests further that our representation of the universal property *red* is not in terms of a symbol in Visionese. Another, less exotic phenomenon that suggests the same thing is that patches of color look different against different color backgrounds (this is known as *simultaneous color contrast*). A nice illustration of this phenomenon can be found in Donald Hoffman's book (1998 [pg.112]), in which two sets of various identical color patches are side by side, but in different arrangements, making it extremely difficult to correctly match the patches across the sets. Note that such cases suggest that there is no sense to be made of the idea of there being *normal* or *ideal* viewing conditions, such that there is a Visionese symbol that represents the universal *red* under those conditions. This is because it is not at all clear how to determine which of our qualia-inverts *has it right*, and it is likewise not clear what would count as the ideal background for viewing a color or the ideal age, or ideal yellowness of the lens.

The next feature to notice about visual representations is that they are structured. Following the framework developed by David Marr (1982), the visual image (the output of the visual system) is understood to be produced in a series of stages, going from the many pairs of 2-dimensional retinal inputs, sensitive to light at various wavelengths, to the representation of a 3-dimensional world of objects with properties and relations. Really roughly, through the stages of processing, there are computations that detect blobs of color in the retinal inputs, and then lines and their concavities, and then edges. Eventually the system interprets the features of the retinal inputs as objects and their parts, properties, and relations. The output of vision, on this framework, is understood to be a structured array consisting of object representations linked to various property and relation representations. The visual representation can hence be thought to be very much like a representation of a proposition, or a bunch of propositions. These visual proposition representations have a syntax. There are object representations that take up the subject positions, and syntactically distinguishable property and relation representations that are linked up to the object representations, taking up the predicate positions.

If these general features of Marr's framework are right, we have some understanding of the representations that make up the visual representation. Now we can consider how concepts, i.e., representations of *kinds* or *categories*, might be acquired through such perceptual representations, and also how concepts might be caused to be *triggered* or *activated* by such perceptual representations.

Next let us consider the contents of the visual representations a bit further, and also their content-determinations. We will see that there is strong evidence that visual

object representations represent their contents in a *direct* way, by demonstrative pointing. In the concluding remarks of the paper, we will consider some avenues for extrapolating to an account on which visual property and relation representations also represent their contents directly.

5.1.2 Visual Object Representations

Let us start with visual object representations, and later we will consider property and relation representations. The most important feature of object representations for our purposes is whether they represent objects in a way that can be understood as basic, and not subject to further interpretation for their meaningfulness.

Vision researcher Zenon Pylyshyn (2000) gives empirical arguments that visual representations contain object representations that represent objects directly, by demonstrative pointing. He describes visual object representations, what he calls FINSTs (FINgers of INSTantiation), as attended visual object representations that are caused to be instantiated by properties in the world, but that are not themselves composed of representations of those properties. He also describes them as a sort of visual demonstrative or pointer, that is much like demonstratives in language (‘this’, ‘that’, and perhaps, ‘the other’). Such representations, he argues, represent objects directly, in virtue of a relationship between the object in the world and the agent. For FINSTs, the causal relationship from the world to the instantiation of the FINST is fully responsible for the FINST’s representing the object. Pylyshyn describes the FINST as, in effect, *attaching* to the object that is its referent, in the way that might happen if the eye had rigid “fingers” (or strings or stretchy plastic limbs) connected to the objects.

In defending this view about the structure and reference of these object representations, Pylyshyn cites data from vision science, some of which come from his own research on multiple object tracking (MOT). In a typical MOT experiment (Pylyshyn, 2000), 8 dots appear on a screen, 4 of them flicker, and then they all move around the screen for about 10 seconds. When they stop, the subjects' task is to identify the ones that had flickered.

Data from such experiments show that subjects consistently track the 4 objects (87% accurate). Pylyshyn argues that the best explanation for the MOT data is that we represent the objects with FINSTs, representations that stay fixed to the (proto-)objects in the way that fingers would. He argues that it cannot be that we track them by encoding their locations and repeatedly updating information about their locations –that would take up more attention and scanning than the subjects have time for in the experiments. In other versions of the tracking experiment, it is found that the dots can't be tracked when they are connected to non-target dots with a bar (the visual system seems to treat the barbells as objects). Also, it is found that the dots can be tracked through (proto)-occluders and through changes in color and shape, suggesting further that the object representations tend to track the dots under the conditions that would be present if and only if (ceteris paribus) the dots are objects. Further data on MOT show that these object representations are indeed attended, or salient, representations. For example, it takes less time to find a property among the tracked objects than among the non-tracked ones.

Some further support for FINSTs as representing directly might be found in experiments on infant cognition. It has been observed (Káldy and Leslie, 2005) that infants can think about objects independently of their properties, perhaps suggesting that

there are visual object representations that directly "lock on" to objects. In his experiments, infants were habituated to a scene in which there was a screen, out from which came a red ball and then a blue ball, repeatedly in alternation. Once habituated, the infants were shown either (a) a red ball and a blue ball side-by-side, out from behind the screen at once, or (b) two red balls, out from the screen at once, or (c) a single red ball. The children in condition (c) were surprised (as though they had expected two objects, not just one), but the infants in neither conditions (a) nor (b) were surprised (as if their expectation for two objects, was met). The striking thing about this set of data is that the infants seemed to be sensitive to the difference in color in *coming to have* the expectation that there are two objects, but not in *recognizing* the objects *in terms of* their colors. Demonstrative object representations in vision might be able to explain these data. Consider the following illustration of the causal relationship between the world and object representations a and b:

(i) While the conditions in the world are [*red*(x) & *round*(x) & *small*(x)], then object representation a is on.

(ii) While the conditions in the world are [*blue*(y) & *round*(y) & *small*(y)], then object representation b is on.

The object representations, a and b, are/remain triggered while the causal conditions in the world hold. But those conditions do not get *represented* with the object representations. As the infant repeatedly sees a red ball and a blue ball, she entertains the

representations a,b,a,b,a,b,a,b,a,b,a,b,…, perhaps without representing the properties of the objects within her visual representation. In other words, perhaps the properties of the objects are used in early stages of visual processing in order to establish object representations, but the child herself has no access whatsoever to what is causing the object representations to fire since the property representations are not present in her visual representation. This allows her to expect there to be two objects without knowing anything about them.

The point of going through these data is to show that visual object representations seem to represent objects directly, in a way that does not involve any interpretation in terms of other representations. Let us now consider property and relation representations in vision. We will see that it is reasonable to think of these representations as similarly representing their contents directly by demonstration.

5.1.3 Visual Property and Relation Representations

Currently, we attempt to liken property and relation representations in the visual representation to visual object representations regarding the directness of their meaningfulness. Recall that what is meant by 'visual property representation' here is something like the qualia of our inverts, Amy's R* and Emma's G*, the raw looks of an object. Of course, it should be recalled that the observations above suggest there is really a *range* of looks that each person has corresponding to the concept RED. The representations are some kind of color value, some symbol for a particular shade. Examples of relation representations in vision might be restricted to spatial relations, like

IN_FRONT_OF*, and ABOVE*, but may also include other relations, like BRIGHTER* and BIGGER*.

Along these lines, a ‘proposition representation’ in Visionese might have the form ‘R*a’ or ‘aABOVE*b’. And, now the question is, in virtue of what do R* and ABOVE* represent their contents? In the case of object representations, it was easier to rule them as direct, for we could focus on their not being interpreted in terms of property or relation representations. But here we are taking it for granted that R* is a property representation and ABOVE* is a relation representation, and still, we want to argue that these and other visual property and relation representations within the visual representation represent their contents directly.

Now, we’ve seen repeatedly that R* is not the same as RED in that they have different contents. RED is supposed to be a concept that is a representation of the universal property *red*. So what is the content of R*? One suggestion to consider presently is that although visual representations plausibly include such representations of properties and relations, they are not necessarily representations of *kinds* or *categories*. The present suggestion is that they do not predicate of their subjects that they are members of any category; instead, they predicate of their subjects that they have some *particular* property, and they do so demonstratively. In other words, the visual representation is entirely demonstrative. A representation of the form ‘R*a’ can be glossed as ‘that is that way’, where the contents of the indexicals are to be filled in by what they point to. So long as the property and relation representations in vision are representations of particulars, they are not concepts in the traditional sense.

5.2 Contingent A Priori Inferences

Perceptual representations convey conceptual information about the world for further cognition. There is no *logical* connection, however, between the way things appear and what kind they fall under (sometimes we misjudge). The inference has to be both warranted and informative. An inference that is a priori is one that is warranted *internally* by an agent's understanding of the beliefs involved, without needing to consult the perceptible world. A contingent inference is one that involves connections between ideas that are non-necessary connections.

What is needed for perception to convey interesting information about the world is a kind of contingent a priori inference. The Baptizing Meanings for Concepts framework (BMC) model yields exactly this. There are two kinds of a priori contingency that result from The BMC framework. First consider a rather boring kind.

Suppose an agent introduces a mental term, M, with a reference-fixing description that picks out apple. The agent is then automatically allowed the a priori inference to (7).

(7) If anything is an M, then some Ms are reddish, and roundish, and approximately 3Inches.

The inference to (7) is a priori at the baptism stage. Yet (7) is contingent –it could have been the case that some things are apples but no apples are reddish and roundish and approximately 3Inches.

The inference to (7) is similar to Kripke's (1972) Neptune case. He suggests that since the name 'Neptune' was assigned through the description 'the planet that causes

such and such perturbations', Kripke claims that the inference to (8) is a priori, but clearly not necessary [fn33,pg. 79].

(8) If such and such perturbations are caused by a unique planet, they are caused by Neptune.

Kripke's case is controversial. The worry is that the introduction of the name 'Neptune' yields (8) in a way that seems like 'easy knowledge'. The BMC framework has a slight advantage here because the agent has perceptual acquaintance with instances. Such perceptual acquaintance seems more clearly to ground our concepts than theoretical acquaintance does.

However, The BMC framework yields a second kind of a priori inference. First consider (9), which is analogous to Kripke's example.

(9) Apple is the unique natural-kind property that explains the similarities in the sample.

The inference to (9) is an a priori inference. Although contingent, (9) then licenses our second, more interesting, inference from (10) to (11).

(10) Object x has the unique natural kind property that all of those samples have that explains their similarity in appearance.

When our agent encounters a new object, x, and comes to believe (10), she would be justified in the contingent inference to (11).

(11) x is an apple.

5.3 Refining the Reference-Fixing Description

Suppose that the objects in a given cluster more-or-less share the property-kinds, p_1, … p_n. The most obvious description the agent could use to pick out the reference is the following description, D1.

(D1) The property-kind of having the conjunction of property-kinds (p_1 & p_2 &… & p_n)

According to the advocate of BMC who proposes Description 1 as the reference-fixing description, an agent who introduces the mental symbol APPLE when presented with a sample of apples, exploits a reference-fixing description such as (for example) 'the property-kind of being red and round'.

Description D1 is not, however, a plausible reference-fixing description. The property of being an apple is not even close to being extensionally equivalent to the property of being red and round. When we think about apples, we are thinking about that property; we are not thinking about red and round things.

A better approach is to say that when confronted with a set of sample objects, (o_1, … o_n), that are similar in appearance, there is a unique property that they all share. This

motivates the following reference-fixing description, D2, which is clearly ostensive in Soames' sense:

(D2) The property-kind that is shared by objects (o_1, &…& o_n)

Notice that D2 requires that there is a unique property that is shared by the objects in the cluster. The idea is that when presented with some apples, the agent infers that there is a unique property that the apples have in common, namely that they are apples. However, D2 will not work to pick out the right property. There may well be a unique property that the objects in the sample have in common (for example 'being in front of me and being an apple and being red and being round and having slight indentations either on the top or the bottom, etc'). But this property is not extensionally equivalent to the property of being an apple. Presumably, few (if any) other apples have this conjunctive property.

The advocate of D2 could respond by restricting her attention to properties that are non ad-hoc in various respects, say simple non-conjunctive properties. On this view of properties, being red and being round are properties, but being the conjunctive property of being in front of me and having some indentations on the top is not a property. However, so understood, D2 fails even more obviously. For there are too many properties the objects in the cluster have in common besides their being apples. For example, they are all red, and all round, and since they are apples, they are also all sweet, and ripe, and edible. In other words, there is no *unique* simple property that is shared by all of the objects.

One way to eliminate these extraneous common properties that the reference-fixing description be restricted to pick out only *natural-kinds*. The properties of being

red, round, sweet, and so on, are all simple properties of the objects in the sample, but, unlike the property of being an apple, they are not *natural kinds*. This leads us to description D3 as a way of uniquely picking out appleness.

(D3) The natural kind that is shared by objects (o_1, &…& o_n)

At first, D3 seems to be a good candidate reference-fixing description to be used in The BMC framework. Two worries arise however. First, the insertion of 'natural-kind' in the description seems to suggest that the agent has to have the concept NATURAL KIND in order to acquire new concepts in this way. This means that the agent must either have NATURAL KIND innately or acquire it in some other way. For now, let us suppose that this concept is indeed innate. Notice also that using 'natural kind' in the description limits the scope of the account, for it cannot be used to pick out non-natural yet simple kinds, such as artifact kinds, social kinds, etiological kinds and so on. We will return to this issue shortly.

A more urgent worry is that D3 still fails to satisfy the uniqueness condition, in spite of eliminating simple properties that are not natural-kind properties. If in fact the members of the sample all have the property of being apples, they also have other natural-kind properties, like being fruit, and being organic. The linguistic version of this worry is well-known in the philosophy of language as the 'Qua-problem'. Michael Devitt and Kim Sterelny discuss this problem in their (1999) textbook *Language and Reality,* when a kind-term is being grounded, or assigned a meaning.

> It seems that the grounder must, in effect and at some level, "think of" the sample as a member of a natural kind, and intend to apply the term to the sample as such a member… The term is applied to the sample not only

> *qua* member of a natural kind but also *qua* member of a particular natural kind. Any sample of a natural kind is likely to be a member of many natural kinds; for example, the sample is not only echidna, but also monotreme, a mammal, a vertebrate, and so on [2nd ed, Pg. 91].

What is needed is a way to isolate the property of being an apple from all of the other natural-kind properties that are shared by the objects in the sample. For the Qua-problem, I propose that the agent must explicitly invoke the notion of explanation in baptizing the mental term. As was suggested above in the sketch of the model, the detected clustering should be *an indication to the agent that there is a property that is shared by the objects in the sample and is responsible for the similarity in appearance*. Of course, as with the notion of natural kind, including the notion of explanation in the description seems to suggest that the agent needs to have the concept EXPLANATION in order to acquire concepts through this process. For now, assume the agent has such a concept, so we can see the work that it can do. Consider description D4.

(D4) the property-kind that objects (o_1, &…& o_n) have that explains their similarity in properties (p_1, &…& p_n)

This fourth description eliminates many of the non-intended properties. Although the set of objects in the sample of apples would have many other natural-kind properties, it is presumably only their property of being apples that *explains* the observed clustering over redness and roundness. The property of being a fruit would be explained by a larger cluster, one that includes blueberries, bananas, and peaches. Likewise, the property of being organic would be explained by an even larger cluster, one that includes fruit as well

as other plant and also animal matter. See image 5.1 for a 3-dimensional space with six clusters, within two 'clusters of clusters', or super-clusters.

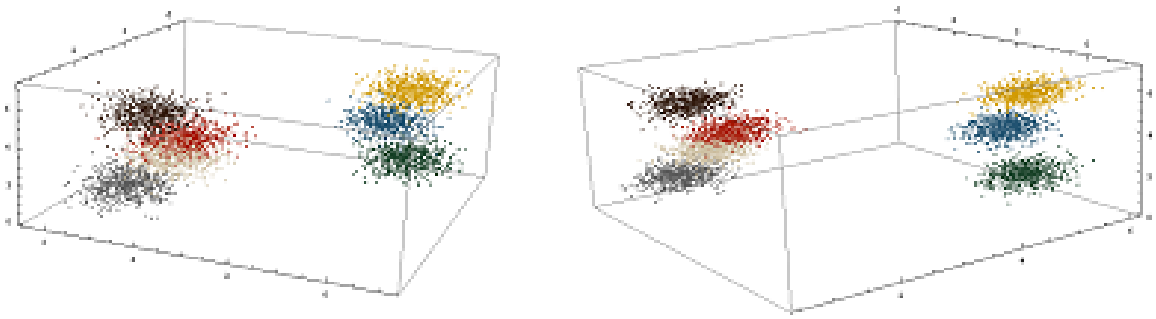


Figure 5.1., super-clusters.[17]

Description D4 works quite well, but it doesn't eliminate all of the unintended natural-kind properties. For, it only eliminates the super-ordinate natural kinds. Suppose that all of the apples in the sample were in fact of a particular variety of apple, like being a McIntosh. The kind *McIntosh apple* is a natural-kind, so that is the property that would explain the clustering of the sample of objects in the region of the property space they occupy. A set of observed apples that included other kinds of apples, like Granny Smiths, which appear greenish in color, would presumably form a super-cluster, over these two apple varieties.

To deal with this McIntosh apple problem, I submit that it is an instance of a more general problem that occurs frequently during human concept acquisition. I call it the *Limited-Sample* problem, and it is a problem that faces the cognitive agent, not the concept theorist. If a child is exposed only to a limited sample of apples, all of which are

17 Image created by Rob Zinkov, Spring semester, 2008.

McIntosh apples, and their being McIntosh apples is what explains the similarity that leads the child to coin a mental term, the mental term thereby coined means *McIntosh apple*. Perhaps later, if the child encounters other varieties of apple and clusters them along with the original set, her mental term will come to mean *apple*. Alternatively, if the child sees a new set of apples, all of which are Granny Smiths, the child may form two separate clusters. If a third dimension of features, say a taste dimension, is added, resulting in a super-cluster of McIntosh and Granny Smith apples, the child may then pick out the property of being an apple via D4.

A more immediate problem for D4 is that without the notion of a natural-kind explicitly involved in the description, it will not serve to pick out the property of being an apple.[18] The property of having similar surfaces, having apple skins, is what explains the similarity in appearance among the reddish and roundish objects in the observed sample. The notion of a natural kind will have to be used in the reference-fixing description along with the notion of explanation. This brings us to description D5.

(D5) The natural-kind that the objects (o_1, &…& o_n) have that explains their similarity in properties (p_1, &…& p_n)

Description D5 seems to do all of the needed work. The kind apple is the unique natural kind that all of the objects in the sample have that explains the clustering around being red and being round. If the agent can formulate such a mental description, she can thereby come to think about the property of being an apple. From this, the agent is a mere baptismal step from having a concept that means *apple*, and is both simple in representational structure and from other representations.

18 This observation is due to Adam Sennet.

Instead of D5, however, we may be able to adjust D4 without the naturalness requirement. This would allow the description to generalize to artifact kinds, social kinds, etiological kinds, and so on. I suspect that the notion of a *best explanation* is involved when human beings try to categorize the world. This doesn't mean that the agent has to know *what* the best explanation is; the agent needs only to suppose there is one and pick it out. I suggest D6 can do this.

(D6) The property-kind that objects (o_1, &…& o_n) share that *best* explains their similarity in properties (p_1, &…& p_n)

Toward one final refinement, consider the following scenario.[19] Suppose that a set of objects appears similar to an agent, *under certain environmental conditions*. If the agent takes the conditions to be stable and notices a clustering over features, this is enough of an indication that the objects share a natural-kind property. It is enough of an indication, even if the objects don't in fact have the properties they appear, under those conditions, to have. This consideration generalizes to include the cases in which samples appear similar in normal conditions. The result is description D7.

(D7) The property-kind that objects (o_1, &…& o_n) share that best explains their similarity in appearances (a_1, &…& a_n) under environmental conditions (c_1, &…& c_n)

I propose that we adopt D7 as the form for the reference-fixing description to be used in the BMC framework process for concept acquisition.

---

19 This consideration was brought to my attention by Jeff King.

5.4 The Existence and Uniqueness Conditions on Descriptions

As Putnam notes in the passage above, the linguistic baptism of kind terms carries empirical presuppositions. This is the same for The Baptizing Meanings for Concepts framework. For example, the introduction of the mental term APPLE involves presupposing that all of the perceived samples in the similarity cluster share at least one and at most one shared property-kind that explains the clustering in appearance. What is the status of baptized terms that fail to satisfy these conditions?

Let us call a description 'The F' a *proper* description if and only if the following two conditions hold:

(i) There is at least one F (Existence)

(ii) There is at most one F (Uniqueness)

If a description of the form 'the F' is to play any role at all in fixing the reference of a term, it seems the description had better be a *proper* description. In cases where it is proper, the description will pick out a property. But any theory that requires descriptions in the acquisition of concepts faces the problem of descriptions that fail to be proper.

First let us consider the Uniqueness condition. A case that comes to mind immediately is the concept JADE, because it was recently discovered that there in fact are *two very different* minerals that people have been treating as a single kind. It is important, however, to notice that JADE violates the *Existence* condition, not Uniqueness. That is, there does not exist *any* property-kind that explains the similarity in

appearance of the samples. JADE will be discussed later on in this section, when we discuss cases that violate the Existence condition.

The Uniqueness condition has to do with isolating one level of abstraction, or *kind*, when there exist multiple kinds within a sample. This condition played a central role in guiding our refinement of the reference-fixing description in the preceding section (section 5.3). I suspect that much more work is needed to really isolate, for example, the kind *apple*, but let us use description D7 for this dissertation.

With the Uniqueness condition aside, we can now turn directly to the Existence condition. Scott Soames discusses the failure of these conditions for the linguistic baptism of natural-kind terms in his (2002) extension of Kripke's project [pgs 281-284]. The considerations involved in arriving at description D7 were already guided by the uniqueness condition. We presently consider several ways in which the existence condition might fail and the options available for the failures, as guided by Soames' treatment for the linguistic case.

The case of JADE comes to mind immediately, and is a common example of a term for which the existence condition failed. When the linguistic term 'jade' was initially introduced, and for a long time afterwards, it was assumed that all of the samples in an observed set shared a property-kind that was responsible for the similarity in appearance. From this assumption, a (mental) description was formulated to pick out that purported property-kind, a description like 'The property-kind that these samples share that explains these similarities in greenish-blue, stone-like appearances under normal conditions'. More recently, it has been discovered that the samples we have been

considering under the term JADE in fact have two very different molecular structures. The case of the concept ARYAN might be another interesting example. For the set of underlying properties responsible for the similarity in look seems to be extremely gerrymandered.

It is not clear how this new knowledge affects our concepts. Some claim that we weren't wrong to think that there was a property-kind there, which all of the samples share; all we have learned is that there are two kinds of jade. If this is the case, however, it is not all clear what it would take to be mistaken about natural kinds. More plausibly, what happened is that we mistook the superficial similarities among the samples to reflect a deeper similarity of natural kind. This would lead us to revise our carving of the world, eliminating the jade carvings and instead carving at the molecularly distinct kinds, jadeite and nephrite. If this latter is the right treatment, then we had a mistaken assumption underwriting our conceptualization of JADE during all those times before the new discovery. We thought there was a property-kind that explains the similarity in appearance, but in fact there was no such property-kind. In other words, the existence condition failed. What was the status of JADE during all that time? Is appropriate to say that it was a concept? If so, what was its meaning?

Option 1: the concept JADE didn't have any meaning at all; it was empty.

There is a sense in which this seems right. There is in fact no such property as the one that JADE was supposed to pick out. However, as Soames points out for the linguistic case, this verdict seems a bit harsh. After all, there does seem to be some

information that was expressed by the term ‘jade’. People managed to interact with one another and with the world in ways that suited their beliefs and desires, buying and selling the set of stuff they collectively carved as jade, indifferent to the falsity of the existence condition. Likewise, the concept JADE seems to have had cognitive significance to these people, regardless of what the world turned out to be like. The advocate of Option 1 could perhaps explain these facts solely in terms of syntactic properties of the concept JADE. Again, consider ARYAN as another example, for which this option may seem much more plausible.

Option 2: the concept JADE picked out the disjunction of natural-kind properties that in fact were in the sample. The thought here is that although it was assumed that what JADE picked out was a simple property, in fact it was picking out a complex, disjunctive property.

At first glance this option might seem crazy. It looks like it is suggesting that the world determines the simplicity or complexity of our concepts. Notice, however, that the suggestion is not that the concept itself is complex and disjunctive, unbeknownst to the agent. The externalist factor determines the complexity in the *semantics* of the concept, not its *syntax*. That is, from the inside, the agent thought that her concept JADE was picking out a simple property, but in fact it was picking out a disjunction. This is just a version of semantic externalism, which is already well understood in the philosophy of language and in the philosophy of mind, and at least some degree of semantic externalism is increasingly becoming endorsed in both fields. Still, for the concept ARYAN, this

option seems quite implausible. The disjunction is far too gerrymandered to be a useful category.

> Option 3: the meaning of JADE, when it was coined via a description that fails at the existence condition, was the complex meaning of the description involved in its acquisition.

As in the case of Option 2, this suggestion is not that the concept itself is complex, but rather that its *meaning* is complex. From the inside, the agent takes JADE to pick out a simple property. But if the world does not comply, then the meaning of JADE is complex. In particular it is the meaning of a definite description that has the structure of description D7.

There is, however, a serious problem with this option. Descriptions of the form of description D7 involve ostensive reference to a presented sample of objects. Different agents will have acquired the concept JADE on the basis of different samples. If the meaning of a particular introduction of JADE is the meaning of an instance description D7, then different agents will have JADE concepts with different meanings. If concepts are individuated in part by their semantic properties, the concepts themselves will be different. For example, if I formed the concept JADE via some observed similarity in samples, I ought to be able to attribute a concept with the same meaning to a sales person at a gem show. However, presumably her concept was acquired via some very different set of observations. The meaning of my token of the concept JADE cannot be identical with the meaning of the mental description that I used in acquiring the concept, if it is

also supposed to be identical with the meaning of the sales person's token of the concept JADE.[20]

I have considered three options here for dealing with cases like JADE, in which the existence condition fails. Any of these is available of an advocate of BMC, depending on the costs she is willing to incur. I have a modest preference for Option 1, as I find that the informativeness of the concept can fully be explained in terms of syntactic processes internal to the agent, in spite of the failure of the concept to hook onto any external meaning. The agent treats the representation as if there were in fact a unique property that it picks out.

5.5 Natural Kinds, Artifact Kinds, Social Kinds

In refining the reference-fixing description to be used in conceptual baptisms, it seemed at one point that the notion of *natural* kind needed to be part of the description in order to pick out lexical properties (see section 5.3 (D5) above). My hope is that the notion of a *best explanation* does the needed work to isolate the right property-kinds. Let us consider the similarities and differences between *natural kinds, artifact kinds, and social kinds*, as pertaining to reference-fixing. I will then argue that the BMC can use the notion of *kind that best explains*, as part of the reference-fixing description D7.

The notion of *natural kind* is a technical, theory-laden notion in philosophy. The theory-laden notion is that natural kinds are the kinds that form *real* categories in the world, in contrast with *man-made* ways of categorizing the world. Depending on your metaphysics, you might take as real only the most fundamental properties in the world (at

20 This response came from discussion of a related issue with Cody Gilmore.

the level of *quark*, *spin*, *string*, and such), or you might include as real the properties that are regular, non-arbitrary combinations of these fundamental properties in the world. This latter metaphysics might take as real some of the lexical properties, like *atom*, *molecule*, and *cell*, some may take higher-level properties, like *water* and *gold*, as similarly regular combinations of fundamental properties, and some may go further to include *apple*, *tiger*, and *ocean*.

Philosophical tensions about the reality of a category grow stronger when we consider man-made categories. It is often taken for granted that artifact kinds are non-real, in the sense that they are arbitrary combinations of real properties. But there is a distinction to be drawn between man-made categories and categories *of* man-made entities. Artifact kinds are the kinds of things that are made by human beings, in the sense that they are literally *built*, from material substances in the world. Typical examples are *chair*, *table*, and *car*. There is no a priori reason, however, to think that all entities that are built by human beings are non-natural. Indeed, consider molecules that are created by human beings, and man-made lakes. Likewise, there is no a priori reason to think that all man-made categories are artifact kinds. Social kinds, for example, are usually thought to be man-made categories, and some philosophers argue of some categories of things that are not literally constructed by human beings, like *jade*, *tiger*, *molecule,* and *string*, are man-made in the sense that the category is man-made. More precisely, there is a distinction between (i) and (ii).

(i) The man-made category X, where X is the category of entities $x_1 \ldots x_n$.

(ii) The category X, where X is the category of man-made entities $x_1 \ldots x_n$.

Now, we can clarify the issue of natural vs. non-natural kinds. Recall the Uniqueness and Existence conditions from the previous section (section 5.4). If the reference-fixing description that is used in the acquisition of a concept for a non-natural kind, we have to worry about the meaning of the concept. As we determined in the discussion of JADE in section 5.4, and in following Soames' treatment, there seem to be several options a baptism theory could take. One option was that the meaning of the description is the meaning of the concept in such cases. But recall that this does not require that the concept itself is a complex description; it may remain simple, and treated by the cognitive system *as if* there does exist such a property. The important thing to notice is that

This approach comports well with the data. Concepts for artifacts, like PENCIL, UMBRELLA, COMPUTER, and GLASS, don't have as their meanings the reference-fixing descriptions that are used in their acquisition. A child might see and even interact with instances of the categories picked out by these concepts without knowing their functions. Indeed, most adults who interact with computers have very little understanding of what makes something a computer, and scientists today still don't understand the nature of glass. This doesn't prevent us from thinking about glass. Likewise, social concepts, like PHARMACY and PHILOSOPHER don't have their acquisition and recognition features as their meanings. Most ordinary people have very little understanding of what a philosopher does. Even philosophers themselves, and

philosophers *of philosophy*, struggle to understand what makes something a philosopher. Regardless of this lack of understanding, they are able to refer to the kind.[21]

5.6 The Innateness of EXPLANATION and KIND

Let us consider now a set of objections that have to do with the terms 'natural-kind property' and 'explanation' being in the reference-fixing description. The way the account was stated, it looks to be committed to the agent explicitly formulating such a mental description during the concept-acquisition process. If this is the case, then the agent must have the concepts KIND and EXPLANATION in order to acquire concepts in the way suggested by The BMC framework model. This means either that these concepts have to be possessed innately or that they have to be acquired via some other process.

For the issue of the reference-fixing description being *explicitly* represented by the agent during the concept-acquisition process, a couple of issues arise. First, it is not clear what it is for something to be explicitly represented. It seems clear enough that the description is not consciously represented in a way that is introspectible, since we don't notice ourselves formulating such descriptions. But conscious representation is taken to be different from explicit representation. Perhaps all the work that needs to be done by the description can be done the mechanics of the agent. However, it is not clear to me what the difference is, and when considering how to implement such an agent the

21 Many of these observations about functional and social kinds were brought to my attention by Michael Johnson in a talk at Rutgers University.

distinction only becomes more confusing. I will put this issue aside for now, and assume that in fact the description is explicitly represented.[22]

This means that if Description 7 is the right description, and it involves the terms 'natural kind' and 'explanation', we have to consider the plausibility of the agent having these corresponding concepts either innately or by some other acquisition process that occurs before The BMC framework process is carried out.

For the current proposal, I shall take the concepts KIND and EXPLANATION to be innate. Both of these concepts are involved in the naïve/psychological essentialism observed in infants (see section 4.4). Humans seem very early in development to explain observed properties in an entity by its essence, its necessary properties, and these posited essences are used in the classification of entities.

This innateness postulation should not be uncomfortable for either the Empiricist or the Nativist. The Lexical Concept Nativist is already comfortable taking the majority of our lexical concepts to be innate. For example, according to the Lexical Concept Nativist, concepts such as CARBURETOR and APPLE cannot be acquired, and are therefore innate. It is difficult to see why these sorts of considerations do not also extend to the concepts KIND and EXPLANATION. If CARBURETOR cannot be acquired, then surely KIND also cannot be acquired. The Lexical Concept Nativist is in no position to criticize alternative views for postulating innate possession of concepts.

The Concept Empiricist also should not balk at the innateness of KIND and EXPLANATION should likewise be comfortable for the Lexical Concept Empiricist.

---

22 Fodor discusses the explicit/implicit issue in his (1983). Robert Matthews challenges the distinction in his (2007).

Classical Empiricists must postulate a great deal of innate structure. For example, Empiricists already allow for the innate representations that make up the perceptual space, as well as an innate mechanism for doing statistics over the perceptual experiences in order to acquire new concepts. In short, Lexical Concept Empiricists attribute to agents an innate mechanism for constructing theories about the world, given perceptual stimuli. It is not difficult to see how an innate mechanism for forming hypotheses would need to employ concepts of KIND and EXPLANATION.

5.7 Acquiring Concepts through Language

The majority of this dissertation on the Baptism of Meanings for Concepts focuses on the acquisition of lexical concepts from reference-fixing descriptions that involve *perceptual* demonstrative representations. The agent infers from perceptible patterns to an unobserved property-kind that is shared among some perceived samples. Perceptual demonstratives, however, are not the only demonstratives that allow an agent to fix onto a lexical kind. Experiences with language can likewise connect an agent to the appropriate referent, thereby allowing for the acquisition of a mental term for that referent.

Fodor (1975) has argued that word learning is impossible without already having a corresponding mental term for the meaning of the word. Language, therefore, cannot precede thought in some sense. Nonetheless, exposure to words in a language can, via an inference-cum-baptism process, allow for the acquisition of lexical concepts. In some recent work (DeVault, Oved, and Stone 2006), this kind of concept-acquisition was discussed in the context of grounding meanings in Artificial Intelligence (AI) systems.

Our argument linked questions about meaning in AI systems to the notion of meaning-borrowing that is well known in the philosophy of language, since the 1970s work of Kripke (1972), Putnam (1975) and Burge (1979). Compelling arguments have emerged suggesting that an agent can point through words used by other speakers to a meaning, in spite of having almost no other beliefs about the meanings of those words. In Putnam's (1975), he showed that a human speaker can use the words 'elm' and 'beech' to mean *elm* and *beech* (the two types of tree) despite being unaware of any other property that distinguishes elm trees from beech trees. Indeed you and I, and most ordinary people are in this position. We have the concepts ELM and BEECH (we can think about each of these kinds of tree as such), and we can refer to them by their linguistic terms, even though most of us have almost no other beliefs about those trees. Not only can we lack perceptual beliefs about a kind, but we can have *false* beliefs it, while nonetheless possessing the concept. Burge showed this in his (1979) in which he describes a patient in a doctor's office complaining about 'arthritis in the thigh'. The doctor had to correct the patient by pointing out that arthritis is a disease of the joints, not the muscles. In doing so, the doctor was pointing out that *what the patient was referring to with 'arthritis'* is a disease of the joints. This is because the patient had 'borrowed' the meaning from other, more expert, users of the word.

It is tempting to object that using this kind of reference-fixing description doesn't give an agent enough information about a kind for it to be sufficient for concept acquisition. It seems too easy of a way to come to know contingent information about a kind. This objection is addressed in chapter 12 of this dissertation, section 2.1.

5.8 Core Claims of the BMC

1. An entity x is a concept token if and only if x is a token mental entity that is stored to be used by the agent during thought to represent/ be about/ have as its meaning /be a symbol for, a (purported) *kind* to which entities may belong.

2. Kinds cannot be directly perceived; only particulars can. Perceptual representations are fully non-conceptual in this sense.

3. An agent A has a concept token x only if it is part of A's design goal to use x to represent a *kind* to which entities may belong.

4. A concept token x is tokenized anytime it is used during occurrent thought processing.

5. For any concept token x and any concept token y, where x and y are not identical, x and y are of the same concept type if and only if x and y have the same meaning.

6. For any concept token x and any concept token y, where x and y are not identical, x and y have the same meaning if and only if they have the same referent.

7. A semantic science of thought-processing is impossible, even if it is partly framed in terms of syntax. Neither meaning nor syntax can be generalized across agents with different experiences.

8. A concept token x has a meaning if and only if x manages to refer.

9. A concept token x has its referent innately or brute-causally only if it is a perceptual demonstrative. (In contrast, Lexical Concept Nativists treat lexical concepts as perceptual demonstratives for kinds.)
10. A concept token x is rationally/inferentially acquired if and only if semantic features of other representations played a role in making x mean what it means.
11. Most lexical concepts are acquired through a baptism process in which the agent makes an inference (usually inference to the best explanation) to the presence of a kind that is picked out by a mental description that picks out that kind.

Part III: Other Major Theories of Lexical Concepts

6 The Building-Blocks Framework

The majority of the concepts debate divides into views that treat lexical concepts as *complex* in representational structure versus views that threat them as representationally *simple*. Lexical Concept Empiricism and Lexical Concept Nativism are theories that divide on the representational structure of lexical concepts in this way. I find it helpful to think about these two theories in terms of a common background assumption which I call *The Building-Blocks Framework*. This will also set up the discussion of a few other theories that don't quite adopt that framework.

The Building-Blocks model is a natural way to account for these observations. Primitive representations compose together into Composites, building more and more complex meaningful representations. Since the Building-Blocks framework makes sense for the composition of propositional and phrasal representations, it seems natural to suppose that the same model of composition would be used when considering the lexical concepts.

According to the RTT, thoughts are structured representations, analogous to linguistic representations, with the simplest units of meaning composing together (in accordance with some 'grammar', or composition rules) into more and more complex units of meaning. The Building-Blocks Assumption is a picture of how mental representations combine that furthers the analogy. On this picture, there are Primitive representations, which are the basic building blocks of thought, and there are Composite representations built up from the Primitives. The Composites inherit their meanings completely from their representational parts, and they do so by a *rational*-causal process,

such that they are inferentially related to their parts. Most theories under the RTT agree that full propositional thoughts (for example A BLUE APPLE IS NEXT TO A ROUND RED SNAIL) as well as phrasal concepts (for example, BLUE APPLE, IS NEXT TO, and ROUND RED SNAIL) are among the Composites.

The Primitives ground all of the meanings in some basic way. There are several different accounts of that basic way, but almost all of them appeal to a relationship between the representation and its cause. The cause is usually taken to be either a distal stimulus (something in the outside world) or else a proximal stimulus (retinal stimulation or sensations themselves). These accounts further spell out just what that causal relationship is, as in Dretske's (1981) informational account, or Fodor's (1990) Asymmetric Dependence account, or Millikan's (1984) Teleo-functional account. The important thing is, almost everyone under the RTT agrees that the Primitives are *brute*-causally related to their meanings –the world and our brains are innately built so that certain properties in the world cause these representations to trigger. They are the representations we get directly from perception (for example, the taste of sourness or the experience of redness, if those are representational, or else whatever representations come as a direct result).

I find that the Building-Blocks picture forces a conflation of the acquisition and meaning-determination factors for representations. I pull these factors apart in my characterizations of Primitive and Composite representations.

Composite Representations:

(Rationally Acquired)	Acquired from more primitive representations through a rational/inferential process.
(Complex Symbol)	Complex in representational structure, in the sense that it represents a complex meaning that is composed from the meanings of its representational parts.

Primitive Representations:

(Innate or Brute-Causal)	Possessed at birth, or else acquired only by brute-causal (i.e., non-rational, non-inferential) processes.
(Simple Symbol)	Simple in representational structure, in the sense of directly representing a meaning, rather than having the complex meaning of its representational parts.

Note that even John Locke (1690) distinguished Primitive and Composite representations in this way –the sense in which he took perceptual representations to be 'acquired' is the same as the brute-causal sense of 'innate' that we are using here.[23]

For researchers within the Building-Blocks framework, questions about concepts are framed in terms of the framework. In particular, the debate between Lexical Concept Empiricism and Lexical Concept Nativism is almost completely about where to draw the line between the Primitives and the Composites, focusing on the lexical concepts. Recall that the lexical concepts are, roughly, the ones that tend to be expressed by single words, the smallest units of meaning in most natural languages. These include the concepts BLUE, APPLE, NEXT-TO, and SNAIL.[24] Lexical Concept Empiricists usually put lexical concepts with the phrasal representations, claiming that they are Composite (acquired from other representations and built up from the representations involved in their acquisition). Lexical Concept Nativists put them on the side of perceptual representations, taking them to be Primitive. Empiricists believe, for example, that the concept APPLE is acquired and built from more simple representations, perhaps RED, ROUND, CRUNCHY, and SWEET, which in turn may be acquired and built from further Primitives. For Nativists, APPLE is among the sensory representations, set up at birth, or through brain development, to be triggered by apples in the world.

---

23 Jerry Fodor's (1981, 1985, 1998) notion of innate simples and John Locke's (1690) notion of acquired simples both count as 'Primitives' in this sense. Their views diverge in ways that will become clear later –Fodor takes most *lexical* concepts to be among the Primitives while Locke takes them to be among the Composites.

24 Note that the notion of a 'lexical concept' involves no commitment to the agent's having a natural language term that corresponds to it. Lexical concepts are simply the mental representations for properties that tend, in most languages, to be expressed by single words *when and if expressed*.

(Lexical Concept Empiricism) Most lexical concepts are Composite.

(Lexical Concept Nativism) Most lexical concepts are Primitive.

Recall the imagined agent that was used in the description of the BMC in Part II of this dissertation. Lexical Concept Empiricists and Lexical Concept Nativists have always assumed that human beings have something like an innate perceptual space, that we are born able to represent the world perceptually. Both views usually also agree that patterns in perceived entities, such as the clusters involved in the BMC, lead to the acquisition of new Composite concepts. They views disagree, however, about where the *lexical* concepts appear on the picture. Lexical Concept Empiricists claim that most lexical concepts are the Composite clusters of representations, whereas Lexical Concept Nativists claim that most lexical concepts are themselves among the Primitives that make up the agent's innate perceptual space. Of course, the BMC framework departs from both of these views, treating lexical concepts as neither Composite nor Primitive as defined here. On the BMC, most lexical concepts are acquired from other representations but are simple in their resulting representational structure.

It is important to note that there are different ways in which a concept may be Complex. First, recall that we have distinguished the entity, *x*, that is the concept and the entity, *y*, that is the meaning. Typically, when a concept is thought to be Complex, theorists take both *x*, the concept, and *y*, the meaning, to be complex entities, and that is how we have defined 'Complex' here. The temptation that should be avoided, however, is to then treat the complex entity that is the concept, *x*, and the complex entity that is the

meaning, *y*, as *one and the same entity*. Of course, one may hypothesize that in fact these are the same entity, but one should do so explicitly, since it is not the default.

The Building-Blocks framework helps to set up our discussion of the current space of theories for lexical concepts. Not all theories completely fit within that framework, but almost all of them involve some aspects of it.

7 Lexical Concepts as Composite

In this chapter we step through some theories that treat lexical concepts as Composite in the sense of being acquired from, and composed by, other representations. There are three major variations on this view.

7.1 Complex and Acquired by Composing Definitional Inferential Relations

> *Lexical Concepts as Definitional Sets of Representations:* Lexical concepts have (1) complex structure, (2) complex meanings, (3) by virtue of inferential connections to symbols that make up their definitions (necessary and sufficient conditions for application), (4) established by experience and inference. (5) Concept token *x* is of the concept type *T* if and only if *x* means *T*.

The traditional version of Lexical Concept Empiricism is the Definitions Theory (also known as the Classical Theory). On the Definitions Theory, lexical concepts have definitions, and they are acquired and built up from the representations that make up their definitions. In other words, for most lexical concepts there are necessary and sufficient conditions for the correct application of the concept, and the concept is acquired and built from representations for those conditions. The concept BACHELOR, for example, if taken to be defined by UNMARRIED and MAN, would be thought to have been inferred and built from those representations; the concept APPLE, if defined as RED, ROUND,

SWEET, and FRUIT, would be thought to have been inferred and built from those representations; the concept CAT, if defined as FURRY, QUADRUPED, WHISKERS, and ANIMAL, was thought to be inferred and built from those representations; and so on. Various theorists under this classical version of Empiricism fight over the definition of a given concept, but it is agreed that most lexical concepts have definitions and that the concepts are inferred and composed from the representations that make up their definitions. The Definitions Theory was the assumed view throughout much of the history of philosophy and psychology. The view is defended explicitly by Locke and Hume in the 17th and 18th Centuries, and it is still defended by contemporary authors. In his *Essay* (1690), Locke describes his theory of simple and complex ideas as follows:

> Thus, the idea of the sun,- what is it but an aggregate of those several simple ideas, bright, hot, roundish, having a constant regular motion, at a certain distance from us, and perhaps some other…" [Essay, Book II, part xxiii, sec 6], and "the greatest part of the ideas that make our complex idea of gold are yellowness, great weight, ductility, fusibility, and solubility in aqua regia, &c., all united together in an unknown substratum… [Essay, Book II, xxiii, sec 37].

Hume defends it in his *Treatise of Human Nature* (1739), where he constructs a theory of simple and complex ideas,

> The complex [ideas] … may be distinguished into parts… a particular colour, taste, and smell, are qualities all united together in this apple... *all our simple ideas in their first appearance are deriv'd from simple impressions*... [Treatise, Book I, part I, sec I].

The view is echoed in contemporary philosophy, as in Jerrold Katz (1972),

> [T]he English noun "chair" can be decomposed into a set of concepts which might be represented by the semantic markers… OBJECT, PHYSICAL, NON-LIVING, …, SOMETHING WITH LEGS, SOMETHING WITH A BACK, … [pg. 40].

Most people seem to find it *obvious* that many of our lexical concepts are acquired by a rational process that responds to patterns in our experiences. Consider the concepts SNAIL, MUSHROOM, DINOSAUR, and COYOTE. There is a strong intuition that these concepts have been acquired as a result of our experiential contact with the world. It is almost impossible to imagine how we could have these concepts innately, or why we would. Moreover, it seems like the world has given us good reasons to think that things in the world fall into these categories. Of course, on the Building-Blocks framework, if lexical concepts are acquired from patterns in sensory experience, they must have those sensory representations as part of their representational structure.

When adopting the tenet that lexical concepts are rationally acquired, that tenet is often coupled with the tenet Complex. This is partly because of the Building-Blocks model, but it is also because Complex gives an easily comprehensible account of *how* concepts can be acquired. That is, we notice patterns in terms of representations we already have, and then we infer that there is a category that is composed of the features in the recognized pattern. The Building Blocks framework is a natural way to understand how concepts can be used to generate new ones –they are generated by composition.

This is a large part of the reasoning that brings Lexical Concept Empiricists to believe that lexical concepts are representationally complex. That forces them against the strong simplicity intuition. Proponents of lexical concepts being complex have to say that even though it *seems* like we think about the world directly as comprised of apples and mushrooms, in actuality we think about these properties by thinking about complex compositions of properties, like red, round, crunchy..., or white, smooth, curved... This result is somewhat counter-intuitive, but not impossible to swallow. These theories

usually take the representational structure of lexical concepts to be essential the representation-making relation, the metaphysics of representation, for lexical concepts. Views about the complex structure of lexical concepts divide into two categories –those that take the concepts to be built from their definitions, and those that do not.

7.2 Complex and Acquired by Composing Non-Definitional Inferential Relations

> Lexical concepts have (1) complex weighted conjunctive structure, (2) complex weighted conjunctive meanings, (3) by virtue of the representations in the complex weighted conjunction is the subset of its inferential connections to other symbols, (4) established by experience and inference. (5) Concept token *x* is of the concept type *T* if and only if *x* means *T*.

In reaction to the un-definability observations, many contemporary Lexical Concept Empiricists adopt revised versions of Lexical Concept Empiricism, ones that cut the link between a concept's being un-definable and its lacking representational complexity. These take lexical concepts to be composed by non-definitional beliefs, such that the meaning of the concept is, roughly, the entity kind of which most of, or some weighted sum of, beliefs $b_1 \ldots b_n$ are true. Beliefs $b_1 \ldots b_n$ are some proper subset of the beliefs the agent holds about entities to which the concept applies.

Versions of non-definitional complexity include Prototype Theory[25], Exemplar Theory[26], and Theory-Theory[27]. The most popular of these is Prototype Theory, on which lexical concepts have as parts the sets of representations of properties that are *typical* of their instances, rather than the sets of properties that make up the definition of the concepts. This way, the APPLE concept, for example, is composed of representations of properties that are typical of apples, like RED and ROUND, stored with a representation of the probability distribution over these properties across instances, rather than these properties being represented as necessary and sufficient for the concept. The view still maintains the connection between lexical concepts being acquired and their being complex and built from the representations used in their acquisition.

7.3 Complex and Acquired by Composing All Inferential and/or Causal Relations

> *Lexical Concept Holism*: Lexical concepts have (1) complex structure, with (2) complex meanings, (3) determined by its inferential connections to other symbols, (4) which get set up as a result of experience and inference. (5) Concept token *x* is of the concept type *T* if and only if *x* means *T*.

Strictly speaking, Holism does not endorse the Building-Blocks picture because there's no foundation, but the view does take lexical concepts to be complex in representational

---

25 Wittgenstein (1953, 1958), Rosch and Mervis (1975), Rosch (1978), Murphy (2002), Barsalou (1999), Prinz (2002).

26 Smith and Medin (1981).

27 Carey (1985), Gopnik and Meltzoff (1997), Keil (1989), Spelke (1994).

structure. Again, this view is in the first place a view about the meanings of linguistic terms, but a corresponding account for mental terms is almost always adopted alongside.

All lexical concepts, according to one version of mental meaning holism, are composed by their (inferential and/or causal) relations to all other representations, that is, they are composed by their place in the whole network of representations (Quine, 1951; Davidson, 1991; Block 1991). On another version (Dennett, 1991), all lexical concepts are composed by their causal relations, not only to other representations, but to non-representational, external features of the world as well.

One major motivation for meaning holisms is that they avoid the need for concepts to have definitions and principled analytic/synthetic dictions in their inferential relations. Since definitions and the analytic/synthetic distinction have been challenged, meaning holism is a way to maintain the representational complexity of lexical concepts without having to address those challenges.[28]

28 Special thanks to Joshua Howard for helping me wade through the literature on Meaning Holism.

8 Lexical Concepts as Primitive

There are two major accounts that treat lexical concepts as Primitive. They both take them to be representationally simple, but the one takes them to be innate, while the other takes them to be acquired, perhaps even through perception, but only via brute-causal, non-rational/non-inferential processes.

8.1 Simple and Innate

> *Lexical Concepts as Simple and Innate*: Lexical concepts have (1) simple representational structure, with (2) simple, referential meanings, (3) by virtue of direct reference, (4) present by birth. (5) They are typed by their meanings.

Lexical Concept Nativism is the view that most lexical concepts are Primitive. Like Lexical Concept Empiricism, this view has a lot of initial appeal. Indeed, the initial intuition when we experience and think about the world is that we do so in terms that are Simple. We feel as though we have direct access to the properties and kinds we represent with lexical concepts. The world presents itself to us as being comprised of apples and trees and tables, as simple objects with simple properties.

Also like Lexical Concept Empiricism, this view maintains a connection between a concept's being Acquired and its being Complex. Nativists usually begin with the intuition that many lexical concepts are simple and by pointing out problems for the tenet that lexical concepts are Complex. From the tenet Simple, then, Nativists argue that lexical concepts are Innate, with meanings determined directly by the external simple properties represented.

Between simplicity and complexity, initial intuitions favor simplicity. Keeping with the analogy between thought and language, it might seem natural to regard representations for these word-sized properties as the mental Primitives. We certainly *feel* as though we perceive and think about the world at that level. We see cars and people on the street, and hear trains and dogs, and our thoughts seem to be about things in terms of such categories. Historically this has always been the intuitive, default view. The trouble is, on the Building-Blocks Framework the concepts that are Simple in representational structure in this way are also supposed to be Innate. There is not any obvious way for Primitives to be Acquired. Giving in to the Simplicity intuition for lexical concepts, Nativists take on the innateness that comes along with it.

8.2 Simple and Acquired Brute-Causally

> *Lexical Concepts as Simple and Acquired Brute-Causally*: Lexical concepts have (1) no representational structure, with (2) simple meanings, (3) by virtue of direct reference, (4) set up after birth, possibly as a result of experiencing instances, but only via a brute-causal, mechanical, non-inferential process (at the level of the physics of brain tissue rather than a result of cognition). (5) They are typed by their meanings.

Fodor's early (1975, 1981) Nativist thesis seemed to suggest that these concepts are present at birth, just like our sensory representations. He later (1998, and in LOT2 2008) suggests that the acquisition might be achieved after birth, and even through perception, but only by normal brain development or by brute-causal interactions with the

world. That is, concept acquisition is not part of the story of thought or thought processing; it is a matter to be explained at the level of brain-tissue. We just have the sorts of brains with the sorts of physical structure that result, in normal conditions, to have primitively representational entities. A lot of the BMC story comports well with Fodor's (1998 [pg.136], 2008 LOT2) view. On his view, perceptual acquaintance with one or two good instances is enough to lock a concept. Here we have added a *mechanism* for such concept-locking. Good instances of a concept will naturally look more similar to one another and less similar to things that we take to be instances of other universals. It is because of this connection between good instances and similarity that the perceptual acquaintance with one or two good instances is enough to get us to designate a mental symbol to represent the (purported) universal. Of course, one upshot of this view of acquisition is that the visual system constrains which similarities human beings will pick up on, so it likewise constrains which concepts we will acquire. Let us consider both the brain-development and brute-causal accounts of lexical conept acquisition. Later we will see why the BMC, with its *rational/inferential* acquisition process, is superior to these accounts.

First, consider the brain-developmental acquisition account. Robert Rupert, in his (2001) *Journal of Philosophy* article, offers an account of concept acquisition that is compatible Robert Cummins' (1997) objection to brute-causal acquisition.[29] Cummins argues that lexical concepts can't be brute-causally acquired, on the grounds that it is incompatible with the evidence that *theories* mediate the acquisition process. In response, Rupert tries to show that the kind of theory-mediation that Cummins discusses

29 These authors talk about Fodor's 'Causal Theory', which is just another name for the view that lexical concepts are simple and acquired brute-causally.

is in fact compatible with brute-causal acquisition. Rupert discusses the notions of 'acquired' that he takes Cummins to have cited. Although he does offer plausible compatibility arguments for all of these notions, none of them are the notion of theory-mediated acquisition, i.e., rational/inferential acquisition, that is at the heart of Lexical Concept Empiricism. Rupert says rather clearly, "Cummins seems to assume that if a concept is not acquired by cognitively described means (as the result of learning, in particular), then the concept is innate. We should reject this assumption." He then goes on to show that brute-causation is compatible with a concept that is acquired (i) as a result of brain development, (ii) a theory that is implicit in the functional architecture resulting in the right causal relation, and (iii) not knowing anything about the concept's extensions at birth.

The second approach to the brute-causal acquisition of lexical concepts comes from Eric Margolis (1998), and is later endorsed by Fodor (LOT2 2008). It was the hope of Margolis to appease the tensions between Lexical Concept Nativism and Lexical Concept Empiricism, and it was his hope to do so by describing a process for the acquisition of representationally simple concepts. The trouble is, just as with Rupert, the sense of acquisition that is described doesn't satisfy the Lexical Concept Empiricist who insists that concept acquisition is a rational/inferential process.

Because the inferential acquisition of concepts holds our intuitions so strongly, most concept theorists ignore pressures against the complexity of concepts. To really preserve the inferential acquisition, however, we must meet what I call *Fodor's Challenge.*

> ***Fodor's Challenge*:** we must either (a) find a set of simple properties that necessarily applies to all and only the apples, and is plausibly the meaning of the concept APPLE, or (b) find a way for a simple concept that directly refers to the property of being an apple to be acquired not by brute-causation, but by an inferential process.

This is a challenge because (a) has been a struggle and option (b) seems impossible. The present work, of course, pursues option (b).

9 Two-Dimensional Semantics / Two-Factor Theories of Meaning

Two-Dimensional Semantics: Lexical concepts have (1) complex structure, with (2) two dimensions of meaning, both of which are functions from possible worlds to referents, where (3) the primary meaning of the symbol is a function from the world the agent is in when (acquiring and tokenizing) the symbol to the referent in that world, and the secondary meaning is a function from the world in which the symbol is being evaluated to its referent in that world. The representation-making relation is (4) set up by experience and inference. (5) They can be typed by each dimension or by the diagonal.

Two-Dimensional Semantics, as defended by David Chalmers (2005), is in the first place a theory about the meanings of linguistic terms. It is easy, however, to imagine that its proponents endorse a mental version of the view. Indeed, the view is a result of observations about linguistic terms that have been used in discussions of their corresponding mental terms. The observations behind Two-Dimensional Semantics are centered at a tension that arises between cases in which a single term in a language has different referents when uttered in different situations (a.k.a., Putnam's Twin-Earth cases) and cases in which two terms in a language refer to a single entity (a.k.a., Frege cases). Many philosophers of language, Chalmers included, discuss the observations in terms of possible worlds.

In his (1975), Putnam asks us to imagine a world that is almost exactly like our own, except that instead of H_2O, the watery substance in that world has some other chemical structure, XYZ. Indeed, for what we knew 300 years ago, we could have been in an XYZ world. In that world, when English speakers use the word 'water' they are

referring to XYZ instead of H_2O. There is no internal mental difference between a person in our world and a person in this twin world, but there is a difference in the meaning of the term 'water'. Putnam used this observation to argue that the meanings of our linguistic terms are not inside of our minds; their meanings are in the external world in which we happen to speak. The case translates directly to the mental version, in which a person in our world and a person in the twin world have identical mental terms with different meanings.

Frege cases raise tensions for such referential theories of meaning. Consider two different linguistic kind-terms that have identical referents, like the words 'water' and 'H_2O'. Both of these terms pick out the same substance-kind. All the samples of water are samples of H_2O, and vice-versa. At the same time, however, there does seem to be something *informative* about the statement 'water is H_2O'. It isn't trivially true as is the statement 'water is water'. The truth of 'water is water' is apparent from the syntax of the statement alone, whereas it was an empirical discovery that 'water' and 'H_2O' refer to the same substance. For Frege (1892) this informativeness observation suggests that there is a semantic element other than the referent of a term that accounts for such differences in information. Again, we might imagine a mental version of this case by imagining a person who has two different mental terms, two symbols, acquired via different reference-fixing descriptions. The mental terms may in fact refer to the same entity, unbeknownst to the person who possesses them.

So a tension arises. The Putnam cases suggest that meaning is a matter entirely of the external world in which the term is baptized, making it *impossible* for water not to be H_2O (or, rather, for 'water' and 'H_2O' to have different meanings). Frege cases, however,

suggest that meaning is entirely a matter of what beliefs are associated with the term by the agent and/or the linguistic community, making it quite *possible* for water to be H_2O (or, rather, for 'water' and 'H_2O' to have different meanings).

Two-dimensional semantics aims to resolve this tension by positing two dimensions of meaning. On the horizontal dimension, the term 'water', for example, has its meaning as determined by the world in which term is uttered (or baptized). In our world, there is a certain substance, call it *a*, that turns out to be the referent of the description involved in assigning a referent to 'water'. The watery stuff turns out to be substance *a*. Once assigned the referent, *a*, that referent sticks, or is rigidified, in the sense that the term 'water' refers to that same substance, even if the description associated with it would have picked out different substances, call them *b* or *c*, in another situation or possible world. The vertical dimension shows that whatever possible world the term is baptized in, the referent of the description used in the baptism in that world is sticks as the referent of the word.

'water'	Evaluated at W_0	Evaluated at W_2	Evaluated at W_2
Baptized at W_0	*a*	*a*	*a*
Baptized at W_1	*b*	*b*	*b*
Baptized at W_2	*c*	*c*	*c*

Table 9.1: Two-dimensional semantics of 'water'

The Frege cases are explained by the two-dimensional semantics being different for the term 'H_2O'. See Table 9.2 below. The first horizontal row of the two tables, *9.1* and *9.2*, are identical, reflecting the fact that when baptized *in our world* the descriptions used for 'water' and 'H_2O' picked out the same substance-kind, *a*. The second and third

horizontals of the two tables are different, however. This reflects the idea that the descriptions associated with ‘water’ and 'H_2O’ might not have picked out the same substance.

‘H_2O’	Evaluated at W_0	Evaluated at W_2	Evaluated at W_2
Baptized at W_0	*a*	*a*	*a*
Baptized at W_1	*d*	*d*	*d*
Baptized at W_2	*e*	*e*	*e*

Table 9.2: Two-dimensional semantics of ‘H_2O’

So far, we have been discussing Two-Dimensional Semantics as an account of *linguistic* terms. We must be careful here in spelling out the observations and translating them into their mental analogues. Notice that the linguistic cases have to do with (tokenizings of) public word *types*. That is, a public language like English has terms that are shared by its users. Linguistic term types, however, are largely *syntactic* types. What makes English utterances of 'water' the same is, in the first place, their same physical/syntactic features. It is far from obvious that *mental* terms have syntactic types in this same sense.

This observation may be exactly what drives some theorists to treat many lexical concepts as Composite. Composite concepts might be thought to be comparable across agents by their inferential roles. This is problematic, however, because the inferential roles presumably bottom out in Primitives (except on Holism), and the Primitives presumably lack syntactic structure that can be typed across agents. See the discussion of COYOTE in 10.2.4.

Part IV: Defense of the BMC

This part of the dissertation steps through the arguments presently on offer (some borrowed, some original) in support of the Baptizing Meanings for Concepts (BMC) framework. There are three chapters that address each of the controversial claims of the BMC. Chapter 10 defends the claim that most lexical concepts are simple in representational structure, chapter 11 defends the rational/inferential acquisition of lexical concepts, and chapter 12 defends the baptism process for coming to have these rationally acquired simples.

10 Why Lexical Concepts have to be Simple

This chapter has two sections. In section 10.1, I step through a list of observations, and show for each how it favors the representational simplicity of lexical concepts over their representational complexity. In 10.2, I step through the major arguments that have been made in favor of the representational complexity of lexical concepts and show how the BMC accommodates them. Refer back to the terms and distinctions in Part I or the discussion of simplicity and complexity in Part III for a reminder of these precise claims.

10.1 Observations Best Explained by Representational Simplicity

In this section I step through a list of observations about the representational structure of lexical concepts. I use these observations to show how theories on which lexical concepts are simple in representational structure, such as the BMC, offer better explanations than views on which lexical concepts are thought to be composed by other (usually more primitive) representations.

10.1.1 Almost No Definitions have been Found

Recall the Definitions Theory, on which most lexical concepts are composed by representations for the kinds that are necessary and sufficient for the concept to apply to an entity. Although highly appealing at first, this theory has been under attack for millennia. Even Plato observed that many lexical concepts lack definitions, as demonstrated through his Socratic dialogues, raising questions about definitions, like "what is piety?" in his *Euthyphro*. More recently, Ludwig Wittgenstein used the concept GAME in his (1953) to show that many lexical concepts lack a set of necessary and

sufficient conditions; that all and only the things that are games have, at best, some overlapping commonalities, or 'family resemblances'. Similarly, W.V.O. Quine (1951) raised problems for drawing a principled distinction between beliefs about a kind that are true by virtue of the meaning analytic/synthetic distinction that would be expected if concepts had definitions.

Trying to analyze a concept for PIETY or a concept for APPLE into a set of representations that clearly make up the necessary and sufficient conditions on the concept's application has proven extremely difficult, if not impossible. BACHELOR as defined into UNMARRIED & MAN, is the one example of definition that is almost always used by philosophers, plausibly because it is as good a candidate as there is for a definitional concept. Even with BACHELOR, however, counter-examples are easy to find –the pope is an unmarried man but not a bachelor, some bachelors are separated from their wives yet legally married, and many unmarried homosexual men are not bachelors because they are in life-long relationships. The situation is the same with most other lexical concepts. In the case of APPLE, for example, it is possible for something to be an apple even if it is blue, furry, rubbery, and talkative; it is possible for something to be red, round, and sweet (and even have that apply flavor) without being an apple.

Sometimes researchers propose that the true essence, the true necessary and sufficient conditions for a concept to apply to an entity, are the deeper features of the kind. Perhaps there is some particular DNA sequence that is shared by all and only apples, and we know that water is H_2O and gold is the substance with atomic number 79. The trouble with that approach, however, is that the deeper features of a kind are usually not known before the possession of the concept. We typically have a hunch that *there is*

some deep necessary and sufficient set of features for the kind, but we do not know what those features are as a way of coming to represent the kind. Water is H_2O, but the concept H2O is certainly not (part of) the concept WATER. It is obvious that children are able to tokenize one mental entity without thereby tokenizing the other mental entity, even if they have the same meaning. This is what makes it a *scientific* discovery that the linguistic terms 'water' and 'H_2O', and their corresponding mental terms, pick out the same substance-kind. Moreover, this is what allows linguistic agents to communicate about the *same* kinds and categories in their shared world, and teach one another what they believe about entities of those kinds.

Another attempt at definitions is to define a concept in terms of its *super-ordinate* kinds (Jackendoff, 1998). The concept APPLE, for example, would be defined in part by FRUIT, which in turn would be defined in part by ORGANIC, which would be further defined in part by OBJECT. The trouble with this approach, however, is that nobody has any plausible suggestions for what needs to be added to the super-ordinate kinds, like FRUIT, in order to be sufficient for the lexical concept, APPLE. Of course, being a fruit is a *necessary* condition for being an apple, and it would be great if a theory of concepts could explain that. But definitions require the full set of necessary *and sufficient* conditions for satisfying the concept.

10.1.2 Compositionality

The problems observed for trying to find definitions for lexical concepts in terms of other concepts (and/or non-conceptual representations) have led many concept theorists, beginning with Wittgenstein, to formulate theories that treat lexical concepts as *non-definitional*, more probabilistic, fuzzy sets of representations. With the work of Eleanor

Rosch (1978), the idea of abandoning definitions while maintaining representational complexity has inspired the approach to the study of concepts that is dominant today across the philosophy of mind, epistemology, cognitive psychology, and artificial intelligence (Prototype, Exemplar, and Theory Theories).

Theories that take lexical concepts to be composed from non-definitional sets of representations face the well-known Compositionality Argument (Fodor, 1998). Again, the idea is that lexical concepts are complex in representational structure, where they are built from (all or a proper subset of the) representations that encode the agent's beliefs involving the concept. Recall that compositionality, at least the composition of lexical concepts into phrasal concepts and thoughts, was one of the original motivations behind the Representational Theory of Concepts. Compositionality is the best explanation we have for the observed systematicity and productivity of thoughts and the logical character of much of the processing of thoughts. For example, by knowing what it is for something to be *talking* and knowing what it is for something to be an *apple*, we are able to know, in a very straight-forward way, what it is for something to be a *talking apple*. It is, simply, for it to be talking and also be an apple.

The most natural, albeit flawed, proposal is that such compositionality continues down through the lexical concepts TALKING and APPLE, into some still more primitive representations. As we just saw in 10.1.1, however the composition can't be quite the same, since TALKING APPLE is *defined* by TALKING and APPLE, whereas TALKING and APPLE don not so seem to have definitions.

The non-definitional versions of Lexical Concept Empiricism struggle, however, to account for the way in which TALKING and APPLE compose. TALKING and APPLE

can't be sets of prototype representations, or exemplars, or theories, because none of these representations compose to produce the concept TALKING APPLE. That is, for compositionality to work, TALKING would have to be a complex representation for the probabilities over the kinds of properties are found in its instances, APPLE would have to be such a complex representation, and when they compose they would have to be a complex representation built from those two sets of complex representations. The trouble is, TALKING APPLE does not have this complex-of-complexes. To be a talking apple is, simply, to be one of *those* that is also one of *those*. The probabilities might play a central role in *grounding* the simple representation APPLE, but going further to say that those probabilities *are part of* the representation APPLE brings unneeded trouble.

Of course, according to probability theory, of any entity, the probability of it being talking multiplied by the probability of it being an apple is equal to the probability of its being something that is both talking and an apple:

(i) prob(A) * prob(B) = prob(A and B)

Probability theory, however, makes assumptions that cannot transfer to a theory of concepts. In particular, probability theory assumes that the two probabilities being combined are *independent*.[30] That is, the theory assumes that the probabilities don't interact with one another, in such a way that, for example, something's being furry lowers the probability of its being an apple. Lexical concepts don't in fact come with the assumption that their probabilities are independent of one another.

Some researchers (for example, Smith and Medin, 1981) have gone to pains to find ways to compose the probabilistic sets of representations, but these attempt often

30 This observation is due to John Pollock.

yield messy theories with epicycles upon epicycles of processing and arbitrary-seeming exceptions. Composing concepts like TALKING and APPLE is something we should be able to do simply from having TALKING and having APPLE. There is no a priori way, however, to compute from (ii) to (iii):

(ii) prob(x is Talking | x is Humanoid) , prob(x is Apple| x is Red)

(iii) prob(x is Talking and x is Apple | x is Humanoid and x is Red)

That is, there is no a priori way to know the probability of something's being a talking apple given that it has a humanoid shape and is red, even if we know the probability of something's being a talking thing given that it is humanoid and the probability of something's being an apple given that it is red.

Another way of making this point is simply to notice that prototypical talking apples, if there are such things, certainly are neither prototypical apples nor prototypical talking things. Indeed, we need to know nothing about talking apples in order to have the concept TALKING APPLE, other than that they are apples and they also talk. What composes is what the prototypes are prototypes *of*, not the prototypes themselves. Theories on which APPLE and TALKING are representationally simple, in spite of having stored associated prototypes, are able to explain the straight-forward composition of the two concepts. Now, conjunction is only one among many ways in which concepts may compose, but a theory of concepts should at the very least be able to handle these simple conjunctive examples.

10.1.3 Lack of Conceptual Priorities

Many of the conceptual priorities that are predicted by the claim that lexical concepts are complex seem highly unlikely. Often the examples of complexity come from intuitions about necessary relations between concepts. Consider, for example, the (apparent) necessity that all cats are animals. Defenders of lexical concepts as complex explain this intuitive necessity by claiming that the concept ANIMAL is a representational part of the concept CAT. If ANIMAL is indeed part of CAT, however, then we should expect children to have the concept ANIMAL before they can have the concept CAT. This seems unlikely.

Even more unlikely is the complexity explanation for the more super-ordinate necessary classifications. Consider the (equally apparent) necessity that all cats are mammals. However likely it is that children have ANIMAL before they have CAT, it seems *extremely* unlikely that they have MAMMAL before CAT. In a similar vein, there doesn't seem to be any *particular* set of representations that are required in order to possess any given concept. It seems a blind child may have the concept CAT without ever having seen a cat and a deaf child may have the concept without ever having heard a cat, as long as they have had *some* perceptual contact with cats and noticed that they form a certain clustering of similar features. Lexical concepts being representationally simple explains the lack of conceptual priorities that their being complex in this way predicts.

It should be noted that accounts that treat lexical concepts as complex and built only from *perceptual* representations are even more threatened by the lack of predicted conceptual priority. Some such views claim that a concept that means *apple* has to be built from *certain kinds of* perceptual representations. Such views cannot explain how a blind person might be able to have a concept that means *apple*. Such views also ignore

observations about concepts like ELM, which Putnam used to show (1975) that we can have concepts for kinds without having had *any* perceptual contact with their instances.

10.1.4 Error / Misrepresentation

One of the major motivations for the claim that lexical concepts are representationally complex is to explain the (rationality of the) inferences made during the recognition of entities as falling under a certain concept on the basis of other representations that apply to the entity, and during the prediction of representations that will apply to an entity on the basis of the entity's being of the kind represented by the concept.

One problem that comes along with the complexity explanation for these recognition and prediction inferences is that it doesn't leave room for the *errors* that are observed in concept acquisition as well as in those recognition and prediction inferences. There are two sorts of error that need to be explained. First, there has to be room for error in the recognition and prediction inferences. We sometimes are wrong in thinking that something is an apple on the basis of its appearance, and we are sometimes wrong in our expectations about how a given apple will appear. The second kind of error that needs to be explained is the error in forming the concept in the first place. A lot of the time the similarities we observe in the world really are evidence of a similar underlying cause, but some of the time we learn that those observed similarities are merely superficial. This was discussed somewhat in Part II, in the motivations for the BMC. The concepts ARYAN, and RACE, have been shown to lack underlying bases, so we were mistaken in carving the world by them. Likewise, we coined JADE to represent the underlying cause of the similar greenish appearances of a set of gemstones, when in fact there was no such unifying similar cause. What is needed is a theory of concepts that

allows *both* for the rationality of the acquisition and inferences without making them trivial. Concepts carry contingent information about the world.

If it is part of the meaning of the concept APPLE that for an entity to be an apple it has to be (probably) red, and (probably) round, and (probably) fist-sized, *and so on*, then agents that have APPLE can never mis-classify an entity as an apple on the basis of its satisfy those constitutive representations. Error in representation cannot be ignored by a theory of concepts, since error is a phenomenon that is essential to something's being a representation.

It is important to note that theories on which lexical concepts are simple but acquired brute-causally also have trouble handling error/mis-representation. This problem will be discussed in section 1 of chapter 11, where I consider arguments that favor the rational/inferential acquisition of lexical concepts over their brute-causal acquisition.

The BMC, however, not only allows for errors, but it explains and predicts them. With limited knowledge at any time, entities that appear to cluster together may turn out to be quite divergent, as discovered when more information is gained. Again, this seems to have happened in the cases of ARYAN, RACE, and JADE. Our initial carvings at these purported kinds were rational, yet mistaken and revisable. Likewise, we might see an entity that appears very nearby a cluster we explained with APPLE, and therefore be very reasonable in categorizing the entity as having the same property kind that the apples have, and be entirely mistaken (which we may realize only upon lifting the object). Errors in prediction are similarly explained. Upon being told that there is an apple in the next room, we might use our stored associated beliefs about how apples tend

to look, and predict that it will be red and shiny, but find only an object that is brown and shriveled.

10.1.5 Concepts without Knowledge

Theories on which lexical concepts are representationally complex often assume that concepts are composed by representations in such a way that the concept encodes the agent's beliefs about entities to which the concept applies. It is my suspicion that this assumption comes from a conflation of 'S has concept C that means M' and 'S has conception C of M' (see Part I, chapter 3, where I list my uses of some terms). Recall from chapter 7 in Part II, that theories that take concepts to be composed by non-definitional beliefs take the meaning of the concept to be, roughly, the entity kind of which most of, or some weighted sum of, beliefs $b_1 \ldots b_n$ are true, where beliefs $b_1 \ldots b_n$ are (all or some of) the beliefs the agent holds about the entities to which the concept applies.

Hillary Putnam (1975) and Tyler Burge (1979) have argued persuasively that it is possible, and indeed common, for people to have concepts for kinds even though they have almost no knowledge about entities of the kind. Putnam points out that a speaker can use the words 'elm' and 'beech' to mean elm and beech (the two types of tree) without having any beliefs that distinguishes the two kinds of trees. Plausibly, then, that same human can have two mental terms, ELM and BEECH, and know nothing more than that the property-kind elm is called 'elm' in English and the property-kind beech is called 'beech'. A speaker may meaningfully ask of a particular tree 'Is that an elm?' The speaker would be thinking about the property-kind elm. Of course, we might hesitate to say that the person *knows what an elm is* or that she *has a conception of elms*, or maybe

even that she has *the* concept ELM, but these are importantly different from whether she *has a concept that means elm* (see the terms and distinctions in Part I, chapter 3). Similarly, Burge has shown that it is possible to have a concept even if most of the agent's beliefs involving the concept are false. He uses the example of ARTHRITIS, a concept that a patient may talk about meaningfully with his or her doctor under the false belief that it is a disease that can affect the thigh. Again, this person may not be in concept-related states in the vicinity of concept possession, but she does indeed have a mental entity that means arthritis.

Theories on which lexical concepts are representationally simple easily explain this wedge between concept possession and knowledge. As Putnam and Burge have suggested, all that is needed for the possession of the concepts in these cases is that there is a certain kind of causal chain, by which the agents in question can refer to the kind. (See DeVault, Oved, and Stone, 2006).

10.1.6 Circularity of Concepts as Beliefs

On the Representational Theory of Thought, under which our discussion about concepts begins, treats thoughts as complex propositional representations, many of which are partially constituted by lexical concepts, representations for lexicalized kinds. For example, the thought FIRE IS HOT is a complex representation that has the concepts FIRE and HOT as constituent parts. Thoughts that are believed, then, are simply propositional representations that are not only entertained by the agent, but are also endorsed.

Theories that treat lexical concepts as composed by beliefs involving the concept face a circularity problem. If beliefs are attitudes toward propositional mental representations, and some propositional mental representations are composed by lexical concepts, it is circular to then claim that lexical concepts are composed by those beliefs. It is circular, that is, *unless the beliefs are definitional*.

Defenders of the claim that lexical concepts are complex and composed from non-definitional beliefs may deny that their claim is that concepts are sets of beliefs. Instead, they may insist that their claim is that concepts are sets of *concepts*. The concept FIRE, for example, is composed in part by HOT. It is by virtue of that composition, goes the defense, that it is the case that we endorse the thought FIRE IS HOT. That is, one's concept FIRE is an encoding of one's beliefs about fire.

The trouble is, that defense begs our primary question about concepts. By virtue of what is the set of beliefs about *fire*? The answer to that question cannot be, without circularity, that the beliefs involve the concept FIRE.

10.1.7 Shared Meanings vs. Shared Beliefs (Publicity)

Different people can have radically different beliefs about one and the same kind. This is apparent from the human phenomenon of teaching and sharing information about the world. In order to do so, of course, people have to be able to think about the same kind. For example, human beings 3000 years ago had radically different beliefs about celestial objects from people today. The only belief in common between these two groups of people is that they were beliefs about *those* things, which we point to in the sky. Perhaps people 3000 years ago believed that those entities were small holes in the sky, and that they were fixed to spheres that move at various speeds around Earth. Most adult humans

today have almost entirely different beliefs. We have learned through other human beings that those entities, the very same ones, are huge and far away, that they are not holes but are made of matter, that some of them are on fire and that others merely reflect light from the fire of others, and so on. The observations made by Putnam and Burge in the previous sub-section (10.1.6) also suggest that agents with different beliefs may have concepts for the same kinds, as some people are experts, for example, on elm trees or arthritis, knowing much more than most other people about the kinds, and therefore being much more capable of distinguishing them from other kinds.

Theories on which concepts are (most, a weighted some of) the set of beliefs an agent has about a kind fail to explain how human beings with radically different, even radically incompatible, beliefs can have concepts for the same kind. This is easily handled by theories on which lexical concepts are representationally simple, and in particular by the BMC theory, which explains how it is that such radically different beliefs can manage to refer to the same kind. For the BMC, different reference-fixing descriptions can result in the acquisition of concepts with the same meaning. Beliefs about a kind are extremely relevant to coming to have a concept that represents the kind, but different sets of beliefs can play this concept-acquisition role.

10.1.8 Mental Twin Earth Cases (Putnam's WATER/TWATER)

Recall Putnam's (1975) Twin Earth case as discussed in chapter 9 of this dissertation. Putnam asks us to imagine a world that is almost exactly like our own, except that instead of H_2O, the watery substance in that world has some other chemical structure, XYZ. In that twin world, when English speakers use the word 'water' they are referring to the

XYZ substance instead of H_2O. Recall that there is no internal mental difference between a person in our world and a person in this twin world. Putnam used this observation to argue that the meanings of our linguistic terms are not inside of our minds; their meanings are, at least in part, determined by the external world in which we happen to speak. At least, there is *some* semantic feature of the terms that is fully external.

The case translates directly to the mental version, in which a person in our world and a person in the twin world have identical internal mental entities but with different referents. In spite of having exactly the same set of beliefs and associating them with exactly the same internal mental term, the semantic content of my term is different from that of my Twin-Earth-ling. Both of our mental terms pick out *that kind of stuff*, whatever the stuff is in each of our separate environments, that we have locked onto with the terms. Likewise, our linguistic terms, 'water' and 'H_2O', pick out *the kind of stuff* they are linked to, in spite of our associating very different beliefs with the two terms.

The BMC handles such cases with ease. The baptized meaning of a concept is the property kind, whatever it is, that satisfies the reference-fixing description used in its acquisition. The *epistemic* possibility that water is not H_2O simply reflects the wedge between representations and their meanings, a wedge that is part of the nature of representation.

10.1.9 Processing isn't faster for MARRIED than for BACHELOR

If some lexical concepts are composed from others, then the former should take more cognitive effort, more processing time, than the latter. For example, if BACHELOR is composed by UNMARRIED and MAN, then it should take longer to process BACHELOR than MARRIED. However, Fodor, Fodor, and Garrett (1975) showed that

sentences involving the word ‘married’ were processed no faster than sentences involving the word ‘bachelor’. Moreover, they found that UNMARRIED took longer to process than BACHELOR, reflecting the surface representational complexity of UNMARRIED.

10.2 Problems with Arguments Against Simplicity

This section considers arguments that have been made against the representational simplicity claim that is part of the BMC. Such arguments can be regarded as reasons to think that lexical concepts have other representations as constitutive parts. I step through several major such arguments to show why they fail to threaten the simplicity claim of the BMC framework.

10.2.1 (Rationality of) Recognition and Prediction

In our everyday lives, we interact with the world through perception and action. On the assumption that many of these interactions are mediated by concepts, symbols for categories, it seems reasonable to suppose that the mediation involves *inference.* From the way a given entity appears perceptually and the results of our motor behaviors upon it, it seems we infer to the entity’s belonging to a particular category, represented by a particular mental symbol. I call this *Recognition*. Likewise, we seem to infer from an entity’s belonging to a particular category to the expectation that it will appear some way perceptually and that certain actions upon the entity will have certain effects on it. *I call this Prediction*. For example, a given object may be recognized as belonging to the category represented with mental symbol APPLE on the basis of the object’s reddish, roundish, shiny appearance, and the sweetish taste that results from biting into it.

Likewise, an object's being represented as belonging to a category represented with APPLE may be predicted to appear that way perceptually and have that effect upon being bitten into.

The most common explanations of these inferences involve variations on the view that lexical concepts themselves are representationally complex, built up from those perceptual and motor representations involved in their recognition and prediction inferences. If the concept APPLE just *is* the complex (perhaps weighted) conjunction of representations for perceptible and motor properties used in their inferences, then the inferences are simply a matter of logic (albeit defeasible logic for non-definitional complexity).

The recognition and prediction observations are just as easily explained by views that treat lexical concepts as representationally simple. Inference relations can be encoded between simple lexical symbols and perceptual and motor representations without those relations being taken as making up the representational structure of the symbols. Views on which lexical concepts are complex have to either endorse Holism or else allow that there are many inference relations that don't make up the representational structure of the concepts themselves. Unless there is a principled distinction between inference relations that make up a given lexical concept, parsimony seems to suggest either Holism or representational simplicity.

The Lexical Concept Nativist takes the view that lexical concepts are among the representational primitives, and that there are inference relations *between* the simple lexical concepts but not because those inference relations make up the concepts' representational structures. The *perceptual and motor* representations, however, do not

trigger the lexical concepts *by inference*; they are at best merely part of the brute-causal chain that results in the triggering of the concept (Fodor 2008). It is not clear whether this position denies that we have such inferential relations, but if there are such relations it is hard to see why we would need them.

On the BMC framework, the perceptual and motor representations are indeed inferentially connected to the lexical concept, but not by being part of the representational structure of the concept. The concept directly refers to a property kind; the perceptual and motor representations are treated as contingent *symptoms* of the property.

Moreover, the recognitional and predictive inferences we make between lexical concepts and perceptual and motor representations seem to carry a kind of *normative property* --they seem to be *rational* or *justified* in some sense. There are very many theories of rationality and justified inference in philosophy and cognitive science, and this is not the place to canvass them. The observation here is simply that the category judgments that are made on the basis of perceptual and motor representations, and the perceptual and motor representations expected on the basis of category judgments, seem to operate on representations in a way that is appropriately sensitive to their relationships in *meaning*.

Theories on which lexical concepts are complex, in the sense that they have their inferential relations to perceptual and motor representations *by virtue of meaning*, seem to explain the observed rationality of the inferences in a very powerful way. On these theories, part of what the symbol APPLE *means* is (necessarily or probably) reddish, and roundish, and sweetish if bitten, etc. The inference relations are encoded directly on the syntax of the symbols. The trouble is, explaining the rationality of these inferences with

representational complexity falsely predicts a higher degree of rationality than is observed. That is, the explanation is mistaken in treating the inferentially related properties as part of the meaning of the symbol.

On the other hand, the view that the concepts are simple, it might seem, cannot explain the apparent rationality of the inferences at all. If there is no representational structure to concept, it isn't clear how computational operations, which are sensitive only to syntax, can be sensitive to the meanings of the symbols. The Lexical Concept Nativist, taking the concepts to be innate, has to take the apparent rationality as an illusion. For innate concepts, even if perceptual and motor representations appear in the chain that leads to the triggering of the lexical concept, they do not do so as parts of inferences; they are at best only brute-causally related. Certainly, on that view, the representations do not appear as parts of inferences that are appropriately sensitive to their meanings.

My alternative model, the BMC, strikes the right balance. The lexical concept is Simple in representational structure, while having inferential recognition and prediction relationships encoded between it and perceptual and motor representations. The rationality of the inference relations is not a matter of the meanings of the symbols. Instead, it is explained by similar representations having been used in the Baptism of the concept. On the BMC, the agent makes a reasonable default inference from something's looking like an apple to its being an apple, while maintaining the informativeness of that connection and the retraction of the inference upon future evidence.

10.2.2 Typicality Effects

Prototype Theory (as discussed in Part III) is one of the most popular accounts on which lexical concepts are composed by non-definitional, probabilistic inferential relations to other representations. Rosch (1973) and her followers defend this representational complexity view by appeal to psychological experiments that reveal *typicality effects* in accuracy and speed of reactions in classification tasks. The found that items that are more typical of a category, or at least ones that are believed to be, are more easily classified as belonging to the category. Table 10.1 below lists the categories that were used in some of these experiments, and their more typical (central) members versus their less typical (peripheral) members.

Categories and members used in reaction time experiment. (From Rosch 1973.)

	Member	
Category	Central	Peripheral
Toy	Doll	Skates
	Ball	Swing
Bird	Robin	Chicken
	Sparrow	Duck
Fruit	Pear	Strawberry
	Banana	Prune
Sickness	Cancer	Rheumatism
	Measles	Rickets
Relative	Aunt	Wife
	Uncle	Daughter
Metal	Copper	Magnesium
	Aluminum	Platinum
Crime	Rape	Treason
	Robbery	Fraud
Sport	Baseball	Fishing
	Basketball	Diving
Vehicle	Car	Tank
	Bus	Carriage
Science	Chemistry	Medicine
	Physics	Engineering
Vegetable	Carrot	Onion
	Spinach	Mushroom
Part of the body	Arm	Lips
	Leg	Skin

Table 10.1

These typicality effects have been used by Prototype Theorists as a reason to think that concepts are stored probabilistic sets of representations, where the more typical features

of the category represented are stored as more probable and the less typical features are stored as more probable.

Other psychologists, however, (Armstrong, Gleitman, and Gleitman, 1983) have undercut these typicality effects as a good reason for thinking that a concept is a stored probabilistic set of representations. Their experiments showed that typicality effects can occur for concepts that clearly are not stored probability representations. They used clearly definitional examples of number categories, like PRIME, ODD, and EVEN. These concepts have clear necessary and sufficient conditions, and for any number, each concept either applies or does not apply. The experiments, nonetheless, showed typicality effects, much like those that were used to defend prototype theory. For example, they found that '3' and '5' were more easily classified under PRIME and ODD than '9', and that '2' and '4' were more easily classified under EVEN than '8'.

Such experiments show that concepts can have associated representations without those representations making up the representational structure of the concept itself. Theories on which lexical concepts are representationally simple, therefore, are not threatened by typicality effects.

10.2.3 Necessity Intuitions can't be explained by Complexity

Recall the necessity intuitions that were discussed in section 10.1.3., such as the (apparent) necessity that all cats are animals. Defenders of lexical concepts as complex explain such intuitive necessities by claiming that, for example, the concept ANIMAL is a representational part of the concept CAT. Property kinds that are necessary for a concept's application are often used to argue for the concept's representational

complexity. The argument is that the only way to explain the necessary relationships between representations is if the concepts are partly constituted by the representations for their necessary features. For example, being fruit is necessary for being an apple, goes the explanation, because being a fruit is part of being an apple, and the only way to explain *that* is if the concept APPLE has the concept FRUIT as a representational part.

Recall also from that section that many of the conceptual priorities that are predicted by this explanation of necessity intuitions are unlikely, such as MAMMAL being part of CAT. Not only do the lack of conceptual priorities suggest that the lexical concepts are simple, they also undercut the arguments from necessity to complexity.

Besides the cases of necessity that involve superordinate categories, there are plenty of cases of concepts that have necessary property kinds for their application cannot be part of the concept because the concept is often possessed before the discovery of its *underlying* necessary property kinds. It is necessary for something to be water that it is H_2O, but people were long able to entertain the concept WATER before the discovery of the necessity. Other examples of necessities that were discovered *a posteriori* include Atomic number 79 being necessary for being gold, electrical discharge being necessary for being lightning, cells being necessary for being a tree, and gills being necessary for being a fish.

Many philosophers, of course, deny that there are a posteriori necessities, or else they make a distinction between epistemic necessity and metaphysical necessity (see 2-Dimensional Semantics in Part III). The BMC explains away such analyticity intuitions with clusters of clusters. See Figure 5.1. Clusters of clusters (super-clusters) are plausibly acquired after clusters, rather than before. This needn't always be the case;

sub-clusters may be found with later information. In any case, the necessity relations are discovered empirically rather than being part of the concept's meaning.

10.2.4 Co-referring Mental Terms (Mental Frege Cases)

Consider the following variant on Frege's puzzle about identity statements.[31] Sammy is a young child who has two mental names that refer to the property/kind *coyote*. One of these mental names, F, was acquired through his visual perceptions of coyotes (suppose he saw some of these skinny-legged wolf-like animals at a zoo). The other mental name, G, was acquired through Sammy's auditory perceptions of coyotes (suppose he heard some screeching howls through his window at night). If Sammy were to eventually learn that Fs and Gs were the same kind of animal, and form the thought 'Fs are Gs', he would presumably have learned something new. The thought 'Fs are Gs' is informative in a way that the thought 'Fs are Fs' is not.

Since the external referents of F and G are the same, it seems like the only way to account for the difference in information is that the concepts are representationally complex and built up from two different sets of perceptual representations. That is, we can have two lexical representations that have the same referent but carry different information, so the representation must be complex. It has to have representational parts from which to obtain its meaning.

This might be a good argument if you want meanings to be typed to explain folk concept ascriptions. But these appeal to syntax as well as meanings. There are two major problems with this argument. First is that the cases of co-referring terms doesn't work in the case of mental terms as well as it does in the case of linguistic terms. This is because

31 Frege's Puzzle is introduced in his (1892).

the mental terms are not *public*, and there is no clear way to type mental terms by their inferential relations. There are unimaginably many ways for the concept to be acquired. Sammy acquired a concept for coyote through certain auditory experiences and he acquired another concept for coyote with certain visual experiences. These experiences could have been very different, and there are indefinitely many other possible perceptual links to the kind. There is no sense to be made of the question of *which concept, F or G, is <u>the</u> concept COYOTE?*. No theory that attempts to give lexical concepts a complex meaning (even if that is only one part of the term's semantic features, as in Two-Dimensional Semantics) will be able to find a clean way of handling the wide variety of beliefs associated concepts that pick out a kind.

The second problem with the co-referring terms argument is that we can explain this with an account that appeals only to concepts and their referents, and we can do so easily when we notice that we are taking concepts to be symbols, in the minimal sense that they are the things in our minds that are about kinds. This distinction between concepts and meanings leaves room to appeal to a difference in syntax, the mere fact that there are two distinct token symbols, F and G, to explain the informativeness of the thought 'Fs are Gs'. In other words, agents can have two concepts with the same meaning without realizing it, because the two concepts are stored as different syntactic entities. The informativeness of identities involving co-referring concepts has to do with the difference in syntax or the representations, not the difference in semantics, so the observation is explained by lexical concepts being simple at least as well as by their being complex.

Moreover, with The BMC framework as a clear alternative, we can deal with the informativeness in a way that is superior to the complexity approach. While under the Representational Theory of Concepts, we can turn to the syntactic difference in the simple mental terms –one term is F and one is G. There were two different descriptions used in baptizing the terms, and the agent had no reason initially to think that the descriptions picked out the same property.

Notice that The BMC framework's syntactic explanation for the informativeness is also superior to the view that lexical concepts are simple and *brute-causally acquired*. Although the difference in syntax would explain the intuition that identities involving co-referential concepts are informative, it is not at all clear *how or why* an agent would have two Innate mental representations with the same referent. There doesn't seem to be much room, on the Nativist's account, for an agent to come to have two different mental names for the same property. The BMC framework model, however, does make sense of there being two mental names. Even though the meanings of both names are the same external natural kind, the mental symbols come to have that meaning via very different reference-fixing descriptions. The two pathways of acquisition lead to the baptism of two different mental symbols.

10.2.5 Can't Type Concepts by Meanings

One might complain that the arguments for the BMC suggest that lexical concepts can't be typed by meanings, which makes a semantic psychology impossible. This is because the only semantic feature a lexical concept has, on the BMC, is its external referent. The success of Folk Psychology, however, which is clearly (at least partially) semantic, seems

to suggest otherwise. Meanings play a role in folk psychological explanations of behavior, so one might infer that they must play a causal role in producing the behavior, and they must therefore be (at least partially) internal to the agent.

The BMC does indeed claim that the only semantic feature of a lexical concept is its external referent. Meanings do play a role in explaining behavior, but it doesn't follow that the meanings themselves *cause* those behaviors. It is because we are on Earth that it is H_2O and not XYZ that we want to drink, even though we and our twins have identical internal entities, WATER, to represent the substance. Moreover, I suspect that we try to convey syntactic information with our folk ascriptions of thought, at least syntactic distinctions that do indeed play a causal role in producing an subject's behavior. Perhaps folk ascriptions of thoughts are similar to reports of what somebody *said*. When we report what was said by a subject, we sometimes convey the meanings of what was said and we sometimes convey the syntax by inserting quotation marks. Contrast, for example, 'Peter said he likes H_2O' and 'Peter said "I like water"'.[32] Of course, since mental syntax is private, the best we can do is try to use language in a way that gestures our listeners towards syntactic distinctions that the subject may have made. The success of Folk Psychology can be explained by conventions in a shared language, along with shared beliefs between interlocutors.

10.2.6 Rigid Descriptions / Actualism

Another objection that might be raised to the claim that lexical concepts are representationally simple is that the rigidity observations on which it is based can be captured by complex representations. Perhaps all that is needed for the rigidity of a

32 This observation is due to Ken Shan.

lexical concept is to insert the term ‘actual’ into the descriptions used in their acquisition. For example, just as in the BMC process, when acquiring APPLE an agent may observe the similarity in reddish, roundish, and sweetish appearance, and form a description that picks out the underlying property kind that the objects share. But instead of then assigning a simple name, fix the description itself to the kind and let the description be the concept. The rigid description might look like (RD):

(RD) The actual property kind that these objects share that explains this actual observed similarity in appearance.

This description will pick out the same property kind when tokenized at later times in the agent’s life. It will always point back to appleness, if in fact that is the property kind that satisfied the description.

This approach is the best I’ve considered for maintaining the representational complexity of lexical concepts. The trouble is, having that description can’t be what it is to have APPLE. Different agents, and even the same agent in different contexts, may use very experiences, and thus very different descriptions, in coming to acquire APPLE. Moreover, in spite of the rigidity of the description, it doesn’t capture the right modal intuitions. Appleness might not have been the actual property kind that explained that similarity. There is no way for RD to capture this, since the syntax of the description is fixed. That is, it is not the case that *the property kind that actually explains that similarity* might not have been *the property kind that actually explains that similarity*, just as it isn’t the case that *x* might not have been *x*.

There may yet be hope for this Actualist approach to saving the representational complexity of lexical concepts. I suspect, however, that it will be difficult and messy. Once the rigidity of lexical concepts has been conceded, I see no reason to reject the assignment of a simple mental term to the property kind. Again, all of the information about the kind that was involved in the acquisition of the concept can be stored in association with the concept without thereby becoming part of the concept itself.

11 Why Lexical Concepts have to be Rationally Acquired

11.1 Observations Best Explained by Rational Acquisition

In this section I step through a list of observations and show that a view like the BMC, on which lexical concepts are acquired by an inferential, rational process, offers better explanations than its competitors.

11.1.1 Adaptibility

Another observation behind the Rational Acquisition claim is that human beings are highly adaptable to various environments. Any given human could have been born into a world with very different properties, and organisms, and man-made contraptions, from the world they were in fact born into; the aardvarks and pizzas and carburetors that surround us may well have been replaced with other kinds of objects. Furthermore, among the kinds of things that in fact appear in a human's environment, there are many that would never be thought about, as our ancestors 200 years ago didn't think about electrons or ultra violet light. Given that humans only think about a small subset of all of the possible properties and kinds, *it would be wasteful to encode innately concepts for all of the possible ones*. Given that we need to be adaptable to various environments, *we need to be able to come to encode concepts for any of the possible properties we may encounter*. This pair of observations suggests that we encode innately some primitive building blocks from which we can learn through experience which properties surround us. This is not a point about evolution; it's a point about good design. In trying to build an artificial cognitive agent, this latter design would be preferred over trying either to build in all possible concepts or build in some fixed and un-flexible subset of them.

11.1.2 Error/Mis-representation

The BMC allows for two sorts of error that need to be explained by a theory of concepts, which are difficult to explain by brute-causal theories. The first sort of error occurs when a concept to be mistakenly applied to an entity. The second sort occurs when a conceptual carving of the world is mistaken, as evidenced by the observation that we revise our conceptual carvings of the world as we experience more and more features on which observed entities cluster.

Let us start with the first sort of error. A generally accepted observation about the application of concepts is that human beings are sometimes are wrong. This can occur in several different ways. A concept that represents the property-kind horse may be applied to an object seen in the distance, when in fact the object is a cow. Or else, when observing that an entity appears near a cluster of features, we might mistakenly infer that it is of the same property-kind as the previously clustered entities. We saw in section 10.1.4 that this first kind of error is problematic for theories on which concepts like HORSE and COW are representationally complex.

Theories on which these concepts are either innate or else brute-causally acquired also struggle to make sense of this first kind of error. Fodor's (1987) Asymmetric Dependence view is an observation about the relationship between a concept and its referent. It does not *explain* the relationship. One might say that the BMC framework just pushes the same problem to the perceptual and other innate representations it posits. But the BMC framework leaves far less un-explained. The problem of error for

perceptual representations seems different from that of concepts, since it is not clear what it is to mis-represent perceptually. These matters are worth exploring, and the BMC framework turns attention to them.

The second sort of error that is observed is the errors we make in carving the world into kinds. This was discussed somewhat in Part II, in the motivations for the BMC. The concepts ARYAN, and RACE, have been shown to lack underlying bases, so we were mistaken in carving the world by them. Likewise, we coined JADE to represent the underlying cause of the similar greenish appearances of a set of gemstones, when in fact there was no such unifying similar cause. The BMC handles these observations by positing concepts that are representationally simple without being acquired brute-causally from entities that fall under the purported kind.

11.1.3 Fissioned Meanings

Consider again the concept JADE. This concept was acquired under the assumption that a set of samples with similar greenish appearance shared a property-kind that was responsible for their observed similarities. It might be said that people *intended* to refer to a unique existing property-kind with JADE, but in fact they failed to do so. Upon the discovery that there are two underlying mineral structures in the samples, we come to realize that we had a mistaken carving of the world.

11.1.4 There is Room for Inference between Perception and Concept

There are various stages in which the agent can be said to play an active role in the processing between perception and concept acquisition, and subsequently in the processing between perception and conceptual belief. See the Appendix to this dissertation for a detailed discussion.

11.2 Problems with Arguments against Rational Acquisition

This section considers arguments that might seem to be reasons to reject the claim that lexical concepts are rationally acquired. Such arguments, if sound, would support accounts on which lexical concepts are either innate or acquired by a merely brute-causal process. I step through these arguments to show that the BMC, on which lexical concepts are acquired by an inferential, rational, process, is the superior view.

11.2.1 Circularity of Hypotheses about Meanings of Concepts

Fodor raises a circularity worry for the inferential acquisition of lexical concepts. It is stated the most clearly in his (2008). It is circular to propose that we learn concepts in the sense that we formulate mental hypotheses like 'concept x means y', because the mental hypothesis, X MEANS Y is itself representational, containing Y, a representation for the meaning presumably being learned. For example, in order to acquire a concept x that means *apple*, the agent would have to entertain the representation X MEANS APPLE. But what it is to acquire a concept x that means *apple* is to come to have APPLE. It is circular to propose that we come to have APPLE by a process that requires the possession of APPLE. (Some psychologists do seem to be trying to do something like this when they ask subjects to figure out the meaning of a linguistic term, "blik". That,

however, is not what it is to rationally acquire a concept as we have characterized it. There is a genuine question about whether lexical concepts, like APPLE, are acquired by a rational process that is akin to the acquisition of phrasal concepts, like TALKING APPLE).

Moreover, the composition of concepts is not the only rational/inferential process for concept acquisition. The BMC offers an alternate process, compatible with almost all of the observations about concepts.

11.2.2 Neither Rupert nor Margolis Meets Fodor's Challenge

Recall what I have been calling *Fodor's Challenge*.

> ***Fodor's Challenge*:** we must either (a) find a set of simple properties that necessarily applies to all and only the apples, and is plausibly the meaning of the concept APPLE, or (b) find a way for a simple concept that directly refers to the property of being an apple to be acquired not by brute-causation, but by an inferential process.

Robert Rupert, in his (2001) *Journal of Philosophy* article, offers an account of concept acquisition in aim of saving the representational simplicity of lexical concepts from an argument that was given by Robert Cummins (1997). (They call the thesis that lexical concepts are simple 'the Causal Theory' and the notion of inferential acquisition 'Theory Mediation'.) Cummins argues that lexical concepts ca not be representationally simple on the grounds that it is incompatible with their being acquired later in development. In response, Rupert replies by pointing out that lexeical concepts may be acquired later in

development in a way that *is compatible* with their representational simplicity. Rupert discusses the notions of 'acquired' that he takes Cummins to have cited as incompatible with representational simplicity. Although he does offer plausible compatibility arguments for the acquisition processes that Cummins considers, none of the processes are *inferential* acquisition processes, so the debate misses the point. Rupert says rather clearly, "Cummins seems to assume that if a concept is not acquired by cognitively described means (as the result of learning, in particular), then the concept is innate. We should reject this assumption" [pg 10]. He then goes on to show how representational simplicity is compatible with a concept that is acquired (i) as a result of brain development, (ii) a theory that is implicit in the functional architecture resulting in the right causal relation, and (iii) not knowing anything about the concept's extensions at birth.

These acquistion processes are not ones that will meet Fodor's Challenge, and they fail to introduce the normative aspects of our concept acquisition and use. The BMC framework, in contrast, explains why we are *justified* in our beliefs about the world (see Pollock and Oved, 2005). But the BMC framework allows for justification on many views of what justification involves.

Another concept acquisition approach is offered by Eric Margolis (1998), and is indeed later endorsed by Fodor himself (LOT2 2008). It was the hope of Margolis to appease the tensions between Lexical Concept Nativism and Lexical Concept Empiricism, and it was his hope to do so by describing a process for the acquisition of representationally simple concepts. The trouble is, just as with Rupert, the sense of acquisition that is described fails to satisfy the Lexical Concept Empiricist who insists

that concept acquisition is a rational/inferential process. The process he sketches yields an inferential connection between perceptual representations and conceptual beliefs, but the concept-acquisition process itself is a brute-causal one.

Fodor's challenge is a challenge because option (a) has been proven incredibly difficult if not impossible, and option (b) seems impossible on the face of it. The BMC, of course, pursues option (b), and illustrates its possibility as well as its plausibility.

11.2.3 Fast Word Learning and Fast Locking from Stereotypes

Next, consider an argument from the speed of word learning. Children seem to lock onto the right properties very quickly when explicitly being taught words for them. The perception of just one or two good examples of an apple, seems to be enough for children to lock onto the property of appleness. Such little experience doesn't seem to be enough for learning to occur, certainly not by the formation and testing of hypotheses, so children must be innately wired to detect the property-kind appleness. We lock onto the right properties with very little experience, goes the argument, too little experience for learning to occur, so the concepts must be innate.

It seems to take only a few good, stereotypical instances of a property kind for people to lock onto the kind. This seems to suggest that by the time a few good, stereotypical instances are perceived, there is a concept waiting to lock onto the property-kind the instances are stereotypical *of*. A few samples doesn't seem like nearly enough for the learning of a concept, which requires the formulation and (dis)confirmation of hypotheses about the kind.

On the BMC, a single instance can allow for the acquisition of a concept. An entity that appears on a part of the perceptual space that is far away from any of the other entities (again, where far enough is determined by the agent's similarity algorithm), may lead the agent to infer that there is an underlying property that the entity has that explains it appearance. The agent may then abstract on this property, considering that other entities might have that same property kind. This is enough to form a reference-fixing description from which to acquire a concept for that kind. What is needed for this to happen frequently is that the agent expects to be in a world in which there are some few kinds of entities, and the kinds consistently cause similar humanly perceptible features.

11.2.4 Un-learnability of Simples

The Building-Blocks framework does not allow for representationally simple concepts to be rationally/inferentially acquired. From the conclusion that Fodor draws about the simplicity of lexical concepts, he argues that they cannot be rationally acquired. He writes, "Lexical concepts are typically undefinable, hence typically unstructured, hence typically primitive, hence typically unlearned" (1981, pg. 298). Part of the idea here is that an agent can be designed to learn about the contingent regularities in its world, but it can only learn those contingencies in terms of the concepts it already has. For example, Fodor (1975) argues that learning involves the formation and testing of hypotheses, suggesting that there is no way for an agent to learn the meaning of a concept unless the agent already has it. Learning a concept would require the agent to have the mental terms by which to form hypotheses about the concept's meaning. For example, in order to form the hypothesis that 'the concept APPLE means *the property of being an apple*', the agent must have some way of representing *the property of being an apple*. But having some

way of representing *the property of being an apple* is exactly what it is to have the concept APPLE. Clearly it would be circular to require the agent to have the concept APPLE in order to acquire it. Indeed, there is not any other obvious way for concepts to be acquired by a *rational/inferential* process.

The BMC, however, turns out to be a process that allows the acquisition of APPLE from representations already in possession, where the representations already in possession allow the agent to represent the property-kind apple, in its context, with demonstrative perceptual representations, before having a concept that *always* means apple.

12 Why Lexical Concepts have to get their Meanings by Baptism

This chapter considers a third, and final controversial claim of the BMC. This is the claim that most lexical concepts are acquired by a baptism process. As in the two previous chapters, I have a section on the explanatory power of the BMC process and a section that responds to anticipated objections.

12.1 Observations Best Explained by Baptism

In this section I list observations that are better explained by the BMC framework than any of the alternative theories. Some of these observations are fleshed out here, and the others were already discussed in the motivations for the BMC in Chapter 4.

12.1.1 Baptism allows for Rationally Acquired Simples

On balance, the observations about concepts suggest directly either that lexical concepts are rationally/inferentially acquired or that they are representationally simple. The majority of the concepts debate has the Building-Blocks assumption in the background. This assumption allows only two possible kinds of representation –Primitive representations, which are simple and innate (or brute-causally acquired), and Composite representations, which are complex and acquired by composition of more primitive representations. As a result, observations suggesting the rational acquisition of lexical concepts were used to argue for their being Composite, and thus representationally complex. Likewise, observations suggesting that lexical concepts are representationally

simple were used to argue for their being Primitive, and thus not acquired by a rational/inferential process.

The BMC goes beyond building blocks, adding a third kind of representation – ones that are rationally acquired yet representationally simple. The baptism process for these concepts coming to have their meanings is the only known process that allows concepts to be rationally acquired and representationally simple.

12.1.2 Discovery of Diseases from Symptoms

See section 4.2 in Part II.

12.1.3 Inferences about the World are Rational and Informative

This observation was partially discussed in section 4.4 in Part II of the dissertation. Epistemologists have struggled, most explicitly after Descartes (1641), to explain how human beings can infer from their perceptual experiences of entities, like the visual sensation you might be having now, to conceptual beliefs about those entities, like the belief that there is a *cup* on a *table*.

The Cartesian project was based on the 17th Century (Descartes, Locke, Hume) notion of epistemic justification as requiring complete, air-tight indubitability. To establish a solid and stable science about the outside world that would enjoy the epistemic status that mathematics seems to have. The approach was to derive beliefs about the outside world by use of purely deductive reasoning, starting with a foundation of indubitable beliefs about our sensory experiences along with built-in first principles. While Cartesian skepticism loomed, however, the veil between the world and our

experiences held thick. Two approaches then emerged for eliminating the veil by reconsidering the situation.

One approach was to rethink the external world by, in effect, making it internal. The hope was to show that we can define all of our concepts about the external world in terms of our sensory experiences. Beliefs about apples, for example, would be understood as beliefs about, say, our reddish, roundish, and sweetish experiences; beliefs about dogs would be understood as beliefs about, if you will, our softish, furryish, and barkish experiences. Beliefs about apples and dogs could thereby enjoy the high epistemic status they would inherit from beliefs about sensory experiences. The success of the approach would, of course, rely on the possibility of indeed analyzing our external world concepts in terms of sensory experiences. Attempts to do so have been shown, however, to be impossible (see Quine 1951; Fodor 1987).

It seemed the only alternative for escaping epistemic skepticism was to rethink the notion of epistemic justification itself. On this track, the new approach begins with the claim that we are forced by many of the belief-producing faculties (perception, reasoning, memory, etc.) to endorse the beliefs that are produced by them. Not only are the beliefs produced automatically, goes the claim, but it happens *immediately*, so there is no *gap* in the processing at which we might willfully jump in and disrupt the automatic production of beliefs by these faculties. From this observation, then, it is argued that we are epistemically justified in endorsing beliefs that are so bound by our faculties. Authors who can (and will for present purposes) be interpreted as taking this approach include

Reid, Moore, Bealer, Alston, Goodman, Pollock, among others.[33] I call this approach *Cognitive Architecture Justification*.

The BMC offers a third alternative. The contingent a priori inferences (section 5.2) allow for the right balance between the rationality and informativeness of the inferences between perceptual representations and conceptual ones.

12.1.4 Naïve/Psychological Essentialism

See section 4.5 in Part II.

12.1.5 We Revise our Carvings

See the observations in section 4.5 in Part II.

12.1.6 How Linguistic Terms Get Their Meanings

See the observations in section 4.6 in Part II.

12.1.7 Concept Acquisition for Robots

See the observations in section 4.7 in Part II.

12.2 Problems with Arguments against Baptism

33 Many of these authors have, or might well be, argued not to take this approach.

In this subsection I consider several anticipated objections to the baptism process as a concept-acquisition process. For each objection, I am able either to show how the BMC explains the phenomena that are used in the objection or how it explains them away.

12.2.1 Easy Knowledge / Latitudinarianism

Using a reference-fixing description doesn't seem to give an agent enough information about a kind for it to be sufficient for concept acquisition. It seems to make concept acquisition too easy, because such descriptions can be formulated willy nilly. Not only that, but it is too easy of a way to come to know contingent information about a kind. Consider the following case to illustrate.[34]

(Case 1) Sally is playing a guessing game with her older sister, Janet. Janet asks Sally to guess what her favorite food is. Suppose that Sally uses a concept K to think about that kind of food, the kind that is Janet's favorite, or the kind that Janet is thinking of.

It seems appropriate in this case to say that K means apple. However, we are not at all comfortable saying that K allows Sally to think about appleness as such. We are also uncomfortable saying that K is the concept of appleness, or that K is a token of the APPLE concept, or that having K in this way is sufficient for Sally to know what apples are.

34 This objection was brought to my attention by Karen Lewis.

One way to deal with this objection is to point out that the BMC framework, as illustrated in this dissertation, uses only a special kind of reference-fixing description – the description involves perceptual demonstrative representations. These demonstratives serve to ground the concept with real instances of the kind that the concept represents.

However, even appealing to demonstrative perceptual representations doesn't help. Consider this second case:

(Case 2) Suppose that Sally is fed applesauce frequently, but she does not know that applesauce comes from apples. She uses another concept token N to think about the kind of food in front of her, the kind that makes this stuff look whitish/yellow and feel mushy and taste sweet.

It seems appropriate in Case 2 to say that N means apple, and that N allows Sally to think about appleness, even as such. However, in this case, we hesitate to say that N is the concept of appleness, or that N is a token of the APPLE concept, or that having N in this way is sufficient for Sally to know what apples are.

My preferred response is that intuitions about, and subsequent ascriptions of, concept possession are due to massive ambiguities. To address this objection, we need to recall distinctions between the following:

(i) having a concept C that means apple

(ii) having a concept C that allows us to think about apples *as such*

(iii) having a concept C that is *the* concept of an apple

(iv) having a concept C that is a token of *the concept APPLE*

(v) knowing what apples are

I claim that in Case 1 and Case 2 the intuitions that deny concept possession are in fact about (iii) or (iv) or (v), which have to do directly with whether Sally would be in agreement with the general population upon distinguishing apples from non-apples. This issue of agreement explicitly involves sharing *beliefs* with the general population. To use these intuitions is to beg the question of whether or not concepts are sets of beliefs.

(Case 3) Suppose that 2-yr-old Sally sees her older sister, Janet, biting into an apple, and uses a concept token M to think about that kind of food, the kind that makes those things looks reddish and roundish and make a crunching sound when bitten.

It seems reasonable to say that Sally's M means apple. She is able to think about appleness, indeed it seems she is able to think about appleness as such. It also seems appropriate to say that she has the concept of appleness, that M is a token of the APPLE concept, and that Sally knows what apples are.

This third case is clearly a case of having APPLE and the first case is clearly a case of not having APPLE, and this is perhaps the reason people turn to perceptual demonstratives as a necessary condition on having such concepts. But Case 2 presents problems with that answer. Case 3 seems so much better than Case 2, even though they both involve perceptual demonstratives. This suggests to me that we have three choices:

(1) Revise our intuitions about Case 2 so that it is a case of having the concept APPLE.

(2) Keep our intuitions fixed and appeal to something other than perceptual demonstratives to explain the superiority of Case 3 over Cases 1 and 2.

(3) Revise our intuitions about Case 3 and accept latitudinarianism.

With the distinctions in hand, and advantages illustrated, option (3) seems clearly to be the right choice. Reference-fixing via various perceptual pathways seem to allow for the acquisition of *a* mental term for lexical property-kinds. It remains to be explored, of course, whether *all* reference-fixing descriptions are sufficient connectors to meanings for concept acquisition.

12.2.2 We Know What we are Thinking About

The BMC framework has the strange result that agents don't know what they are thinking about. When baptizing a mental term, the agent picks out some property via a mental description, but the agent doesn't know which property is thereby picked out. How does the agent know, for example, that it is the property of being an *apple*, rather than, say, that of being a *pear*, that satisfies the reference-fixing description? This might be a reason to think that humans have to be innately connected to the properties that their concepts represent, for that is the only way to make sense of the intuitive access we have to the meanings of our representations.

This objection seems powerful at first, but on closer inspection we see that the observation is irrelevant. There is a phenomenology associated with many of our representations, what it's like, for example, to be thinking about appleness rather that

pearness. But this phenomenological difference should not be taken into considerations about the contents of our representations. Indeed, if we assume the Representational Theory of Mind, the question is almost impossible to frame. Notice that the only way for an agent to represent any property, according to the RTM, is to entertain its symbol for that property. Within this framework, it is not clear what it is for an agent to know which properties its own mental symbols represent, other than in terms of other symbols. The best an agent can do, 'from the inside', is use the symbols it has to think about the meanings of its symbols. There is no representation-independent access to the meanings of our concepts under the Representational Theory of Mind. Perhaps it feels meaningful from the inside because our symbols are introspectible, and they have different syntactic features. It is counter-intuitive, but the counter-intuition itself is explainable through the BMC Framework. The intuition that we know what we are thinking about in the sense that we have *direct* contact with meanings is explained away by the umbrella representationalist theory that we access meanings through representations.

12.2.3 Concepts like JADE and ARYAN Seem Meaningful

The concept JADE seems to be meaningful, but the BMC suggests that it doesn't really manage to refer (it doesn't satisfy the Existence Condition), and on the BMC, if it doesn't manage to refer it doesn't manage to have a meaning.

The BMC explains this seeming of meaningfulness away. Just as in cases of co-referring concepts (mental Frege Cases), the agent's cognition is sensitive only to meanings insofar as the agent's syntax corresponds to meanings. In the case of JADE, the agent *takes* the mental term to be meaningful, to pick out an existing and unique kind, that is the kind that explains the believed similarities in the samples. Even after

discovering that in fact there isn't such a kind that satisfies the description, we can keep the mental term JADE and use it as if it did pick out a kind, and it makes sense to do so in some contexts. Even in contexts of jewelry exchange, however, the term may lose relevance. For example, if the discovery of the lack of a kind that explains all of the superficial similarities, leads to the discovery that one of the underlying minerals is more rare, and therefore more valuable in the market, JADE maybe become a useless concept.

12.2.4 There is no Room for a Semantic Science of Cognition

Because the BMC has lexical concepts with meanings that are fully external to the agent, it is impossible to give laws about cognition in terms of the meanings of their lexical concepts. An agent may have two concepts with the same meaning (as in mental Frege Cases) while its rules of cognition treat them as having different meanings. Without a semantic science of cognition, how can the BMC explain the success of the semantic Folk Psychology?

In response, I offer that the semantic Folk Psychology is taken, at best, to offer *ceteris paribus* generalizations. However, while ceteris paribus generalizations are common in many highly respected sciences (laws of biology and gravity, for example), the sciences *explain* and *predict* the failures of their laws in the exception cases. Folk Psychology cannot explain the cases in which its laws fail. In contrast, the BMC framework offers very clear explanations of the successes and failures of semantic cognitive generalizations --in usual circumstances, human-beings have quite similar carvings of the world. This results from (i) beginning with similar perceptual apparatuses and generalization algorithms, (ii) a shared world that is mostly stable and therefore gives

similar experiences to human beings, and (iii) language allows the syntactic features of thoughts to be shared (words correspond to concepts) as well as the sharing of beliefs.

Part V: Concept Acquisition in Artificial Intelligence

13 Artificial Intelligence and the Representation-Making Relation

Recall Figure 3.1 from the set-up for our discussion of representation in Chapter 3. All cases of representation involve at least (1) an entity, *x,* (2) an entity, *y,* (3) a representation-making relation, *Rxy*, by virtue of which *x* is a symbol/representation and *y* is the meaning/content of *x, and* (4) a process, *P*, through which the state *Rxy* gets established. Most of this dissertation has focused on (4), and the Baptizing Meanings for Concepts framework offered here is an account of the process *P*. In this chapter, we explore some considerations for (3), the metaphysical relationship between two entities, *x* and *y*, when *x* is a lexical concept and *y* is its semantic content.

One way to formulate philosophical questions about conceptual representation is by asking what it would take for *any* entity to have lexical concepts. Artificial Intelligence (AI) is a field whose objective is to build artificial thinking entities, where the focus is on computational devices. So the questions arise, what would it take for a computer or robot to have concepts? What would count as inferential acquisition for a computer or robot? Let us explore through the AI lens some of the philosophical theories of what constitutes the possession of a concept with a given meaning.

The field of Artificial Intelligence struggles with the question of what constitutes an agent's *understanding* the world in terms of its representations. For example, Steven Harnad's (1994-present) work on symbol grounding gives reason to build agents with representations that are causally connected with their referents in the world. However, an agent that learns to discriminate faces from non-faces in visual stimuli, for example, lacks a genuine concept or understanding of the category. This becomes obvious when certain

kinds of mistakes are made, (such as categorizing a blob on the ceiling as a face) Some researchers try to find new features within the visual stimuli for agents to use to improve their classifiers, but the result is a bag of tricks for classifying correctly as deemed by the programmers, rather than a principled model for classification. Bringing observations and theories of meaning from philosophy and cognitive science will help to frame the goal of building computers that understand.

13.1 Functional Role Theories

These theories plausibly have their origin in Wittgenstein's (1953) idea that "meaning is use". The idea here is that the meaning of a mental representation is determined (and in fact, on some accounts, is *identical with*) the role it plays in a system of representations. The best examples to support this view are logical symbols, such as OR and AND where they might be said to mean what they do by the virtue of roles they play in inference.

However, the theory goes further to liken all mental representations to these logical symbols, claiming that all of an agents can be said to be meaningful, in the full-fledged intentional sense, simply by virtue of the role that representation plays in the agent's system of symbols. So, for example, an agent can have a representation like a concept RED, that is meaningful solely by virtue of its relationships with other representations, like COLOR and APPLE. Of course, COLOR and APPLE are likewise meaningful only by virtue of their relationships with other symbols, and so on.

Many AI systems have been regarded as, in effect, implementations of functional role theories, and several objections to early AI systems have to do with their implementing such theories. A generalization of the problem is known as "The Symbol

Grounding Problem" (Harnad, 1990). Searle's (1980) Chinese Room Argument is one such argument. Searle imagines a scenario in which he, a non-speaker of Chinese, is in a room with a large book with instructions for manipulating Chinese symbols. He receives Chinese symbols in the room, looks them up in the book, looks up indexed representations, and so on, until he is led to an output symbol. Searle argues that in such a scenario, he is manipulating symbols in much the same way that a computer would, but that such symbol manipulation clearly is not enough to make the symbols meaningful. Some human interpreters are involved in making the Chinese symbols meaningful. Similarly, as Harnad puts it, functional role theories are like a "Chinese-Chinese dictionary-go-round", where symbols are defined by other symbols, which are defined by still further symbols, and there is never anything outside of the system of symbols to "ground" any of them.

13.2 Causal Theories

Causal theories are based on the observed problems for functional role theories, as well as observations that come from Kripke (1980), Putnam (1957), and Burge (1979). The gist of the Kripke/Putnam/Burge observations is that the contents of mental representations are often real things in the material world, and are connected to mental representations of them by some physical causal interactions with them. One such causal theory is an early version of Fodor's (1986) account, which is designed to be an account of concepts, not mental representations in general. On that account, a representation S represents some content M if, and only if, all and only instances of M cause S to be activated or *tokenized*.

The characterization is not, however, sufficient for something's being a mental representation. Footprints satisfy the requirement, for they are tokened when and only when a foot has been present. But, of course, footprints are not representations that have intentionality. The requirement is also too strong. It is not necessary for being a mental representation that it perfectly reliably tracks its content. Mental representations can *mis*represent things. For example, my CAT representation might be caused to be tokened by a rabbit, and, likewise, it may fail to be caused to be tokened by a cat –suppose the cat is wearing a rabbit costume. Brentano (1874) pointed out this *reality-neutral* feature of mental representations. He also pointed out that mental representations can have *nonexistent* contents. For example, we can have a mental representation of unicorns, even though unicorns do not exist. From this observation that some contents don't exist in the physical world, Brentano argues that mental representations cannot require causal interaction with their contents for their intentionality.

13.3 Teleological and Nomological Theories

The observations about the possibility of mis-representation and empty contents have led to the development of teleological and nomological theories. These theories make claims about what *should*, by design, or *would*, by law of nature (respectively) cause the tokenings of mental representations.

Teleological theories (Millikan, 1984) are attempts to capture the possibilities of misrepresentation and empty contents by suggesting that mental representations have "proper functions" which determine their contents. More spelled out, a mental representation S represents some content M by virtue of having the function of

representing that content. The trouble with teleological theories is that it is not clear how to explicate what a “proper function” is. (We will see one approach, of Millikan’s, shortly).

Fodor’s asymmetric dependency theory aims to capture the possibilities of misrepresentation and empty contents without having to worry about the notion of a 'proper function'. His theory is *nomological*, which is to say that it appeals to the laws of nature. This theory can be characterized by two claims. (1) A representation S represents some content M by virtue of the fact that, *in ideal circumstances*, instances of M and only instances of M cause tokenings of S. (2) If something that is not an instance of M causes a tokening of S, it is only *because* of (1). Now, one trouble with nomological theories is that it is not clear what to say about 'ideal circumstances'.

There is, however, a much more serious problem with these theories than the difficulty in fully fleshing them out. Even if they could be fully fleshed out, they do not state conditions that are in fact sufficient for characterizing mental representations. It is unfortunate that Brentano’s points have been taken to be distinguishing features of mental states, for as a result the philosophical theories have focused on trying to give an account of something such that it can misrepresent and have empty contents. The trouble is, these *aren’t* the distinguishing features because they are not sufficient for characterizing mental representations. Many non-mental representations, like states of a digital clock, can misrepresent and could possibly have false contents. If they are necessary conditions, more is required for a full characterization of the intentionality of mental representation. Next we consider the idea that being *biological* is the missing component.

13.4 Biological Theories

Biological Theories take note of the fact that all heretofore known cases of mental representation have a biological basis, and then posit this feature as a necessary condition on mental mental representation.

Millikan (1984) uses a biological feature to fill out the "proper function" aspect of here teleological account. The theory is, roughly, that the proper function of a mental representation is the role it was designed (i.e., selected) by evolution to play in the fitness of the agent's ancestors. So, in that sense, a mental representation S represents whatever is its proper function to represent. This makes her account more than just teleological, for the evolutionary determinant of the proper function of the representation involves a biological factor. One problem for Millikan's view is the Swampman problem. If a molecule-for-molecule duplicate of me were to emerge by sheer accident (there is *some* chance that this has occurred, however infinitesimal), it seems that this being could at least have *some* thoughts. Indeed, for all I know, I *am* swamp-Iris, at it seems that at the very least I'd be able to have thoughts directed at *those* things directly before me. This suggests that Millikan's biology requirement is not a necessity condition on intentionality. Another biological theory is the one suggested in Searle's "Chinese Room" paper (1980).

Swamp-Iris, though not evolved and perhaps without states that have a "proper function", has a brain. Searle's idea seems to be simply that the only remaining clear difference between mental and non-mental representations is that mental ones seem to have a biological basis. The problem with Searle's account, and biological accounts in

general, is that they fail to offer any suggestion about *what it is* it about biology that makes biological systems necessary for intentional representations.

13.5 Other Theories

Brentano's (1874) offers the following requirement, in conjunction with the possibility of misrepresentation and empty contents. He claims that all and only mental representations are non-material. The problem with this claim is that it is still not sufficient for mental representations. Ordinary computational devices contain representations that are non-material entities; they contain 'virtual objects', like text, documents, and windows (see Pollock, 2006 for a thorough discussion of this software/hardware analogy). These things are distinct from the hardware of the machines that instantiate them. Yet, of course, even ones that can misrepresent and have empty contents as well, such representations are not intentional. If this account were right, we would already have AI, we would have it in our desktop computers.

Dreyfus (1981) has a knowledge requirement for the possession of concepts. His complaint against AI is that its systems do not truly understand utterances because understanding (and plausibly he would say the same for meaning utterances) requires that the agent know information about things that fall under the concepts involved. For example, he would argue that AI dialog systems do not have a representation of the property of being *red* unless they can tell us that red is a color and perhaps even that it is a surface property of reflecting light at particular wavelengths and that it is the color of apples.

Among the problems for this account, it fails to capture intentionality because it is not necessary for the possession of a concept that someone know about things that fall under it. Socially acquired concepts seem to have the feature that agents can know them without knowing anything about objects that fall under them. Consider the concepts ELM and CARBURATOR. Most people know very little about elms and carburetors, yet it seems they may still have the concepts. Indeed, if a person heard someone say, “I saw a pingloo yesterday,” they would plausibly form the concept PINGLOO, assuming that there is *some* class of things being referred to.

Searle (1980) considers and rejects a “Robot” view. On this view, an AI system that manipulates meaningless symbols could be made to manipulate meaningful symbols by attaching cameras and moving body parts so that its symbols are connected to the world.

Searle’s reply to this view is that it is not enough to be connected in the world in this way. He suggests we imagine the same Chinese Room situation he discusses earlier, except that, unbeknownst to him, some of the symbols he receives come from cameras. It seems, in such a scenario, the representations Searle would be manipulating are not intentional. We will see in the next section, however, that this reply plausibly comes from a naïve view about how such representations could get connected up so that intentionality indeed is transferred from the perceptual representations to the ones constitutive of thoughts.

Harnad (1990) offers a view on which symbolic representations can get 'grounded', and thereby come to be intentional, by having a causal connection with sub-symbolic, perceptual representations that track the world. This view is very close to the

one that will be suggested and defended in the next section. The trouble with what Harnad offers, however, is that it is not in any way *illuminating* of the property of intentionality. He does not suggest *what it is* about sub-symbolic representations that makes them intentional in the basic sense, and not requiring of any further interpretation for their meaningfulness.

Dretske (1981) offers a similar account to that of Harnad, using the terms 'analog' and 'digital' in a way that is similar to Harnad's 'sub-symbolic/symbolic' distinction. Dretske adds a role for the *acquisition* of representation-content links on the basis of perceptual, 'analog' representations. This acquisition point is also central to the view defended in the next section. However, again, on its own it is not illuminating. The acquisition of representation-content links may indeed be a necessary factor for intentionality, but it is not shown *why* this makes mental representations meaningful in the basic sense characteristic of intentionality.

In summary of this section, the features that Brentano focused on are *not* the distinguishing features of mental states that makes them intentional. The possibility of misrepresentation and empty contents are not characteristic, since man-made gauges and tracking systems clearly have these features. The feature of not being material is not characteristic either, because the virtual objects that are the representations in a hand-held calculator have that feature. The theories that capture features that are unique to mental phenomena don't seem to go the distance either, for they don't serve to *illuminate* the notion of intentionality. They leave it a complete mystery what it could be about those factors –correlating with flesh, being in agents that interact with the world perceptually and motorically, or involving sub-symbolic representations – that make mental

representations meaningful in the basic, intrinsic sense that is characteristic of intentionality.

Appendix: Further Considerations on the Rationality of Concept Acquisition

1. Introduction

The present discussion covers an approach to epistemic justification that arose in response to the failure of the Cartesian project. The Cartesian project was based on the 17th Century (Descartes, Locke, Hume) notion of epistemic justification as requiring complete, air-tight indubitability. Descartes' hope was to establish a solid and stable science about the outside world that would enjoy the epistemic status that mathematics seems to have. The approach was to derive beliefs about the outside world by use of purely deductive reasoning, starting with a foundation of indubitable beliefs about our sensory experiences along with built-in first principles. While Cartesian skepticism loomed, however, the veil between the world and our experiences held thick. Two approaches then emerged for eliminating the veil by reconsidering the situation.

One approach was to rethink the external world by, in effect, making it internal. The hope was to show that we can analyze all of our concepts about the external world in terms of our sensory experiences. Beliefs about apples, for example, would be understood as beliefs about, say, our reddish, roundish, and sweetish experiences; beliefs about dogs would be understood as beliefs about, if you will, our softish, furryish, and barkish experiences. Beliefs about apples and dogs could thereby enjoy the high epistemic status they would inherit from beliefs about sensory experiences. The success of the approach would, of course, rely on the possibility of indeed analyzing our external world concepts in

terms of indubitable sensory experiences. Attempts to do so have been shown, however, to be impossible (Quine 1951; Fodor 1987).

It seemed the only hope left for escaping epistemic skepticism was to rethink the notion of epistemic justification itself. On this track, the new approach begins with the claim that we are forced by many of the belief-producing faculties (perception, reasoning, memory, etc.) to endorse the beliefs that are produced by them. Not only are the beliefs produced automatically, goes the claim, but it happens *immediately*, so there is no *gap* in the processing at which we might willfully jump in and disrupt the automatic production of beliefs by these faculties. From this observation, then, it is argued that we are epistemically justified in endorsing beliefs that are so bound by our faculties. Authors who can (and will for present purposes) be interpreted as taking this approach include Reid, Moore, Bealer, Alston, Goodman, Pollock, among others.[35] I will refer to this approach as Cognitive Architecture Justification (henceforth CAJ). The majority of the present paper will address CAJ as it is defended by Thomas Reid.

Let us begin by considering the three central tenets of CAJ, and then we test them against contemporary results from cognitive science. My hope in doing this is primarily to direct research on this issue by framing it as a largely *empirical* question about which aspects of the processing of these faculties are automatic, and forced upon us, and which aspects are subject to the will and therefore are candidates for epistemic praise and blame. In cognitive scientific

35 Many of these authors have and/or might well be argued not to take this approach.

terms, we will frame the issue as asking which aspects of processing, from the inputs to these faculties to the ultimate output conceptual beliefs endorsed about the world, are 'impenetrable' to conscious direction and which aspects are 'penetrable' in the sense that they are within our control. I will focus mostly on visual perception, but I expect that the general observations that are made will apply to other faculties as well. The paper proceeds as follows: Section 2 describes each of the central tenets of the CAJ approach as will be discussed in the present paper. In Section 3 we step through each of the CAJ tenets to show how poorly they hold up in the face of introspective and empirical findings. Section 4 summarizes the discussion as revealing the failure of the CAJ approach for eliminating the Cartesian veil.

2. Cognitive Architecture Justification

The CAJ approach to dealing with the Cartesian veil is to argue that there is no such veil, that our epistemic faculties put us in direct contact with the world, and to argue that we are epistemically justified in beliefs produced by our faculties by mere virtue of the design of our architecture. On this more modern approach, our cognitive architecture 'gives', or, rather *forces on*, us beliefs about the world, and we are therefore epistemically justified, in the strongest, knowledge-conferring sense, in holding beliefs that are so bound by our architecture. The idea is that we are built in such a way that we process information in accordance with some rules, the 'first principles' of our cognitive architecture, so we are not epistemically culpable for holding these beliefs, and indeed, we are

epistemically praiseworthy for doing so, for this is all that being epistemically praiseworthy amounts to. Notice that for the purposes of the present paper my interpretation of these authors takes the CAJ not as a *practical* vindication from skepticism, but rather as an *epistemic* vindication, as an approach to the epistemic justification of beliefs that we are bound to by the workings of our cognitive faculties.

Here are the central tenets of CAJ regarding beliefs produced by visual perception, followed by a description of each, guided by excerpts taken from proponents of the approach:

(1) Direct Seeing: Through vision we are directly acquainted with the external world; we are not acquainted with the world by first being acquainted with sense data. At the very least, we directly/immediately see particular objects, properties, and relations, and at most we directly/immediately see these particulars *as* falling under various conceptual categories, like *red, cat*, *tree, chair, sphere*, and the like.

(2) Epistemic Bondage: The direct seeing then causes (or perhaps is identical with) beliefs about those objects, properties, and relations in the world. This connection between direct seeing and belief is simply the automatic endorsement of the propositions presented by vision.

(3) Justification by Design: We are epistemically justified, in the knowledge-conferring sense, in holding the beliefs about the world that are automatically produced by vision. All it is to be epistemically justified in a belief is, more or less by definition, for the belief to be formed the *right* way, and the right way is, more or less by definition, the way we were designed to do it.

As we will see, it is not so clear where to draw the lines between the tenets. To some extent, some authors seem to equate immediate seeing with immediate believing, and the immediacy of the seeing and believing as leaving no room for things to go differently from how they should go by design. Likewise, there seems to be a blurring between how things *in fact* go automatically and how they *ought* epistemically to go.

2.1 Direct Seeing

Here we focus primarily on Thomas Reid's account of immediate (or direct) perception from his *Essays on the Intellectual Powers of Man* (1941; see D.R. Brookes (ed.), 2002). In this essay Reid lists many "first principles", which he claims cannot but be believed to be true, since they are 'no sooner understood than believed'. Among these first principles are beliefs about the world produced, on his view, *immediately*, through perception. He observes that perceptions come along with the following three components. "*First*, Some conception or notion of the object perceived. *Secondly*, A strong and irresistible conviction of its present existence. And, *thirdly*, That this conviction and belief

are immediate and not the effect of reasoning" (II. V: 96). The present subsection focuses on the first of these components.

Reid notices a distinction between sensations and perceptions. Whereas perceptions point outwards onto the world, i.e., they *represent* things, mere feelings and sensations do not. Pains, for example, and perhaps even the qualitative character of a visual experience, are not *about* anything, they are just sensations. Indeed if all we receive from vision are sensations, as the 17th Century tradition holds, we somehow must find a way to link the sensations with the world in order for them to be a rational basis for the beliefs about the world. But Reid points out that visual perceptions are not mere sensations; they come along with contents. Indeed, it is difficult, if not impossible, as Reid points out, to reflect on the mere qualitative feel of a perceptual experience and ignore the objects in the world that the experience presents to us. If perceptual experiences in fact are representational, and are not mere feelings or sensations, then what we have as the result of perception is more than mere sense data. The perceptions themselves are links to the world. They present the world as being some way.

We can see in other passages that Reid not only takes perception to present particular objects, properties, and relations in the world, say, demonstratively, as *those things being and relating in those ways*, but he seems to take perception automatically to present the objects, properties, and relations as falling under various *concepts* or *categories*. The first of the components of perception may be seen as making this claim. Further support for this

interpretation of his first component comes from his essay *An Inquiry into the Human Mind on the Principles of Common Sense*, where he makes a distinction between "natural" and "acquired" perceptions. "Our perceptions are of two kinds. Some are natural and original, others acquired, and the fruit of experience" (VI.xx: 171). His notion of "natural perception" seems to be akin to a sort of demonstrative presentation of the world, before the further acquisition of concepts associating those experiences with abstract kinds or categories. His notion of "acquired perception" seems to be something like a conceptualized perceptual experience, where the experience itself takes the particular objects, properties, and relations as falling under various categories, like *red, tree,* and *bell*. In further support, we find that in his (1941) he says, "I can say, without impropriety… I hear a great bell, or I hear a small bell; though it is certain that the figure or size of the sounding body is not originally an object of hearing" (II,XIV:182). His idea seems to be that in our very first experiences, we already have more than the sense data theorist suggests. The experience not only has a qualitative feel, but points outward, and presents the world as being some way. Vision has a way of presenting objects, their variations in colors, and the like. Then, through experience we come to acquire the ability to categorize objects. As adults, we have formed very strong habits connecting the natural perceptions, to our acquired concepts, such strong habits that now those new concepts come to be part of how vision presents the world to us. Reid says, "It is experience that teaches me that the variation of color is an effect of spherical convexity… But so rapid is the progress of thought, from the effect to the cause, that we attend only

to the last, and can hardly be persuaded that we do not immediately see the three dimensions of the sphere" (1941, xxi: 236). Reid seems to take many category representations to be involved in the visual experience itself, and these categories are among the things we are immediately acquainted with via visual processing. Contrast this position with a direct seeing view on which the non-conceptual demonstrative proposition is what is always experienced as the result of visual processing, and some inference-like process takes the viewer from the non-conceptual experience to the conceptualized proposition. It is not clear which concepts he takes vision to be capable of containing, or immediately presenting, but his view clearly takes vision automatically to present to us the contents of many of the beliefs that the 17th Century tradition took to be inferred from sense data.

Very central to Reid's account is the *immediacy* of the conceptualized visual presentation of the world. He claims that there is no gap in processing between the sensory experience and the demonstrative outward directedness of the sensory experience as a presentation of the world. Furthermore, he claims that there is no gap between the demonstrative presentation of the world and the conceptualized/categorized presentation of the world. The role of this immediacy in his argument is that there is no gap at which *we*, as conscious responsible agents, can enter and disrupt the process.

In his defence of the immediacy of the conceptualized experience, he makes two points. (a) We do not *notice* first something other than the conceptualized world; it feels immediate. (b) There *can't* be anything that

mediates, for there is no logical link between a sensation and a category. The idea behind the first point seems to be that there is no gap in the processing in which we can penetrate the processing. He notes that we aren't aware of the un-interpreted the sensation, suggesting that we cannot isolate the sensation and block its interpretation. In the second point, he makes the further claim that (even if there were a gap) there *couldn't* have been an *inference* that would take us from the sensation to the interpreted, outward-directed representation, since there is no relationship between a given sensation and an interpretation. Again, for both of these claims, he seems to have in mind *two* seamless connections: (a) from the sensation to the demonstratively parsed experience that has the outward content and (b) from the demonstratively parsed experience to the conceptualized experience.

So Reid seems to hold a kind of Perceptual Direct Realism: the perceptual experience has an outward directedness. This is captured by the idea of demonstrative representations. It is not simply a sensation, a sensory quality, but it points outward; it is representational; it has aboutness; it has content. Furthermore, over time, after having lots of sensory experiences, that presumably are related in some way, our perceptual experiences present to us a world parsed into categories, categories for which we have learned associated looks. These categorized perceptions are automatic and force on us beliefs about objects as falling under categories without any inferential steps (or at least without any that are slow enough for us have control over).

This Perceptual Direct Realism has a lot of intuitive force. When we look around the room and perceive objects, we seem to experience most of them as falling under particular categories. We look around, and the world is presented to us with chairs and tables and people, with their colors and shapes, in various positions relative to ourselves to one another. We do not seem to *infer* from the way things look to something's being a chair or a table, or being near or far.

2.2 Epistemic Bondage

Reid's view about immediate seeing takes him a long way to eliminating the difficulty presented by the 17th Century tradition. If perceptual experiences are representational, presenting particular objects, properties, and relations in the world, and the experiences themselves actually present objects in the world as falling under various conceptualized categories, then perception itself reaches out into the world. This seems, on the face of it, like a step towards eliminating the 17th Century problem of penetrating a veil. We are in direct contact with the world. But it is not clear that this is enough, and indeed, Reid does not stop there. In his *Second* and *Third* elements of perception, he seems to take the perceptual experience not only to contain such presentations of the world, but to contain the *endorsement* of the presentations, the *conviction* or *very compelling belief* that the objects indeed exist in the perceiver-relative locations, with conceptualized properties and relations as presented perceptually. The first things we get from perception are *beliefs* about the world, according to Reid, these come *immediately*; we are not first given a sensation, a mere qualitative

experience, we are given a fully parsed experience. Not only that, we are given a fully parsed *belief* in the accuracy of the proposition presented experience.

Like the Perceptual Direct Realism in the first CAJ tenet, this further step can be regarded as a sort of *Belief* Direct Realism, in which our *beliefs* directly "reach out" into the world, and are not mediated by any other representations, not by sensations, nor by perceptual propositions, demonstrative or conceptual, that reach out into the world. Thus we can carve up two versions of Direct Realism:

> Perceptual DR: The first objects of acquaintance are presentations of propositions *as of* objects, properties, and relations in the world. In natural perception these propositions are fully demonstrative; in acquired perception the propositions are (at least partially) conceptualized.

> Belief DR: The first objects of acquaintance are *beliefs in* propositions *as of* objects, properties, and relations in the world. In natural perception these propositions are fully demonstrative; in acquired perception the propositions are (at least partially) conceptualized.

It should be noted that Belief DR is not to be interpreted as making any claims in itself about the *justification* of the belief. There is a further *normative* notion of

Direct Realism that seems to appear in defenses of CAJ (for example, Pollock 1986), giving rise to the following further distinction:

> Descriptive DR: Propositions and/or endorsed propositions about the world are directly/immediately produced by perception.
>
> Normative DR: Beliefs about the world produced by perception are directly/immediately *justified*.

So there is a normative version, having to do with justification, and a descriptive or psychological one, having to do with the actual production of beliefs from perception.

The Normative version of DR is left for the next subsection. So far we are dealing only with Descriptive DR, as it applies to the processing leading up to perceptual experiences of the world, and as it applies to the processing leading up to beliefs about the world.

Now, it is not clear in Reid's text that he is sensitive to these distinctions. Regarding Perceptual DR, Reid says, "When we see the sun or moon we have no doubt that the very objects which we immediately see are very far from us and from one another…. But how are we astonished when the philosopher informs us, that we are mistaken in all this… because the objects we perceive are only ideas in our own minds…." (EIP II.xiv:172.). Here, Reid actually seems to be running Perceptual DR and Belief DR together. Indeed, he seems to think the

entertaining of a visual proposition more or less *is* the endorsing of that proposition. Regarding Belief DR, Reid says the following. "If the word axiom be put to signify every truth which is known immediately without being deduced from any antecedent truth, then the existence of the objects of sense may be called an axiom" (EIP II.xx:231). Here he is using the normative notion of *knowledge* along with the psychological claim about how we come to have the beliefs. The descriptive version tells us about how natural and automatic it is for us to accept the authority of the senses, and that we do so without any *mediating step* taking us from the perceptual experience to the perceptual beliefs. The beliefs occur automatically as a result of perceptual processing. The connection between the visual experience and the belief is immediate. Presumably the force of the immediacy claim is that it leaves no gap in which *we* could jump in and re-direct the processing so that the belief is not produced. The CAJ theorists may then take this descriptive point and then use it to defend the normative claim that we are justified in our (immediate) perceptual beliefs. It would be *because* beliefs about objects in the world are produced automatically from, and indeed are *forced on us by*, the immediate perception of those objects, that we are justified in endorsing those beliefs. The same argument seems to be used for CAJ theories about the non-perceptual faculties. For example Nelson Goodman (1954) argues that Hume's (1748) riddle about the justification of induction can be solved by comparison to the justification of deduction. Deductive arguments are valid as long as they follow the rules of deduction; likewise, we have socially accepted rules for induction and that all it is to do it right is to follow the

rules (as for grammatical correctness). Reid's "first principles" are similarly thought to be forced upon us by our architecture. He takes these to be justified because of the descriptive point that they are "no sooner understood than believed".

2.3 Epistemic Justification by Design

The third tenet, and final step, for CAJ is that there is no room but for us to be justified in beliefs that are immediately produced from perception. Again, the sense of justification here is the strongest, knowledge-conferring sense. We are not merely justified in these beliefs in that we can't be held practically culpable, according to CAJ, but rather we are justified in the sense that such beliefs, if true (barring Gettier considerations), amount to knowledge.[36] Again, Reid doesn't clearly distinguish the descriptive and the normative claims. However, we may turn to G.E. Moore's defense of CAJ, in his "Proof of an External World", for explicit moves to the justification of such automatic beliefs (1939).

Moore does not offer a justification, in the sense of offering a *proof* for the truth of automatic beliefs from perception, like the belief that he has hands. But he argues that it is clear nonetheless that he *is* justified. In his argument, he seems to weigh the *confidence* in his beliefs that are produced by perception

36 It is interesting to notice the following asymmetry in CAJ: In order to be *unjustified*, it seems you need to have a mediating step, a gap in processing at which "we" can wilfully jump in and take responsibility, whereas such a gap is not necessary in order to be *justified*.

against his confidence in their negations that are produced by reasoning. He finds that perception is more persuasive than reasoning.

> Russell's view that I do not know for certain that this is a pencil or that you are conscious rests, if I am right, on no less than four distinct assumptions. (1) That I don't know these things immediately; (2) That they don't follow logically from any thing or things that I do know immediately; (3) That *if* (1) and (2) are true, my belief in or knowledge of them must be "based on an analogical or inductive argument"; and (4) That what is so based cannot be *certain knowledge*. And what I can't help asking myself is this: Is it, in fact, as certain that these four assumptions are true, as that I *do* know that this is a pencil and that you are conscious?

In other words, Moore seems to have noticed a *paradox*. It seems, intuitively, that (0) we have knowledge about the external world. Assumptions (1) through (4) from the skeptical argument also seem to have some intuitive appeal. But from (1) through (4), we can deduce that we *don't* have knowledge of the external world. Seeing the situation as a paradox, Moore resolves the paradox by denying the conjunction of the intuitions (1) through (4) and holding onto the intuition that (0). Moore seems to include in epistemic cognition, a mechanism for choosing between conflicting beliefs, and he argues that this mechanism chooses in favor of perceptually-produced beliefs over reason-produced beliefs.

Other defenders of CAJ seem to give different arguments for the epistemic justification of the beliefs. Again, for Nelson Goodman, all it *is*, by definition, to be justified in a belief is to come to the belief *in the right way*. The only notion we have of *the right way* comes from introspecting the cases in which beliefs are produced in the right way. Presumably what drives these intuitions is our set of built-in procedures for producing beliefs. So if we're doing it the way we were designed to do it, we're justified. George Bealer

(1996) seems to offer yet another defense. In his paper, "*A Priori* Knowledge and the Scope of Philosophy" is a defense of the justification of our reliance on *intuitions*, but his defense could easily be brought to apply to beliefs produced by perception as well. He seems to offer an evolutionary argument, or an argument from conservatism. His idea seems to be that practices should be adhered to merely because of their being firmly established practices and natural processes.

3. Are We Epistemically Bound by the Senses?

Now we consider in turn each of the tenets of CAJ and show that there is some wiggle room after all. Our approach is to consider observations from introspection and from Cognitive Science. Tenets (1) and (2) seem to be claims about the penetrability of various cognitive processes, an issue that has been dealt with in Cognitive Science. In particular, there have been empirical studies that examine which aspects of the processing from retina to belief are automatic, what parts are impenetrable to higher-level information, what aspects can we *willfully* penetrate? Tenet (3) of CAJ is the normative claim, but it likewise has a large empirical component. The claim is that we are justified in holding the beliefs that are automatically produced, supposing such beliefs exist. Presumably we are only justified in maintaining such beliefs if they cannot be *defeated* by beliefs produced by other processes, such as reasoning. So here again, we must turn to empirical study to find out whether such beliefs, even if automatically endorsed, can be rejected in later epistemic processing. The

sections that follow take on each of the tenets of CAJ in turn, treating each as making claims about different stages of processing from the inputs to our faculties to the ultimate long-term endorsement of beliefs about the world.

3.1 The (Im)penetrability of Vision

In this sub-section, we consider the claims made by CAJ theorists about Perceptual DR. Recall that there are two points in processing at which Reid claims there is *immediacy*. First, he claims that there is no gap between the sensation and the outward directed, demonstrative perceptual presentation of the world. Even in our very first experiences, our "natural experiences", the world is directly presented to us, with objects, properties, and relations. Second, he claims that after having many such experiences we come to learn associations between our natural experiences and the categories that things producing those natural experiences fall into. For example, we learn that certain kinds of shadings on objects is an indication that the objects are spherical. Once we come to have such associations, the perceptual systems immediately present the world to us with objects *as* falling under those acquired categories. Again, he is not explicit about *which* categories can be immediately presented through perception, but it is clear that he takes perception itself to penetrate the veil posited by the 17th Century tradition. In addressing this tenet, there are several considerations to be made. Some of these can be made "from the armchair" in the sense that we can reflect on our memories and generalize about our

experiences, and other considerations can be made by looking into empirical studies from cognitive science.

3.1.1 The first gap

The first observation to make is that while it's true that we don't *usually* notice and attend to our sensations, we *can* notice them, independent of their outward directedness. For example, when I wake up in the morning, I look around my room and see my desk, and chair, and scattered notes. But I notice that it is difficult to see these things, for I am not wearing my glasses. I notice that my visual *sensation* is blurry.

When I do this, I do not automatically interpret the *world* as being blurry, I am directly in contact with my sensation at these times, not the world. There are other times, I can recall, when I attend to my visual sensation and not to the world. For example, when I am considering skeptical possibilities, for example, that I am a bodiless brain in a vat, being fed artificial stimulation via a computer. When considering this possibility, I pretend that is my situation, and notice my visual sensations. These observations suggest that although perception *usually* presents to us the world, we are able to block this presentation upon willfully deciding to do so.

If there is indeed *sometimes* a gap here, when we *willfully* put one in, between the sensation and the outward-directed presentation of the world, then perhaps we can prevent the outward-directed experience from occurring.

Although it takes effort, it seems to be within our control, and perhaps we can train ourselves so that it takes less effort over time.

Work on visual processing from the cognitive science literature deals explicitly with the issue of where visual processing is "encapsulated" and "impenetrable" to higher level cognition. In this work, as in the introspective considerations just made, the issue is presented less as one about the *automatic* or *default* processing of the system, and more as an issue about which aspects of the processing we can "enter" and direct. Zenon Pylyshyn, in his (1998) article on this topic, presents an empirical argument that the processing that occurs in what he calls "early vision" from inputs to the retina up to a demonstrative presentation of the world, of 3D objects and their viewer-centered locations, is impenetrable to higher-level cognition. Among his evidence, he notes cases of visual illusion, like the two Muller-Lyer lines, which although in fact are the same length, appear to extend out to different locations. Even when the viewer understands the illusion and *believes* that the lengths of the lines are the same, they cannot seem to disrupt the processing up to the appearance of them as different lengths. Another observation that leads Pylyshyn to his claim about the impenetrability of vision is the evidence that visual processing involves many sophisticated computations that we cannot compute consciously. We wouldn't know how to redirect the processing of objects and their locations. As further support, he cites evidence from neuroscience for independent pathways for visual processing and higher-level cognitive cortical areas.

The cognitive science data seem to suggest that Reid is right about the seamlessness of processing of visual information to at least the unconceptualized demonstrative presentation of the world. But how can we treat these considerations against the introspective observation of our ability to attend only to the visual sensation and block the outward-directed presentation of the world? Although these observations leave the issue unresolved, they provide direction for the study of direct seeing as an empirical issue.

3.1.2 The second gap

Next consider the issue of there being a second gap, one between the demonstrative presentation of the world that we have in "natural perceptions" and the conceptualized presentation of the world that we have in "acquired perceptions". Can we decide *not* to conceptualize the objects and properties that are demonstratively presented to us? Alternatively, can we decide *how* to conceptualize the objects and properties that are demonstratively presented to us? Again, we can make observations from the armchair as well as in the laboratory.

Although perception does seem, as Reid observes, automatically to present to us a conceptualized presentation of the world, the issue is really about whether we can jump in and direct the processing. There are a couple of observations we can easily make from where we sit. First, notice that we sometimes are creative in how we conceptualize the objects presented to us

visually, even if there is an automatic default conceptualization. We can see a chair as a footstool, a stapler as a paperweight, a hand as a cup, and so on.

Again, we find that the cognitive science literature has a way of discussing this same issue. Eleanor Rosch's (1978) paper "Cognition and Categorization" introduces the notion of the "basic level category", to talk about categories, like *chair, car, tree, bird,* by which we automatically parse objects upon perception. While her discussion does not make it clear *what* this basic level is, she hits on an intuition that seems similar to the one that drives Reid's notion of "acquired perception". Upon seeing an object, say a particular bird, we seem to conceptualize it by default as a *bird*, rather than as an *animal*, its super-ordinate category, or as, say, a *pigeon*, its sub-ordinate category. She notes, further, however, that the default category by which we conceptualize an object can vary with context. If the task is to classify various birds, we automatically see the objects as falling under the more specific categories. This latter observation suggests that there is some penetrability from higher-level cognition over the conceptualized perceptual presentations. According to Pylyshyn, the processing that occurs after early vision, that is after the non-conceptual demonstrative presentation from vision, can indeed be penetrated. Here he cites observations involving ambiguous pictures, such as the famous duck/rabbit picture. While subjects can be primed to see the picture as a duck or as a rabbit, these subjects can be shown how to reinterpret the image.

As a further approach to addressing this issue, we might ask to what extent we are responsible over our "acquired perceptions". Perhaps to some

extent we can train or retrain ourselves to conceptualize objects in various ways. Further research into this issue as an empirical question is due.

3.2 The (Im-)penetrability of Perceptual Beliefs

Recall that Reid takes the *endorsement* of the propositions presented by perception to be among the "elements of perception". Indeed, it is this seamlessness between the conceptualized perceptual presentation and the epistemic commitment to the presented proposition that bring him to his claim about the justification of perceptual beliefs. In evaluating his claims, however, we should consider these as separate stages of processing, whether or not there is an introspectible gap between them and whether or not we can enter that gap and re-direct the processing. Here again, we can consider some observations from the armchair. While there are no empirical studies, to my knowledge, that address this issue, we will see how it could indeed be studied empirically.

As a first observation, we should consider occasions in which we do not trust perception. For example, when we know we are in a situation in which our perceptual faculties are unreliable, it seems we are able to withhold endorsement of the propositions presented by our faculties. Consider, again, the cases of optical illusion, like the Müller-Lyer case. After exposure to the trick, upon looking at the lines we still have the experience of them as being different lengths. However, recall also that the observation that drove the impenetrability claim in the previous subjection was that we *believe* that the lines are the same

length, in spite of the visual appearance. This suggests that there is indeed a gap between the experience and the belief about the world, and it suggests that we can indeed take responsibility and enter this gap to prevent the production of the belief.

A second observation is that even when, as far as we can tell, we are in normal perceptual circumstances, we do not always endorse the propositions presented by perception. For example, suppose you were sitting at your desk in your office, and you have a visual experience of a 5ft grasshopper standing at your doorway. My guess is that you would consider the visually presented proposition, but you would not endorse it. Alternatively, you might endorse it briefly, but quickly find defeaters from your other beliefs about your office environment and about grasshoppers. This second alternative suggests that there indeed is no gap between the perceptual presentation and the belief, but that there is room *after* the initial endorsement for you to rationally override the belief. Again, this is clearly an issue for empirical investigation. If indeed the latter is the correct sequence of events, then we enter into the domain of epistemic normativity.

3.3 Are We Epistemically Bound?

The final potential gap in processing is between beliefs produced as a direct or immediate consequence of perception and the further epistemic cognition that results in the ultimate endorsement or rejection of the belief. Even if the first beliefs from perception are produced automatically, and even if we cannot

penetrate this automatic process, we may not be *bound* to the belief. Clearly beliefs or candidate beliefs produced by various faculties come into conflict. Indeed, this observation was made by Moore in his defence of CAJ. Also, the observation about the unusual perceptual experiences suggest that we do not always believe our eyes. It seems, then, that at the very least we can override beliefs that are produced by perception, at least some of the time. Moore's claim was simply that *in the case* of perceptual beliefs about our hands, or pencils, or other minds, being pitted against their negations from the faculties of reason, we are more compelled by the perceptual beliefs.

The idea now seems to be that epistemic cognition has some way of weighing the beliefs produced by various faculties when the beliefs come into conflict. Suppose there is indeed such a weighing mechanism, and its output is some sense of confidence in one proposition over another. Perhaps that is all there is to the justification of a belief. This seems to be Moore's claim. Alternatively, we may regard this mechanism as another belief-producing mechanism, one that produces higher-order beliefs, beliefs about the relative confidence in lower-order beliefs. In that case, we may want to question the reliability of this mechanism. Here we seem to lose our grip on the issues. In what sense can we be said to be epistemically *responsible*? Can we override the weighing mechanism, and by what norms are we to override it? The CAJ approach seems to lead to an infinite regress before any decision can be made about the endorsement of beliefs. This seems to suggest either the collapse of

our notion of justification or a retreat to some, perhaps more modern, externalist notion of justification.

4. Conclusion

The present discussion considered an approach to dealing with the 17th Century problem of having justified beliefs about the external world on the basis of our internal sensory experiences. The approach, which I named "Cognitive Architecture Justification", makes claims about the seamlessness in the processing through the faculties to our beliefs about the world. The seamlessness is presented as a kind of bondage to our faculties that not only leaves us with beliefs about the world for which we are not epistemically culpable, but for which we are epistemically praiseworthy. In evaluating the claims made in support of CAJ, we brought in observations from introspection and from cognitive science that suggest that the processing is not so seamless after all. There are several gaps along the way at which we might, and indeed presumably do, "penetrate" and redirect or block the processing. Furthermore, we saw that the only way out for the CAJ theorist is to posit further belief-producing mechanisms for evaluating belief-producing mechanisms, and that this approach cannot work for it leads to a regress in epistemic decision-making.

Bibliography

Armstrong, Sharon Lee, Gleitman, Lila R., and Gleitman, Henry. (1983). What some concepts might not be. *Cognition*, 13: 263-308.

Barsalou, Lawrence. (1999). Perceptual Symbol Systems. *Behavioral and Brain Sciences,* 22: 577-660.

Barsalou, Lawrence. (1987). The instability of graded structure: Implications for the nature of concepts. In U.Neisser (Ed.), *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization.* 101-140. New York: Cambridge University Press.

Bealer, G. (1996). A Priori Knowledge and the Scope of Philosophy. *Philosophical Studies*, 81: 121-42.

Block, Ned. (1991). Meaning Holism and Conceptual Role Semantics. In J. Fodor and E. Lepore (eds), *Holism: A Shopper's Guide.* Blackwell.

Bloom, Paul. (2000). *How Children Learn the Meanings of Words.* Cambridge, MA. MIT Press.

Brandom, Robert. (1994) *Making it Explicit*. Cambridge, Massachusetts: Harvard University Press.

Burge, T. (1979). Individualism and the Mental. *Midwest Studies in Philosophy*, 4:73-121.

Keil, Frank. (1986). The acquisition of natural kind and artifact terms. In W. Demopoulos and A. Marras (Eds.), *Language Learning and Concept Acquisition.* 133-153. Norwood, New Jersey: Ablex.

Carey, Susan. (1985). *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.

Chalmers, David (2004). The foundations of *two-dimensional semantics*. In *Two-Dimensional Semantics: Foundations and applications*. M. Garcia-Caprintero and J. Macia, eds. OUP.

Chalmers, David. (2004). The Representational Character of Experience. In B. Leiter, ed., *The Future for Philosophy*. Oxford University Press.

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

Clark, Andy (2008) *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. OUP.

Clark, H.H., and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22, 1-39.

Crane, T. 1992. The Nonconceptual Contents of Experience. *The Contents of Experience: Essays on Perception*. Cambridge University Press.

Cummins, R. (1997). The LOT of the Causal Theory of Mental Content. *Journal of Philosophy* 94:535-42.

Davidson, Donald. (1991). Meaning Holism and Radical Interpretation. In J. Fodor and E. Lepore (eds), *Holism: A Shopper's Guide.* Blackwell.

Dennett, Daniel C. (1991). Meaning Holism and The Normativity of Intentional Ascription (and ALittle More about Davidson). In J. Fodor and E. Lepore (eds), *Holism: A Shopper's Guide.* Blackwell.

Dennett, Daniel C. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.

Dennett, Daniel C. (1977). Review of Jerry Fodor's The Language of Thought. *Mind*, 86 342:265-280.

Descartes, Rene. (1641). *Meditations on First Philosophy*.

DeVault, David, Iris Oved, and Matthew Stone (2006). Societal grounding is essential to meaningful language use. *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*.Boston, MA.

Devitt, M. and Sterelny, K. (1999). *Language and Reality: An Introduction to the Philosophy of Language*. Second ed. MIT Press.

Dretske, Fred. (1988). *Explaining Behaviour*. MIT Press.

Dretske, Fred. (1981). *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Dreyfus, Hubert (1972). *What Computers Can't Do: A Critique of Artificial Reason.* New York: MIT Press.

Dummett, Michael. (1993). *The Seas of Language*. Oxford: Clarendon.

Evans, Gareth. (1973). The Causal Theory of Names. *Aristotelian Society Supplementary Volume xlvii* pp. 187-208.

Eccles, Ron. (2005). Understanding the symptoms of the common cold and influenza. *The Lancet Infections Diseases*, Vol 5, Issue 11, Pages 718-725.

Fodor, Jerry. (2008). *LOT 2: The Language of Thought Revisited*. OUP.

Fodor, Jerry A., and Ernest Leopore. (2002). *The Compositionality Papers.* Oxford University Press.

Fodor, Jerry A.**,** and Ernest Lepore. (1999). All at Sea in Semantic Space: Churchland on Meaning Similarity. *Journal Of Philosophy,* 96 (8):381-403.

Fodor, Jerry. (1998). *Concepts: Where Cognitive Science Went Wrong*. New York: Oxford University Press.

Fodor, Jerry. (1995). *The Elm and the Expert*. MIT Press.

Fodor, Jerry. (1990). A Theory of Content, I: The Problem. In J.A. Fodor. *A Theory of Content and Other Essays*. 51-136. Cambridge, MA: MIT Press.

Fodor, J.A. (1990) A Theory of Content, II: The Theory. in *A Theory of Content and Other Essays*. 51-136. Cambridge, MA: MIT Press.

Fodor, Jerry. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge: MIT Press.

Fodor, Jerry. (1985). Fodor's Guide to Mental Representation. *Mind* 94:76-100.

Fodor, Jerry. (1983). *The Modularity of Mind.* MIT Press.

Fodor, Jerry. (1981). The present status of the innateness controversy. In Fodor, Jerry *Representations: Philosophical Essays on the Foundations of Cognitive Science*. 257-316.

Fodor, Jerry. (1975). *The Language of Thought*. New York: Thomas Y. Crowell.

Fodor, Jerry, J. Fodor, and M. Garrett (1975). The psychological unreality of semantic representations. *Linguistic Inquiry* 6, 515-531.

Frege, G. (1892) Über Sinn und Bedeutung, in *Zeitschrift für Philosophie und philosophische Kritik*,100: 25-50. Translated as 'On Sense and Reference' by M. Black in *Translations from the Philosophical Writings of Gottlob Frege*, P. Geach and M. Black (eds. and trans.), Oxford:Blackwell, third edition, 1980.

Gelman, Susan, and H.M. Wellman. (1991). Insides and essences: Early understandings of the nonobvious. *Cognition*, 38, 213-244.

Gelman, Susan.A., and Markman, M. (1986). Young children's inductions from natural kinds: The role of categories and appearances. *Child Development*, 58, 1532-1541.

Gergely, György, Z. Nadasdy, G. Csibra, and S. Biro. (1995). Taking the intentional stance at 12 months of age (1995) *Cognition*, 56 (2), pp.165-193.

Gleitman, L.R., Cassidy, K., Papafragou, A., Nappa, R., & Trueswell, J.T. (2005) Hard words. *Journal of Language Learning and Development*, 1:1. 23-64.

Goodman, N. (1954). *Fact, Fiction, and Forecast*. University of London: Athlone.

Gopnik, A. and Meltzoff A.N. (1998). Infant cognition. In *The Encyclopedia of Philosophy*. Routledge.

Gopnik, Alison, and A.N. Meltzoff (1997). Words, Thoughts and Theories. Cambridge: MIT Press.

Grice, Herbert P. (1957). Meaning. *Philosophical Review*, 64, 377–388.

Harnad, Stevan. (1990). The Symbol Grounding Problem. *Physica D*, 42: 335-346.

Haugeland, John. (1979). Understanding Natural Language. *Journal of Philosophy.* 76: 619-632.

Hoffman, Donald. (1998). *Visual Intelligence: How we create what we see*. New York: W.W. Norton and Co.

Horgan, Terry, and J. Tienson. (2002). The Intentionality of Phenomenology and the Phenomenology of Intentionality. In D. Chalmers, ed., *Philosophy of Mind: Classical and Contemporary Readings*. Oxford University Press.

Hume, David. (1739). *A Treatise of Human Nature*.

Hume, David. (1748). *An Enquiry into Human Understanding*.

Jackendoff, Ray. (1998). *The Architecture of the Language Faculty*. The MIT Press.

Káldy, Z., and Leslie, A.M. (2005). A memory span of one? Object identification in 6.5 month-old infants. *Cognition,* 97, 153–177.

Katz, J. (1972). *Semantic Theory*. Addison-Wesley Educational Publishers.

Keil, F. C. (1989). *Concepts, Kinds, and Cognitive Development*. Cambridge, MA: MIT Press.

Kripke, S.A. (1972). *Naming and Necessity*. Harvard University Press.

Laurence, Stephen and Eric Margolis. (1999). Concepts and Cognitive Science. In E. Margolis and S. Laurence (eds) *Concepts: Core Readings*. Cambridge, MA: MIT Press.

Leffler, Bethany R. Michael L. Littman, and Timothy Edmunds. (2007). Efficient Reinforcement Learning with Relocatable Action Models. *Proceedings of the Twenty-Second Conference on Artificial Intelligence.* July, 2007.

Lindsey D.T. and A.M. Brown. (2002). Color naming and the phototoxic effects of sunlight on the eye. *Psychological Science*. 13:506–512.

Locke, J. (1690) *An Essay Concerning Human Understanding*.

Ludwig, K. (2001). Phenomenal Consciousness and Intentionality: Comments on *The Significance of Consciousness*. In *Psyche* 7.

Margolis, Eric. (1998). How to Acquire a Concept. *Mind and Language*. Vol.13, no.3. 347-369.

Marr, David (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* New York: Freeman.

Matthews, R. (2007). *The Measure of Mind: Propositional Attitudes and their Attribution.*OUP.

Medin and Ortony. (1989). *Similarity and Analogical Reasoning.* New York: Cambridge University Press.

Millikan, Ruth G. (2000). *On Clear and Confused Ideas*. Cambridge: Cambridge University Press.

Millikan, Ruth G. (1984). *Language, Thought, and Other Biological Categories*: *New Foundations for Realism*. Cambridge, MA: MIT Press.

Mitchell, Thomas. (1997). *Machine Learning*. McGraw Hill.

Moore, G.E. (1939). Proof of an External World. *Proceedings of the British Academy.* 25.

Murphy, G.L. (2002). *The Big Book of Concepts*.MIT Press, A Bradford book.

Pinker, Sephen (1997). *How the Mind Works*. New York: Norton.

Peacocke, Christopher. (1992). *A Study of Concepts*. Cambridge: MIT Press.

Plato. (380 bce). *Euthyphro.*

Pollock, John L. and Iris Oved. (2005). Vision, Knowledge, and the Mystery Link. *Philosophical Perspectives*. 309-351**.**

Pollock, John and Joseph Cruz. (1999). *Contemporary Theories of Knowledge*. 2nd Edition. Rowman and Littlefield.

Pollock, John. (1986). *Contemporary Theories of Knowledge*. Rowman and Littlefield.

Prinz, J. (2002). *Furnishing the Mind: Concepts and their Perceptual Basis*. MIT Press.

Putnam, H. (1975). The meaning of meaning. *Philosophical Papers: Mind, Language and Reality*. 2:215-271. Cambridge University Press.

Pylyshyn, Zenon (2000). Situating vision in the world. *Trends in Cognitive Science*, 4(5): 197-207.

Pylyshyn, Zenon. (1998). Is vision continuous with cognition?: The case for cognitive impenetrability of visual cognition. *Behavioral and Brain Sciences*.

Quine Willard.V.O. (1951). Two Dogmas of Empiricism. *The Philosophical Review* 60: 20-43. Also in *From a Logical Point of View: Nine Logico-philosophical Essays.* 20-46. Cambridge, MA: Harvard University Press.

Reid, Thomas. (1785). *Essays on the Intellectual Powers of Man*. London (Also in 2002. D.R. Brookes (ed.). 23. Edinburgh University Press.

Reid, Thomas. (1764). *An Inquiry into the Human Mind on the Principles of Common Sense*. Edinburgh.

Rosch, Eleanor. (1978). Principles of categorization. In: E. Rosch & B. Lloyd, eds., *Cognition and Categorization*. Hillsdale, N.J.: Erlbaum Associates. 27-48.

Rosch, Eleanor and C.B. Mervis. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7:573—605.

Rosch, Eleanor. (1973). Natural categories. *Cognitive Psychology*, 4- 328-50.

Rupert, R. (2001). Coining Terms in the Language of Thought: Innateness, Emergence, and the LOT of Cummins's Argument Against the Causal Theory of Mental Content. *Journal of Philosophy* 98: 499-530.

Searle, John. (1980). Minds, brains and programs. *Behavioral and Brain Sciences*. 3(3): 417-457.

Smith, E., Osherson D., Rips, L. and Keane, M. (1988). Combining Prototypes: A selective modification model. *Cognitive Science* 12. Alex Publishing Corporation.

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

Soames, Scott. (2002). *Beyond Rigidity: The Unfinished Semantic Agenda of Naming and Necessity*. OUP.

Spelke, Elizabeth. (1994). Initial knowlege: six suggestions. *Cognition*, 50:431-445.

Sosa, Ernest. (2007). Moore's Proof. *Themes from G.E. Moore: New Essays in Epistemology and Ethics*. S. Nuccetelli and G. Seay. (eds). OUP.

Stich, Stephen. (1983). *From Folk Psychology to Cognitive Science*. Cambridge, MA: MIT Press.

Wittgenstein, Ludwig. (1958). *Philosophical Investigations*. Blackwell Publishers. Also in Anscombe, G.E.M., trans. Oxford: Basil Blackwell.

Curriculum Vitae

Iris Oved

Education:

The University of Arizona (1996 – 2001)
Rutgers University (2001 – 2009)

Degrees Awarded:

B.A. Interdisciplinary Studies (Philosophy, Psychology, and Symbolic Systems)
PhD Philosophy (with Cognitive Science Certificate)

Current Position:

Computing Innovations Postdoctoral Fellow, The University of Arizona, School for Information Science, Technology, and the Arts.

Published Papers/Abstracts:

Linda W. Norrix, Iris Oved, and Lawrence D. Rosenblum. (2000) Auditory-visual context effects in a speech and nonspeech condition. *The Journal of the Acoustical Society of America*. Vol. 107, No.5, pg. 2887. [Poster Abstract]

Iris Oved, Linda W. Norrix, and Merrill F. Garrett. (2000) Identity priming using McGurk stimuli as primes. *The Journal of the Acoustical Society of America* Vol 108, No. 5, p.2482. December 2000, Newport Beach, CA. [Poster Abstract]

Iris Oved (2006). Exploring the Role of Qualia in the Intentionality of Thought. *Toward a Science of Consciousness, 2006.* April 2006,Tucson, Arizona. [Poster Abstract]

David DeVault, Natalia Kariaeva, Anubha Kothari, Iris Oved, and Matthew Stone (2005). An Information-State Approach to Collaborative Reference. *Association for Computational Linguistics*. Proceedings Companion Volume, Pg1-4. University of Michigan, June 2005. [Conference Paper]

John Pollock and Iris Oved (2005). Vision, Knowledge, and the Mystery Link. *Philosophical Perspectives*. Pg 309-351. Blackwell. [Journal Paper]

David DeVault, Iris Oved, and Matthew Stone. (2006) Societal Grounding is Essential to Meaningful Language Use. *Proceedings of the Twenty-First National Conference on Artificial Intelligence* (AAAI-06). Boston, July 2006. [Conference Paper]

Iris Oved. (2009). Baptizing Meanings for Simple Concepts. *Proceedings of the Twentieth Midwest Artificial Intelligence and Cognitive Science Conference*. MAICS-09. Fort Wayne, Indiana. April18-19. Pg 104-109. [Conference Paper]