The Nature and Implementation of Representation in Biological Systems

by

Michael Collins

A dissertation submitted to the Graduate Faculty in Philosophy in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2010

© 2010

MICHAEL PATRICK COLLINS

All Rights Reserved

This manuscript has been read and accepted by the Graduate Faculty in Philosophy in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

11/12/09	Michael Devitt
Date	Chair of Examining Committee
11/12/09	Iakovos Vasiliou
Date	Executive Officer

Jesse Prinz

Samir Chopra

Michael Levin

Supervision Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

The Nature and Implementation of Representation in Biological Systems

by

Michael Collins

Advisor: Professor Jesse J. Prinz

In this dissertation I defend a theory of mental representation that satisfies naturalistic constraints. Briefly, we begin by distinguishing (i) what makes something a representation from (ii) given that a thing is a representation, what determines what it represents. Representations are states of biological organisms, so we should expect a unified theoretical framework for explaining both what it is to be a representation as well as what it is to be a heart or a kidney. I follow Millikan in explaining (i) in terms of teleofunction, explicated in terms of natural selection.

To explain (ii), we begin by recognizing that representational states do not have *content*, that is, they are neither *true* nor *false* except insofar as they both "point to" or "refer" to something, as well as "say" something regarding whatever it is they are about. To distinguish veridical from false representations, there must be a way for these separate aspects to come apart; hence, we explain (ii) by providing independent theories of what I call *f-reference* and *f-predication* (the 'f' simply connotes 'fundamental', to distinguish these things from their natural language counterparts).

Causal theories of representation typically founder on error, or on what Fodor has called the *disjunction problem*. Resemblance or isomorphism theories typically founder on what I've called the

non-uniqueness problem, which is that isomorphisms and resemblance are practically unconstrained and so representational content cannot be uniquely determined. These traditional problems provide the motivation for my theory, the *structural preservation theory*, as follows. F-reference, like reference, is a specific, asymmetric relation, as is causation. F-predication, like predication, is a non-specific relation, as predicates typically apply to many things, just as many relational systems can be isomorphic to any given relational system. Putting these observations together, a promising strategy is to explain f-reference via causal history and f-predication via something like isomorphism between relational systems.

This dissertation should be conceptualized as having three parts. After motivating and characterizing the problem in chapter 1, the first part is the negative project, where I review and critique Dretske's, Fodor's, and Millikan's theories in chapters 2-4. Second, I construct my theory about the *nature* of representation in chapter 5 and defend it from objections in chapter 6. In chapters 7-8, which constitute the third and final part, I address the question of how representation is *implemented* in biological systems. In chapter 7 I argue that single-cell intracortical recordings taken from awake Macaque monkeys performing a cognitive task provide empirical evidence for structural preservation theory, and in chapter 8 I use the empirical results to illustrate, clarify, and refine the theory.

Acknowledgments

It is customary to save one's most important acknowledgment for last; however, not being one to follow traditions for their own sake, I will put first things first. First and foremost, I wish to thank my fiancée, Sangeeta Nair. Without her intellectual and emotional support, her encouragement, understanding, and thoughtfulness, not only would this dissertation not be as it is, but I would not be who I am. Sangeeta, I dedicate this work to you, and only you know what that means to me.

I wish to thank my advisor, Prof. Jesse Prinz, both for intellectual advice on technical issues and professional guidance on broader issues. More importantly, Prof. Prinz has, perhaps unknowingly, been a role model to me, and I aspire to be as kind, reliable, good-humored, and thoughtful as he is. I also wish to thank Prof. Michael Levin, who has read, literally, thousands of pages of my work, and always managed to get back to me in less than a week with extremely detailed notes and incisive commentary. I have learned far more about academic writing from Prof. Levin's comments than I have in all of my undergraduate and graduate years combined.

I thank Prof. Samir Chopra, for kindly agreeing to join my committee on short notice, and then becoming a core committee member, entailing greater responsibilities, on even shorter notice. I also thank Prof. Michael Devitt, from whom I have learned a great deal over the years.

I wish to thank Prof. Peter Mandik, who was both a member of my dissertation committee as well as my undergraduate advisor. The project completed herein is one that I began thinking about many years ago as an undergraduate, under the tutelage of Prof. Mandik. The germ of the idea, that representation involves some combination of isomorphism and causation, was developed under his guidance, and I have no doubt that the idea originally came from him. More than that, it was Prof.

Mandik's infectious enthusiasm for combining empirical and conceptual research that set me on the path towards where I am today, for which I will always be grateful.

I wish to thank Prof. Rosamond Rhodes. Although she was not a member of my dissertation committee, Prof. Rhodes gave me a great deal of advice, assistance, and encouragement, on matters involving my dissertation and more. I consider Prof. Rhodes to be another role model, both personal and professional.

There have been many fellow graduate students with whom I have had a great deal of stimulating and challenging conversations over the years, including especially the members of two comprehensive examination study groups. I wish to thank one person in particular, my friend David Pereplyotchik, who is both an extraordinarily gifted philosopher as well as an uncommonly warm and sincere person.

I also thank my new little sister, Anu Nair, for a great deal of friendship and encouragement throughout this process.

Finally, I thank the rest of my family: Mom, Dad, Ed, Bridget, Will, Brian, and Michelle, but not, perhaps, for what they might expect. Over the years, but especially the last few, I have changed, and grown, a great deal. I suppose that is part of what the dissertation process is all about, anyway, as one does not come through it unchanged (or is it *unscathed*?). In subtle ways, each of my family members has signaled a willingness to let me be who I am, and who I want to be, now; that is, to see me as I see myself now. That shows a great deal of respect, and it is something for which I am grateful.

A Note on Gender

It is common practice to use the masculine 'he' and its cognates as gender-neutral pronouns, and to use 'Man' or 'Mankind' to refer to humanity in general. If this is truly gender-neutral, then there would be nothing odd about this: "Man is one of the many species that breast-feeds his young"[1].

There is nothing gender-neutral about using 'he' to mean 'he or she', nor is there anything neutral about using 'Man' to refer to humanity in general. The latter is especially pernicious: If we use 'Man' to refer to *humanity*, it would seem to follow that only *men* are important enough to qualify as part of humanity.

The fact that this is common practice is no justification for its continued existence. The first and sometimes hardest step towards eradicating morally unjustifiable prejudices is simply to recognize them. We cannot change what we do not see. Thus we must first recognize this practice for what it is: It is a manifestation of the all-too-prevalent sexist prejudice that has dominated human societies for thousands of years. That prejudice is so deeply ingrained that it is even part of our language.

In this dissertation, whenever possible I use gender-neutral language (such as 'person' or 'humanity' rather than 'he' or 'mankind'). When that is not possible, I use the feminine pronoun exclusively. The pendulum is now very far to one side, so far that it might seem odd to hear 'she' being used to mean 'he or she', but only familiarity can make this as natural as using 'he' for that purpose. It is a small step perhaps, but one worth taking.

[1] This is taken from Miller, C. & Swift, K. (1976), *Words and women* (Anchor Press/Doubleday), pp. 25-26.

Table of Contents

Chapter 1: Motivation and Characterization of the Problem

1.0 Introduction 1
1.1 Naturalism 2
1.2 Intentionality and Representation: Clarifying the Target Explanandum 3
    1.2.1 Intentionality 3
    1.2.2 Theories that Posit Representations 12
1.3 Challenges 18
    1.3.1 Why Use 'Representation'? 18
    1.3.2 Eliminative Materialism 30
    1.3.3 Dennett, Darwin, and the Distinction between Original and Derived Intentionality 35
    1.3.4 Horgan on Folk Psychology and Cognitive Science 46
1.4 The Plan 62

Chapter 2: Information and Representation

2.0 Introduction 65
2.1 Information Theory 66
2.2 Informational Content 71
2.3 Semantic Content and Belief 75
2.4 Critical Analysis: *Knowledge and the Flow of Information* 77
    2.4.1 The Learning Period and Idealization 77
    2.4.2 The Objectivity of Information 80
        2.4.2.1 Background Conditions Determine Probabilities 81
        2.4.2.2 The Distinction between Signal and Channel 84
        2.4.2.3 Objections and Clarifications 91

Chapter 3: Asymmetric Dependence Theory

3.0 Introduction 97
3.1 Asymmetric Dependence: Version 1 97
    3.1.1 *Only Xs cause 'X' 99
    3.1.2 *All Xs Cause 'X' 101
3.2 Version 1 Doesn't Work 105

3.3 Asymmetric Dependence: Versions 2 and 3 110
3.4 Versions 2 and 3 Don't Work 114

Chapter 4: Biological Categories, Teleofunction, and Teleosemantics

4.0 Introduction: Normativity at the Foundation of Representation 126
4.1 Biological Categories 127
4.1.1 A Preliminary Distinction 127
4.1.2 The Theory of Proper Functions 128
4.2 Millikan's Theory of Intentionality 133
4.2.1 Intentional Icons: Mapping Rules + Function 133
4.2.2 Articulateness of Intentional Icons and the Relation of Sense to Reference 140
4.2.3 The Mapping Rules 143
4.2.3.1 *Language, Thought, and Other Biological Categories* 143
4.2.3.2 *Varieties of Meaning* 147
4.3 Some Extant Critiques and Why They Don't Work 153
4.4 Critique of Millikan 158
4.4.1 Isomorphism 159
4.4.2 Local Information 166
4.4.3 Normativity 171
4.4.3.1 The Basic Claim 171
4.4.3.2 Distinguishing Kinds of Normativity 172
4.4.3.3 Distinguishing Failure to Represent from Representing Falsely 175

Chapter 5: The Nature of Representation I – Structural Preservation Theory

5.0 Introduction 181
5.1 Adequacy Conditions and Review of Explanandum 181
5.2 Foundations 183
5.2.1 Two Preliminary Distinctions 183
5.2.2 Truth and Structure 185
5.3 Theory Schema 194
5.4 Isomorphism and Structural Preservation 196
5.4.1 Distinguishing Picture Theory from System-Isomorphism 196
5.4.2 Measurement Theory 198
5.4.3 Representation Theorem 199
5.4.4 Uniqueness Theorem 202
5.4.5 Extensions and Relaxations 206

5.4.6 Empirical Axioms 210
5.4.7 Typing and Idealization 215
5.4.7.1 What Has Been Proven? 215
5.4.7.2 Typing Empirical Relational Systems 216
5.4.7.3 Idealization 220
5.4.8 Structural Preservation and Representation 223
5.4.9 A New No-Miracle Argument 233
5.5 Causal History and Nomically Grounded Causal Covariation 235
5.6 The Structural Preservation Theory of Original Representation 242

Chapter 6: The Nature of Representation II – Traditional Objections to Resemblance Theories

6.0 Introduction 246
6.1 Goodman's *Languages of Art* 247
6.2 Fodor's *Language of Thought* 255
6.3 Structural Preservation and the Syntactic Structure of Thought 262
6.4 The Causal Chain Problem 273

Chapter 7: The Implementation of Representation I – Evidence and Structural Preservation Theory

7.0 Introduction 277
7.1 Background: Assumptions and Hypotheses 278
7.2 The Relation of Theory to Evidence and the Dual-Approach Strategy 282
7.3 The Neurobiological Mechanisms of Vibrotactile Discrimination 286
7.3.1 The Implicit Theory and Its Critique 287
7.3.2 Review of Empirical Literature 289
7.4 Defense of the Representation and Vehicle Hypotheses 294
7.4.1 The Representation Hypothesis 295
7.4.2 The Vehicle Hypothesis 300
7.5 Applying SPT to the Brain: Claim 1 308
7.5.1 The Abduction 308
7.5.2 Sensory Representations 311
7.5.2.1 Peripheral Burst Code 311
7.5.2.2 Temporal Code in S1 317
7.5.2.3 Positively and Negatively Sloped Rate Codes in S2, PFC, VPC, and MPC 320
7.5.3 Motor Representations 327

7.5.4 The Role of the Empirical Literature 332

Chapter 8: The Implementation of Representation II – Clarifying and Refining Structural Preservation Theory

8.0 Introduction: Applying SPT to the Brain from the Perspective of Claim 2 342
8.1 Working Memory 343
8.2 Computation and Representation: A Comparison and Decision Procedure 345
8.2.1 Computation and Representation 345
8.2.2 A Possible Disconfirmation of SPT 347
8.2.3 The Computational Rule 351
8.3 Artificial Percepts and the Failure to Represent 355
8.4 Noise in Neural Systems 359
8.5 Objections and Clarifications 363
8.5.1 Clarifying the Roles of Each Component 363
8.5.2 Objections: Teleology and Covariance 366
8.5.3 Causal Efficacy 370
8.6 Scaling Up the Theory 374
8.7 Conclusion 380

Appendix A: Measurement Theory and Empirical Relational Systems

A.0 Introduction 382
A.1 Finite Domain 383
A.2 Infinite Countable Domain 393
A.3 Uncountable Domain 399
A.4 Summary of Results 407

Appendix B: The Cognitive and Neurobiological Mechanisms of Vibrotactile Discrimination

B.0 Introduction 410
B.1 The Task and its Psychophysics 410
B.2 Relevant Neuroanatomy 414
B.3 Sensory Encoding Mechanisms 418
B.3.1 Firing Rate and Periodicity in S1 419
B.3.2 Further Measures of Periodicity and Comparison of S1 and S2 423
B.3.3 Subpopulations in S2 428

B.3.4 Artificial Percepts Generated Through Cortical Microstimulation 429
B.4 Working Memory 432
B.5 Comparison and Decision Procedures 435
B.6 Motor Plans 439

Appendix C: Specifying Relational Systems by Neurometric Discrimination Thresholds

C.0 An Alternate Method of Typing Biological Relational Systems 442

Bibliography 449

Chapter 1: Motivation and Characterization of the Problem

1.0 Introduction

The goal of this dissertation is to articulate and defend a thesis about the nature of mental representation. Briefly, representational content is a structured relation involving two parts, and the explanation of how physical systems represent involves the preservation of internal structural relations and causal history. My goal for this opening chapter is to clarify and motivate the problem of explaining representation.

I begin with some comments on naturalism. With this general background in place, I clarify the target explanandum, distinguish intentionality from representation, and introduce various theories that make use of representational posits, although not all in the same way. I then discuss and reply to several objections to the projects of naturalizing intentionality and representation. Finally, I provide an outline of what is to come in the remainder of the dissertation.

1.1 Naturalism

Naturalism is a broad and somewhat imprecise thesis, whose principal ontological commitment is to the entities described by our basic physical sciences. The world consists of the distribution of matter and energy across spacetime, and physical objects are concatenations of the fundamental "building blocks" of matter (quarks, electrons, etc.). Essentially, the naturalist claims that the physical sciences constitute our best understanding of the world, and further that persons and their mental

states are just as much a part of that natural physical world as are protons and carbon molecules. The naturalist is a materialist, and seeks to understand how the mind and its properties are elements of the physical world[2].

Besides her ontology, a second characteristic of the naturalist is her methodology. We begin with the assumption that the mind and its processes are a part of the natural physical world and should be amenable to similar methodological strategies as those used in the sciences. Within that methodology there is a place for intuition and thought experiments, but as Devitt writes, "according to naturalism, semantics is an empirical science like any other. Intuitions and thought experiments do not have this central role elsewhere in science. Why should they in semantics?" (Devitt 1994, 545).

While the naturalist aims to understand mind and its properties in physical terms, there are two properties of mind that stubbornly resist physical characterization: consciousness and intentionality. The investigative approach I adopt, following many before me, is that of "divide and conquer". The hope is that it is possible to understand intentionality without a full-blown theory of subjectivity and qualitative consciousness as well. By way of motivating the problem then, to understand how mind is a part of the natural world, we must understand how it can bear an "aboutness" relation toward certain things, some of which it is physically or causally in contact with (as in veridical perception), others of which it is not (as in goals, desires, misrepresentation, etc.).

Fodor, for example, has remarked, "if the semantic and the intentional are real properties of things, it must be in virtue of their identity with (or maybe of their supervenience on?) properties that are themselves *neither* intentional *nor* semantic. If aboutness is real, it must be really something else" (Fodor 1987, 97). According to Fodor, "what we want at a minimum is something of the form *'R*

[2] Levin (1997, 87) concisely defines naturalism as "the view that all living things, including ourselves, were created by purposeless forces of Darwinian evolution".

represents S' is true iff C where the vocabulary in which condition *C* is couched contains neither intentional nor semantic expressions" (1984, reprinted in Fodor 1990, 32). While this is a good starting point, it is not enough for an adequate preliminary characterization of the project.

1.2 Intentionality and Representation: Clarifying the Target Explanandum

As a basic characterization, intentionality is the property of being directed towards or of something, or, it is aboutness. Dennett (1983, reprinted in Dennett 1987, 240) for example writes, "Intentionality … is – in a word – *aboutness*".[3] Traditionally, there are more properties associated with intentionality than simply aboutness. In this section I discuss those properties as well as some of the theoretical enterprises that make use of intentional states or representational states as ontological posits. My goal in this section, as in this chapter more generally, is to provide a precise characterization of my project and the methodology I propose for tackling it. "To naturalize intentionality" or "to understand mind in the physical world" is hopelessly unconstrained, so we must be more careful about describing the questions that I seek to answer.

1.2.1 Intentionality

The word 'intentionality' originated with Brentano (1874); he considered intentionality to be a characteristic mark of the mental. In this subsection I catalogue the properties that have come to be

[3] This was a fairly random selection; many authors say the same or similar things. Searle (1983, p. 1) writes, "As a preliminary formulation we might say: Intentionality is that property of mental states and events by which they are directed at or about or of objects and states of affairs in the world". Fodor (1987) writes on p. 97 "I suppose that sooner or later the physicists will complete the catalogue they've been compiling of the ultimate and irreducible properties of things. When they do, the likes of *spin*, *charm*, and *charge* will perhaps appear on their list. But *aboutness* surely won't; intentionality simply doesn't go that deep".

associated with intentionality in the literature. While the most often cited characteristic of intentionality is aboutness, there are others.

Millikan (1984) writes, "the traditional earmark of the intentional is the puzzle that what is intentional apparently stands in relation to something else – that which it *intends* or *means* or *means to do* or *is meant to do* – which something can be described, yet which something may or may not *be*". In other words, intentionality is a relation one of whose relata need not exist. In a later writing (2004), she argues that Brentano's term 'intentionality' was actually characterized in two different ways. The first is as the property of aboutness. The second characterization involves what he called "intentional inexistence". This is a peculiarity of the objects of intentional states (that is, the contents or what intentional states are about), that they can be thought about even though they need not exist. By conflating these two interpretations, we have been left with the idea that "to explain how a representation could be of or about something is just the other side of the coin of explaining how it could be empty or false" (2004, 64). Millikan argues that we should keep these separate, and provide different explanations for each.

Dretske (1995, 28-34) provides a helpful list of some of the characteristics associated with intentionality. The first on the list is the ability to misrepresent. A state is not intentional if it does not have the capacity to "say" something false. For example, I can have beliefs that aren't true.

The second characteristic on the list is aboutness, which I mentioned above. The fourth, which I'll mention out of order here, is based on Husserl's term *noema*, and which Miller (1984) calls *directedness*. Dretske argues that that the concept of directedness, if it is coherent at all, is simply aboutness.

The third characteristic is *aspectual shape*. This term originated with Searle (1992), although the idea, as we'll see, traces back (at least) to Frege. In thinking about a thing, we tend to think about it "under an aspect"; that is, with respect to its properties. For example, I can think of a ball with respect to its roundness and its color. "Our mental states not only have a reference … they represent that object in one way rather than another. When on object is represented, there is always an aspect under which it is represented" (1995, 31). A further component of the notion of aspectual shape is the fine-grained individuation of the aspects under which objects get represented. For example, I can think about Venus under the aspect of its being the earliest star visible in the morning (*as* the Morning Star) without thinking about it under the aspect of its being the latest star visible in the evening (*as* the Evening Star), even though the Morning Star is the Evening Star, which is the planet Venus.

Warfield and Stich write (Stich and Warfield 1994, 3-8) that any theory of content[4] must satisfy some minimal constraints, including that it must be naturalistic (in roughly the sense adumbrated above), it must explain misrepresentation, and it must explain *fine-grained meanings*. The idea of the fine-grainedness of meaning (or of content, or of intentionality) is this. I can desire to visit the birthplace of Ben Franklin while lacking the desire to visit the birthplace of the inventor of bifocals. I can (sincerely and seemingly without contradiction) assent to sentences containing 'Mark Twain' while not assenting to the same sentence with 'Samuel Clemens' substituted for 'Mark Twain'. Semantic or intentional properties seemingly cut the world into a finer "grain" than do physical or even modal properties: The same can be said about my assent and refusal to assent to sentences containing 'square' or 'four-sided equiangular closed planar figure' as can be said about 'Mark Twain' and 'Samuel

[4] There are many different phrases floating around the literature, which may or may not mean the same. Is a theory of *intentionality* the same as a theory of *representation*, the same as a theory of *content*? Is being intentional the same as having semantic properties? Are meaning and aboutness the same thing? Are representation and meaning the same? Is there a distinction between reference and representation? I'll not try to provide scholarly analyses of what other people have said for every one of these questions (although in many cases I will), but I will answer them myself.

Clemens'. The problem of intentionality, it seems, is more than the problem of explaining how one state of the universe could be about another. It is also the problem of saying how one state of the universe could be about another *under an aspect* but *not* about the same state of affairs *under a different aspect*.

Concurrent with the fine-grainedness or aspectual shape of intentional states, are the logical properties that sentences about intentional states take on. Sentences about intentional states exhibit *intensionality* (with an 's'), or, *referential opacity*. That is, sentences about intentional states use clauses for which the usual rules of substitution do not hold. Substitution of co-referential terms, in general, preserves the truth value of a sentence. But sentences about intentional states generate intensional contexts, in which co-referential substitution is not necessarily truth-preserving. For example, 'Mark Twain' and 'Samuel Clemens' are co-referential. The sentence, "Mark Twain is the author of *Huckleberry Finn*" would still be true if 'Samuel Clemens' were substituted for 'Mark Twain'. However, within an intensional context, generated by the intentional idiom, that no longer holds: "Fred believes that Mark Twain is the author of *Huckleberry Finn*", if true, does not necessarily remain true upon substituting 'Samuel Clemens' for its co-referential term. Dennett (1983, reprinted in his 1987, 240) claims that the generation of intensional contexts is the *first* mark of intentionality.

Prinz (2002) takes the properties traditionally associated with intentionality and splits them into two categories. *Intentional content*, for Prinz, is reference or standing in for: "Concepts represent, stand in for, or refer to things other than themselves. My AARDVARK concept is about aardvarks; it refers to all and only aardvarks. Philosophers call this property 'intentionality'" (2002, 3). *Cognitive content*, on the other hand, involves the sort of fine-grainedness mentioned above. "We do, however, need *some* kind of content other than reference, or intentional content, as it was called in the last section. I call this further requirement, 'cognitive content'. Cognitive content is what allows two coreferential representations, be they terms or concepts, to seem semantically distinct to a cognitive agent" (p. 7).

The motivation for this distinction between intentional content (mere aboutness) and cognitive content, which is what allows coreferential representations to at least *seem* different, comes from Frege's work in the philosophy of language (Frege 1952), on the difference between sense and reference.

First, statements involving an identity claim using two co-referential terms can be informative. If you hadn't already known it, the statement "Mark Twain is Samuel Clemens" would be informative to you; you would learn something new. This presents a puzzle, since the claim that Mark Twain is Mark Twain is not informative, yet 'Twain' and 'Clemens' refer to the same person. Second, as mentioned above, the introduction of intentional or semantic terms into sentences (such as 'believes that', 'said that', etc.) results in the failure of substitutivity of co-referential terms. Frege's solution was to distinguish sense from reference. While it is not abundantly clear what Frege's term 'sense' amounts to, it is this basic phenomenon that underlies Prinz's cognitive content, Searle's aspectual shape[5], and the fine-grainedness of meaning or intentional content. Further, this phenomenon is what generates referential opacity.

Dennett (1983, reprinted in his 1987, 242) mentions a further requirement. The use of the intentional idiom with respect to some system or agent carries with it a presupposition of minimal rationality. Exactly what this amounts to is not important for present purposes, but it should be mentioned that this introduces a normative component to intentionality. Being rational involves something like: if one believes that *p* and one believes that *p* implies *q* then one *ought* to believe that *q*.

[5] Searle's aspectual shape involves two components. One is the fine-grainedness component and the other is that representations always represent "under an aspect". So far as I understand this, it merely amounts to the fact that representations represent some object *as* having one property or another, and that simply amounts to representations having some content or another. But if a representation didn't have a content then it wouldn't be a representation (I'll have to qualify that statement later on) so we should consider aspectual shape to be, at the least, very closely related to fine-grainedness.

There are various related properties traditionally associated with intentionality: aboutness or directedness, capacity for misrepresentation, a relation one of whose relata need not exist, fine-grained contents, aspectual shape, cognitive content vs. intentional content, the generation of referential opacity, and minimal rationality. These properties are usually associated with what are known as *propositional attitudes*, such as beliefs and desires, but they are also associated with meaningful sentences.

There are some further complications to a preliminary characterization of my project. What is the relationship of intentionality to representation, or representation to reference? What, exactly, do philosophers mean when they say 'content'? What is the relationship of language to thought, and the intentional or semantic properties of each? Is there an original/derived distinction? What is the relationship of folk psychology or other theoretical enterprises to all of this? I'll make some brief comments on each of these in turn, starting with the notion of content.

Frequently philosophers say that the content of a representation is what the representation is about. But this is ambiguous: Suppose some animal has a perceptual representation of its ambient chemical environment. We might then say that the content of its representation just is the chemical environment in which the animal is situated. But suppose that I have a belief that Mark Twain is the author of *Huckleberry Finn*. Is the content of my belief the state of affairs in which Mark Twain has the property of being the author of *Huckleberry Finn*, or is it perhaps some mediating "content" (where now the content is something *other than* Mark Twain's being the author of *Huckleberry Finn*)? What about my belief that unicorns fail to have three horns? Is the content the state of affairs in which unicorns *lack* three horns? What state of world affairs is that? In other words, is representation a binary relation, from representation to represented state in the world, or is it a ternary relation involving representation, propositional content (whatever that is), and some state of affairs in the world, or even

something else? Since I'm looking for an explanation of the most fundamental kind of representation, I presuppose a "'Fido'-Fido", direct reference theory in which representation is a binary relation from representation to represented state in the world. Ultimately, I'm aiming at a solid naturalistic foundation upon which theories of less fundamental kinds of representation can be built, for which the direct reference approach might seem inappropriate.

It is widely but not universally accepted that the intentional content of language is dependent on the intentional content of thought but not vice versa. That is, the content of language *expresses* the content of thought. Conventional signs such as linguistic utterances, stop signs, hand gestures, and so on, each have their meaning only in a derivative fashion. Their meaning is derived from the intentional content of thought. While both these "conventional" signs and non-conventional signs (presumably, *mental* representations) have aboutness, there seems to be an important difference between mental and non-mental representations. Linguistic items and other conventional signs have their aboutness in virtue of the aboutness of psychological states, whereas the aboutness of psychological states is not ontologically dependent on anything else. Haugeland (1985, 25) describes it thus:

> The basic question is: How can thought parcels *mean* anything? The analogy with spoken or written symbols is no help here, since the meanings of these are already *derivative* from the meanings of thoughts. That is, the meaningfulness of words depends on the prior meaningfulness of our thinking…

I accept the original/derived distinction, which I describe as follows. Representations that have content derivatively have their content in virtue of bearing some suitable relation to other intentional

states, whereas representations that have original or non-derived intentionality do not have their content in virtue of bearing a suitable relation to another state that has intentionality. In counterfactual terms, if there were no representations with original intentional content, then there would be no representations with non-original content, although there might be representations with original intentional content in the absence of representations with non-original content. The intentionality of representations with derived content is thus *ontologically dependent* on the existence of representations with original intentional content. The fundamental task, clearly, is to explain the nature of original intentionality.

Notice that I did not say that *linguistic states* have derived intentionality while *mental states* have original intentionality. I remain neutral on that. I'll have little to say specifically about language in this work. However, I do want to mark the prima facie plausible distinction between the kinds of representations that are *basic* and those that aren't. I'll have a little more to say on this in section 1.3.3 when I discuss Dennett's objections to the original/derived distinction.

What is the relationship between representation, intentionality, and the several theoretical endeavors that posit them[6]? Cummins (1989) distinguishes representational content from intentional content. For Cummins (1989, 14), "A system with intentionality is just a system with ordinary propositional attitudes (belief, desire, and so on). Thus construed, intentionality is a commonplace, and so is intentional content". However, he leaves open the possibility that intentional characterizations of physical systems is much like what Dennett says it is. That is, it is a characterization of a system and its behavior *as a whole*, which does not necessarily pick out discrete states with determinate intentional contents. *Representation*, by contrast, is a theoretical construct, not part of our ordinary commonsense

[6] I also mentioned this: What is the relationship of reference to representation? Answering that is part of the theory I'll present, so I won't be discussing it in the cursory style I'm using here to set up the preliminaries.

way of viewing the world and each other, in the way that beliefs and desires are. Further, in the context of different theories, there are different constraints on the nature of representation, as determined by the work that the theory needs it to do. Cummins' goal in his 1989 is to provide an analysis of the properties representations must have, given the classical computational theory of cognition. Hence, he is interested in naturalizing representation, but considers representation only from within the context of a theory that uses it.

Fodor's view is related. He also thinks that representation and intentionality are distinct. However, he argues that the Representational Theory of Mind, a hodge-podge of various theses, is the only remotely plausible theory of mind (Fodor 1975, 1987, 1998). Part of that theory is the claim that intentionality, that is, the aboutness associated with the ordinary propositional attitudes of folk psychology, reduces to representational content, where representations are *symbols* in a language of thought. For Fodor, cognitive science vindicates folk psychology (1987). Cummins prefers to call this thesis the Representational Theory of Intentionality (rather than 'Mind'), because it makes the connection between intentionality and representation explicit.

Cummins' methodological suggestion, that we should seek to understand the nature of representation from within the context of the theories that posit it, is insightful and very important. We should examine not only the insights gained through reflection on language and our ordinary way of viewing the world and each other (our folk psychology), but also the more explicitly worked out theories or theoretical frameworks that use the concept of representation.

Here's another way of making Cummins' methodological point. Quine (1948) gave us perhaps the most lucid manner of affirming an ontology. We ought to posit all and only those entities in the domain of variables that are bound by an existential quantifier in a successful theory, where the success

of a theory is determined by the usual theoretical virtues. Given this insight, when we approach the problem of representation, it makes sense to do so in the context of the theory that posits them.

However, representations are the explanatory posits of several theories. The situation is further complicated by the fact that one of those explanatory enterprises forms part of our commonsense understanding of ourselves. In what follows I discuss several different notions of representation as they are used in different theories. I argue that, while there are indeed important differences among them, there is also a common nucleus, and it is this nucleus that is worthy of philosophical investigation and in need of a naturalistic explanation. Further, this common core seems to underlie much (but not all) of the traditional ways of thinking about intentionality discussed above. The outcome of the following discussion is that I will make a clear distinction between intentionality and representation, and it is *not* intentionality that I am interested in.

1.2.2 Theories that Posit Representations

There are at least five theoretical enterprises that posit representations. I begin with folk psychology. We generally understand, explain, predict and manipulate each other's behavior by means of the mental states, and in particular the propositional attitudes, that cause it. I went to the store because I wanted to obtain some milk, and I believed that by going to the store I could buy milk. This explanation trades on my *propositional attitudes*: I desired that I have milk, I believed that by going to the store I could obtain it, and those two mental states combined are causally efficacious in the production of my behavior. My behavior could also have been predicted or manipulated if one knew that I had a desire for milk.

The propositional attitudes posited by this type of explanation have the following properties. First, they are about or directed towards some state of affairs: my desire that I have milk is about the

state of affairs of my having milk. Second, they are causally efficacious in the production of behavior. Third, they are relevant to the explanation of that behavior (partially in virtue of their causal efficacy, but note that the content plays a crucial role in the explanation as well).

A second enterprise that posits representational states is philosophical semantics, or, the attempt to understand *meaning*. Meaning is paradigmatically a property of linguistic utterances. My spoken and written sentences have meanings, and the philosophical task is to explain that. A fairly standard (though not universally accepted) tactic is to posit meaningful thoughts, where meaningful sentences express thoughts that have meaning. This explanatory enterprise is related to but distinct from folk psychology, because the major concern of the latter is the explanation of behavior, while the major concern of the former is an explanation of meaning. However the posits of folk psychology and philosophical semantic s are, for the most part, the same. They both posit propositional attitudes such as beliefs and desires, which are meaningful psychological states efficacious in the causation of behavior, and have aboutness.

Some further properties of the propositional attitudes may be noted in this context. First, propositional attitudes seem to be structured in some way. Like sentences, both the belief and desire that I have milk have the content that I have milk, or, are about the state of affairs of my having milk. They have a different "mode", however. The different modes are individuated in terms of the functional role that the mental state plays with respect to the content: My belief that I have milk and my desire that I have milk have a different "direction of fit" (Searle 1983). In sentences, the different modes are determined by the sentence structure itself, with "illocutionary force" (Austin 1975) serving as parallel to the type of attitude one takes towards a proposition in thought.

Another aspect of meaning is its fineness of grain. As discussed above, semantic properties cut the world into a finer grain than seemingly any other kind of property, including physical and perhaps even modal properties.

Meaningful thoughts or mental representations are posited to explain the meaningfulness of public languages, gestures, etc. Since languages have these properties (fineness of grain, mode/content distinction, and others, such as productivity and systematicity[7]) it is generally thought that those properties derive from the thoughts which sentences express. The meaningful mental representations posited by philosophical semantics are basically the same entities as those posited by folk psychology. The difference is that folk psychology marshals them for use in the explanation of behavior, while semantics posits them to explain meaning.

Next, let's consider the computational theory of cognition, or the classical model of cognition as computation. On this view, cognition is not merely modeled by computation or computer models, it *is* a species of computation. Computation is the rule-governed manipulation of symbol tokens according to quasi-linguistic rules. Multiple realizability is a critical element of this view: the abstract computational structure defines the elements, not any particular substrate that implements them. The key idea is that it is possible for the semantics, or the aboutness of the symbols, to supervene on the syntax, without the semantics actually being causally efficacious in the manipulation of the symbol. It is only the syntax that is so. However, the semantics "comes along for the ride" so to speak, so that, given the right syntactical structures and manipulations, it is possible to have meaningful symbol tokens that, if they start out true, upon symbol manipulations, remain true. Truth-preserving inferential rules can be defined solely in terms of the syntax, where the syntactical properties supervene on physical properties.

[7] Productivity is the ability to generate new sentences with distinct meanings. Systematicity is the systematic structuring of sentences in such a way that the person who understands one sentence can understand a restructured but distinct sentence that uses the same sentence parts.

The representations posited in this theoretical enterprise are *symbols*: they are syntactically structured, recursively definable entities that can be manipulated according to formal rules.

A competing view is the connectionist theory of computation and cognition. The foundation of this view is to use simple, neuron-like "nodes", connected to each other via inhibitory or excitatory links, to model cognitive processes. The network is defined in terms of the nodes, their connections, and the weights or strengths with which an incoming signal affects its connected node. The activity of the network is defined in terms of activation vectors, which are lists of numbers that define the activation level of each node in the network. Representations in these networks are thought of as *activation vectors*; hence, they are distributed throughout the network and are continually updated over time. Representations are also sometimes thought of as the *weight vectors*, which are lists of numbers that describe the strengths of the inter-node connections. They are apparently not syntactically structured, nor do they operate according to formally defined rules.

Finally we have the several branches of the neurosciences, including ethology. Talk of representations and detectors is prevalent in these disciplines, in various ways. One is in terms of anatomical location, and the connectivity that that location bears to either sensory or motor systems. In primary somatosensory cortex for example, the topographical arrangement of neurons respects adjacency relations to receptors in the skin, so that, for example, the area that is connected to receptors in the hand is adjacent to the area that is connected to receptors in the arm. There are literally dozens of these topographic maps, with several for each sensory modality and some for combined sensory modalities, as well as maps that respect the adjacency relations for the motor systems. There are egocentrically defined spaces, such as retino-centric and head-centered maps, and there are allocentrically defined spaces as in the hippocampus, which apparently has maps defined with respect to

objective coordinate spaces. Each of these spaces is called, for example, the "hand representation area", the "face representation area", etc.

Another type of representation posited here is in terms of differential activation of particular cells in response to specific types of stimuli. What are known as *edge-detectors* in primary visual cortex respond selectively to contrast lines in their receptive fields. Some respond selectively to bright stimuli in the center of their receptive field, when combined with dark on the outside edges, and others respond to the inversion of that. There are cells that respond selectively to faces, to motion, color, etc. There are many more specializations, where particular cells or groups of cells fire in coordinated ways either in response to a specific type of stimulus, or in order to effect a specific type of muscle output.

Each of the above theoretical endeavors posits representations, and their doing so places different constraints on the properties that representations must have to play the roles that they are purported to. Folk psychology and philosophical semantics explicitly posit *beliefs* and *desires* and other propositional attitudes. These states must have propositional content (granting that it is unclear what that is), be capable of error, and be causally efficacious in the production of intelligent, goal-directed, or rational behavior. Some further properties attributed to propositional attitudes include fine-grained content and a distinction between mode and content. Classical computational theories of cognition posit language-like, syntactically structured and recursively definable symbols, whose syntax governs their manipulations, yet whose semantics respects the syntax, in such a way that syntactic manipulations can be truth-preserving. Connectionist theories posit distributed representations, which seemingly are not structured in the way that the syntactical symbols of the classical theory are, and are not manipulated according to formal rules. However, the classical computation and connectionist theories have in common that the representations they posit have "aboutness" or "directedness": The representations point outside of themselves to something else, and are subject to error or

misrepresentation. It is at least consistent with their home theory to call them causally efficacious in both cases, although that doesn't necessarily come up in the context of these theories. With the classical theory, the representation is causally efficacious, but solely in virtue of its syntax, however, the semantics respects that syntax. In connectionism, which is supposed to be at least broadly a biologically plausible model of brain function, the output levels are thought of as connected to effector muscles, and hence, the representations are causally efficacious there as well. Finally, the representations posited by the neurosciences are related to all of the above in that activity in particular areas of the brain is thought to be about or point towards activity or stimulation at some particular part of the body or some egocentrically defined point in space. As above, these representations can be in error, and they are causally efficacious in the production of behavior. Issues of fineness of grain and the distinction between mode and content seem to be specific to philosophical semantics and, to a lesser extent, folk psychology.

The representations in each of the theories share a *common core* with those of every other. Representations, in all of the above theories, are about, of, or directed towards something outside of themselves. Misrepresentation or error is possible for all of them, and they are causally efficacious in the production of behavior.

Following Cummins (1989), I hereby reserve use of the word 'intentionality' for aboutness and other properties associated with the folk psychological propositional attitudes. *Representations*, by contrast, are theoretically motivated posits, which, in every theoretical context that posits them, have aboutness, the possibility of error, and causal efficacy in the production of behavior. I am interested in providing a naturalistic explanation of *the common core*. Let us consider some challenges to my project and in so doing, we will further clarify its scope, aims, and methodology.

1.3 Challenges

1.3.1 Why Use 'Representation'?

Cummins (1989) has argued that the notion of representation is a theoretical construct, whereas intentionality is a commonplace. However, clearly, 'intentionality' is a technical term, and 'representation' has ordinary currency. Additionally, Cummins' project is to explain the concept of representation as it is used in classical cognitive science. The concepts of representation found in neuroscience and connectionism, however, would seem to be very different than either intentionality or the symbolic data structures of cognitive science. Why then, would I use 'representation' to describe my explanandum?

The implicit challenge in this question is not a trivial terminological dispute. Ramsey (2007) illustrates this point as follows. Imagine someone giving a representational theory of disease. Upon analysis, it turns out that that theorist is using the word 'representation' in a way that has nothing in common with the ordinary usage, and instead is using 'representation' to simply refer to infectious agents. The reply that this is a theoretical construct, and hence all that matters is whether the "representational" posits do explanatory work, is not sound:

> This would not be a case where a technical notion of representation is playing some explanatory role. Instead, this would be a scenario where a notion of representation would not be playing *any* explanatory role; it would be completely absent from the theory. All of the work would be done by ordinary notions of infectious agents. This is because there is nothing about the job these states are doing that is intuitively recognizable as representational in nature. Unless a

> posit is in *some* way grounded in our ordinary understanding of representation, it is simply not a representational posit, in any sense (Ramsey 2007, 13-14).

While there can certainly be technical notions of representation that depart from our commonsense use of 'representation', nonetheless, to *be representational posits*, those technical notions must match up with commonsense to some (unspecified) degree. For a different example, I may stipulate that I'll use 'Santa Claus' to refer to the jar of peanut butter in my cabinet. However, when I attempt to prove the existence of Santa Claus by making you a peanut butter and jelly sandwich, we may all rightly agree that *that's* not what we mean when we say 'Santa Claus', and thus I have not in fact proven Santa's existence.

Something similar to the Santa/peanut butter example has occurred in the evolution of cognitive science, Ramsey argues. While classical computational cognitive science makes use of representational posits, the "representations" posited by neuroscience and connectionism are not really representations at all. Rather, researchers in this field are simply misusing the term. Since I am interested in naturalizing the semantics for the representations in cognitive science *and* neuroscience, (admittedly, with an emphasis on the neuroscience) it would seem that my project is ill-considered, because there are no semantics to be naturalized for neuroscience: That field doesn't actually make representational posits. Thus we return to this section's titular question: Why use 'representation'?

I begin by noting my qualified agreement with Ramsey. I cannot prove the existence of Santa Claus by stipulating that 'Santa Claus' refers to something which ordinary usage does not license. Nonetheless, just what "ordinary usage" amounts to is imprecise and open for interpretation and re-interpretation. Caution is advised whenever a philosopher claims to know what "the common person"

means by 'X', and then reaches sweeping philosophical conclusions justified by the assertion that her opponent's philosophical claims are not licensed by the common person's usage of 'X'. Nonetheless, Ramsey is right that if a theory that claims to posit representations uses the word in a way that has *nothing* in common with ordinary usage, then it would turn out that that theory does not make use of representational posits. A whole lot turns on an analysis of "commonsense" or "ordinary usage". I'll argue in the next few paragraphs that Ramsey's arguments are flawed because they have nothing to do with commonsense. As a result, his argument that neuroscience does not posit representations is not sound. Additionally, I argue that the representations of neuroscience and connectionism deserve the name just as much as the representations of cognitive science.

Ramey's analysis of the *commonsense* notion of representation is very similar to what I have argued is the core concept of representation common to the several *theoretical* approaches that use it. Within folk psychology, he argues, the key element of mental representation is intentionality, which he identifies with aboutness[8]. Additionally, he claims that intentionality is original or non-derived, as opposed to the derived intentionality of linguistic signs, and that this is also part of ordinary usage. He mentions the generation of referential opacity, the capacity for error, and that intentionality is a relation one of whose relata need not exist, each as key features of intentionality licensed by our ordinary manner of speaking (2007, 16-17). In addition to intentionality and its features just mentioned, Ramsey argues that causal efficacy is part of the commonsense concept of mental representations. "Beyond these mundane observations about the intentionality and causality of mental representations, what little consensus there is about our commonsense picture of mentality begins to evaporate" (2007, 19).

[8] "Intentionality (in this context) refers to 'aboutness'. Thoughts, desires, ideas, experiences, etc. all *point to* other things, though they could also, it seems, point to themselves. Intentionality is this feature of pointing, or designating, or being about something. Typically, mental representations are about a variety of types of things, including properties, abstract entities, individuals, relations, and states of affairs" (Ramsey 2007, 16).

Ramsey also provides an analysis of the "ordinary" concept of non-mental representation. Ultimately his analysis is, to be a (non-mental) representation a thing must be *used* by a cognitive agent *as a representation*. He bases his analysis of our commonsense use of 'representation' on Pierce (1931-1958). Ramsey accepts Pierce's analysis, which is that "there can be no meaning or representational content unless there is some thing or someone *for whom* the sign is meaningful … What is significant about Pierce's triadic analysis is the idea that representations are things that are *used in a certain way*" (Ramsey 2007, 22-23).

From these considerations, Ramsey arrives at a crucial element of his book, the *job description challenge*:

> There needs to be some unique role or set of causal relations that warrants our saying some structure or state serves a representational function. These roles and relations should enable us to distinguish the representational from the non-representational and should provide us with conditions that delineate the sort of job representations perform, *qua* representations, in a physical system … What we want is a job description that tells us what it is for something to function as a representation in a physical system (Ramsey 2007, 27).

From the job description challenge, Ramsey concludes that the representations of classical cognitive science are indeed representations, but the "representations" of connectionism and neuroscience are not.

Ramsey's analysis is flawed at several points. First, I am amenable to Ramsey's characterization of intentionality and its various properties. However, the generation of referential opacity, the possibility of a relation one of whose relata need not exist, and so forth, are associated with the technical philosophical concept of *intentionality*. It stretches credulity beyond its limit to claim that this is part of the "commonsense" concept, or that an analysis of "ordinary use" leads to these features, when in fact it took many years of careful philosophical investigation to discover and explicate them. Similarly, Ramsey's analysis of the concept of non-mental representation, which plays an essential role in generating the job description challenge, is based on technical philosophical literature (namely, the work of Pierce). To be clear, the philosophical literature is indeed the appropriate place to look in order to gain an understanding of these concepts, and that is what I have done. But it is disingenuous to claim that the average non-philosopher has any of these ideas in mind when she uses the word 'representation'.

Second, Ramsey's analyses amount to a conceptual analysis of the concept of representation, which generates a necessary condition on being a representation (even though he explicitly repudiates conceptual analysis – cf. pp. 8-14). For Ramsey, a thing is not a representation unless that thing functions as a representation. What it is to "function as a representation" is left to intuition:

> If we want to evaluate the different notions of representation posited in scientific theories, a more promising tack [than posing sufficient conditions, generating counterexamples, etc.] would be to carefully examine the different notions of representation that appear in cognitive theories, get as clear as possible about just what serving as a representation in this way amounts to, and then simply ask ourselves – is this thing really functioning in a way that is recognizably

> representational in nature? In other words, instead of trying to compare representational posits against some sort of contrived definition, we can instead compare it directly to whatever complex concept(s) we possess to see what sort of categorization judgment is produced … To some degree, this means our analysis will depend on a judgment call (Ramsey 2007, 10).

His analysis of the commonsense concept of representation leads to the conclusion that only things that function as representations, that is, that *do something* that is "recognizably representational in nature" can *be* representations. Since the "commonsense" concept led us this far, we should think that commonsense provides us with some guidance on what it would be for a thing to serve as a representation, or, could tell us something about what it is that we recognize when we recognize a thing as playing a representational role. On this, Ramsey says

> our ordinary notion of mental representation leaves unexplained a great deal of what a theory-builder should explain about how something actually serves as a representation … Commonsense psychology provides us with little more than a crude outline of mental representations and leaves unanswered several important questions about how representations drive cognition (Ramsey 2007, 20).

If the commonsense concept is to be our guide in discovering which theoretical posits are representational and which aren't, then it follows that commonsense should tell us what it is for a thing to serve as a representation. But, according to Ramsey, it doesn't. The only thing left to demarcate the

representational from the non-representational are the intuitions of individual philosophers; Ramsey doesn't even claim that these are common intuitions or part of the ordinary concept. Thus, Ramsey makes something's being a representation depend on its functioning as a representation, and then explicitly notes that commonsense has nothing to say on what it is to function as a representation. *Nonetheless* Ramsey's intuitions get to tell us what is and what isn't a representation. That is not a way of getting at the truth.

Finally, Ramsey has claimed, correctly, that there is "little consensus" on what commonsense says about representations or mental representations: "[other than intentionality and causality] what *little consensus* there is about our commonsense picture of mentality begins to evaporate" (Ramsey 2007, 19, my emphasis). While it is correct that there is little consensus on what commonsense says about this, it follows immediately that *it isn't commonsense*. If it were a part of the common, ordinary usage, there would be a great deal of consensus. For example, my "Santa is peanut butter" argument doesn't work precisely because there is a great deal of consensus surrounding the concept of Santa Claus. Everyone agrees that the concept involves a man that lives at the North Pole and distributes gifts on Christmas. But whether mental representation involves the generation of referential opacity, or requires the possibility of error (both of which Ramsey has claimed are part of ordinary usage), are technical philosophical debates about which commonsense and ordinary usage do not pronounce.

While Ramsey is misguided in some of his arguments, his basic point still stands. For a theoretical posit to be a representational posit, it should have something in common with the standard ways that 'representation' is used. Ramsey's analysis of the commonsense concept of mental representation is almost perfectly in line with what I have called the core theoretical concept of representation. Where Ramsey is mistaken, is in claiming that these features are part of the ordinary

usage of 'representation'. As I mentioned above, a whole lot rides on just what ordinary use amounts to.

The only authority on the common usage of any term is the dictionary. Merriam-Webster defines 'represent' as follows (from [http://www.merriam-webster.com/dictionary/represent]):

1. To bring clearly before the mind <a book which *represents* the character of early America>
2. To serve as a sign or symbol of <the flag *represents* our country>
3. To portray or exhibit in art
4. To serve as the counterpart or image of <a movie hero who *represents* the ideals of the culture>
5. To produce on the stage; to act the part or role of
6. To take the place of in some respect; to act in the place of or for usually by legal right; to manage the legal and business affairs <athletes *represented* by top lawyers and agents>
7. To serve especially in a legislative body by delegated authority usually resulting from election
8. To describe as having a specified character or quality <*represents* himself as a friend>
9. To give one's impression and judgment of: state in a manner intended to affect action or judgment; to point out in remonstrance or protest
10. To form an image or representation of in the mind; to apprehend (an object) by means of an idea; to recall in memory
11. To correspond to in essence
12. To make representations against something; to protest [intransitive verb form]
13. *Slang*: to perform a task or duty admirably; serve as an outstanding example

Common to most of these senses of 'represent' are the following two things. First, there is the notion of surrogacy, proxy, or standing-in-for. This is common to the legal notion of representation, as in a legislative body or an attorney or agent acting in the stead of some other party, as well as the notion of an actor on stage. Second, there is a notion of aboutness, directedness, or pointing implicit in many of the above senses. For example, in serving as a sign or symbol *of* a thing (such as a flag), or in exhibiting or portraying something (as in a piece of art), and perhaps as well, in forming an image or memory *of*, there is the idea that whatever represents, *points towards* something. Thus, when we ask if

a technical use of 'representation' accords with ordinary usage, we should inquire if it is sufficiently similar to any of the dictionary uses. A reasonable claim would be that the ordinary use of 'representation' with which we are interested involves something like the notion of a proxy or surrogate, *or* involves the idea of aboutness or pointing. Other elements such as fine-grainedness, recursively defined symbols etc. are, while legitimate, technical uses and thus *not* the ordinary usage.

Ramsey has claimed that the purportedly representational states posited by neuroscience and connectionism are not representations because, ultimately, they do not fit his intuition on what it is to play a representational role. While I have argued that his reasoning is flawed, nonetheless, this is not an uncommon challenge. To meet it, let us begin by considering orthodox computationalism.

Few philosophers challenge the claim that the representational posits of classical computational cognitive science are representations on the grounds that the cognitive scientists' use of 'representation' does not accord with ordinary usage. However, considering what ordinary usage (that is, the dictionary) actually licenses, it would seem that the ordinary language argument would cut equally against both cognitive science and neuroscience as representational theories. The dictionary entries make no mention of recursively definable symbolic data structures, whose semantics supervene on their syntax. This is, emphatically, *not* ordinary usage. This use of 'representation' is just as far removed from ordinary use as is the use of 'representation' to describe activation vectors or differential firing rates in response to stimuli. Of course, I don't accept the ordinary language argument with respect to either explanatory approach. Here's why.

The computational notion of representation matches up with the commonsense notion precisely at the two places that I have noted are core features of the dictionary senses: surrogacy and pointing. The idea is that data structures are about or point to something, and in so doing, those data

structures can function as a surrogate for that thing. For example, a computer model of weather patterns has data structures that individually are "about" different elements of a storm system. Further, the elements of the model can be manipulated to predict what would happen in the real storm system, if such and such changes occur. Thus, while there is a significant departure from ordinary usage, 'representation' as used in cognitive science nonetheless enjoys a significant commonality to the ordinary use of 'representation'.

Similarly, 'representation' as used in connectionism and neuroscience also shares significant commonality to ordinary usage. Here, the idea is that there are states of the brain or a connectionist network which are about or point to other states outside of themselves. In so doing, as with classical computationalism, this allows those states to play a surrogacy role in neural or connectionist processing. For example, we will consider a sensory discrimination task in chapters 7 and 8, where an animal must decide which of two vibrating stimuli is faster. The states of its nervous system that are selectively activated as a result of stimulation are thought to be about or point to that stimulation. When a second stimulus is presented, neural activity that is thought to be about the first stimulus (i.e., the "memory") is compared to neural activity that is about the second stimulus (the second "sensory representation"). Since the different stimuli occur at different times, the organism could not "directly" compare them (whatever that amounts to), and so there must be states of its nervous system that play a proxy role, standing in for different kinds and levels of energy that have previously impinged on its periphery, so that they can be compared.

Importantly, I make no claim that every time 'representation' is used by a neuroscientist, connectionist researcher, or computational cognitive scientist, she is using the word in accordance with standard usage. For all I know, edge detectors aren't really representations, nor are activation vectors describing some connectionist model (and for all I know, maybe they are). To fairly assess any particular

usage of 'representation', we need to look at the literature that does so. In chapter 7 we will have a careful look at the experimental paradigm mentioned above. Whether the brain states that I discuss there deserve the title of 'representation' cannot be adequately assessed until we've given that literature a fair look. The point I make here is that in general, the basic idea of representation as used in neuroscience provisionally deserves the name 'representation', because it involves aboutness and surrogacy just as much as the concept of representation as used in cognitive science.

The notion of 'representation' as used in connectionism involves distributed representations continually updated over time, and the notion of representation in neuroscience generally involves neural states that are causally related to some specific non-neural event. These are both very different from the idea of a recursively definable symbol system. But, crucially, the idea of a recursively definable symbol system is a technical notion that departs from ordinary usage just as much as the technical notions in connectionism and neuroscience. As a result, you don't get to say that a connectionist or neural state isn't a representation because the technical use from classical cognitive science doesn't license that claim. You only get to claim that a technical concept is not representational if it bears no similarity to the ordinary usage, noted above. But in fact the connectionist and neural concepts bear just as much similarity to, and just as much difference from, the ordinary use as does the technical concept from classical cognitive science. What really underlies many philosophers' aversion to calling brain states representations is that this does not match up with their intuitive notion *of a symbol system from classical cognitive science*. But this is a non sequitur.

There are several theoretical approaches, each of which correctly uses the word 'representation' to describe its posits. Those theoretical approaches include folk psychology and philosophical semantics, both of which posit propositional attitudes. The intentionality of propositional attitudes is associated with features such as the generation of referential opacity and fine-grainedness,

minimal rationality, and a relation one of whose relata need not exist. Additionally, it includes aboutness, the capacity for error, and causal efficacy. While intentionality may be a commonplace, as Cummins has claimed, the *concept* of intentionality and the investigation of its features are a result of many years of careful, technical philosophy. Classical cognitive science posits representations, which are symbolic data structures over which computations are performed. These representations bear little resemblance to folk psychology's beliefs and desires. Nonetheless, they have aboutness, the possibility of error, and causal efficacy (with semantics supervening on the syntax), in common with the propositional attitudes. Connectionism and neuroscience posit representational states, which bear little resemblance to classical cognitive science's states. Nonetheless, they share aboutness, the possibility of error, and causal efficacy with cognitive science's states.

Each of the above, including intentionality, are technical concepts. Yet they can all be correctly classified as representations, because they bear sufficient similarity to the ordinary use of 'representation', which involves pointing to or standing-in-for. Given that there is a core concept that is common to all of the theoretical approaches as well as the ordinary dictionary use of 'representation', I assume that there is such a thing to which that core concept refers. My task in this dissertation is to explain that thing, on the assumption that it exists. This justifies my use of 'representation' to describe my target. I seek to explain both what representation is, as well as, importantly, how representation is implemented in the brain. This project is distinct from the more traditional project of naturalizing intentionality in that I have identified only the core elements of various theoretical concepts of representation as my target.

1.3.2 Eliminative Materialism

Eliminative materialism is the thesis that folk psychology is false. Further, it is so radically false that, like caloric, phlogiston and the starry sphere, the ontology of that theory will be entirely eliminated. When neuroscience matures, Paul Churchland (Churchland 1979, 1981, 1988) tells us, we will agree that there are no such things as beliefs, desires, fears, wishes, hopes, pains, etc.

Churchland's main argument for eliminative materialism is that (i) it is a theory, and as such should be evaluated with respect to its possession of the usual theoretical virtues, (ii) it suffers from serious explanatory, predictive, and manipulative failures, hence (iii) it should be rejected and with the rejection of the theory goes the rejection of the ontology. Churchland supports (ii) with the following. There are many aspects of ourselves about which folk psychology is either silent or wrong. Commonsense psychology cannot explain the phenomenon of sleep or why we need it, memory, intelligence and intelligence differences, learning, creativity, or mental illness. The paucity of the explanatory resources of folk psychology becomes more evident when we consider not only individuals with relatively normally functioning brains, but individuals with damaged brains and the strange and unexpected phenomena that occur as a result of such damage. Syndromes such as blindsight, hemineglect, prosopagnosia, Anton's syndrome[9] and many others, lack any explanation and are unpredictable from the perspective of folk psychology. Further, folk psychology is stagnant, not having changed in 2,000 years. Because of these difficulties, Churchland concludes that it is false, and along with its ontology, should be rejected in favor of a more neuroscience-oriented account of the mind.

[9] Blindsight and Anton's syndrome are (conceptually but not physiologically) related in a sort of inverse way. In blindsight, patients lose qualitative visual consciousness; it seems to them as if they are blind. However, some processing of visual information nonetheless occurs. In Anton's syndrome, apparently no processing of visual information occurs, but patients insist that they are not blind, and instead confabulate stories explaining their mishaps as a result of their blindness. Prosopagnosia is the inability to recognize faces, but is not accompanied by general loss of visual acuity. Hemineglect is the phenomenon whereby a person neglects everything to one side of their visual field, even to the drastic extent of denying that their limbs on that side are their own. Each of these phenomena occur as a result of damage to a specific part of the brain, and are neither explained, predicted, nor able to be manipulated using concepts from folk psychology.

Eliminative materialism constitutes a prima facie objection to my project because beliefs and desires are paradigmatic cases of representational states. Because of this, perhaps all the talk of "representations" in the various theories outlined above is parasitic on the commonsense, folk notion of beliefs and desires. If eliminative materialism is right, and if the notion of representation used in our theories of the mind is parasitic on the commonsense notions, then we have no reason to believe that the theoretical posits are representational either. And that implies that my project is doomed from the start, since I would be seeking to explain something that doesn't exist.

There is a general debate in the philosophy of science about scientific realism, a thesis which has various formulations but whose basic idea is that the entities posited by successful scientific theories exist and do so independent of mind or theory. Perhaps the most well-known (and in my view, best) argument for scientific realism is the argument from the success of science, or, the no-miracle argument, which goes like this. Successful, well-entrenched theories allow for predictive and manipulative utility. The only explanation for this utility, short of making it a miracle, is that those theories are largely *true*, and hence the terms in the theories refer to real things. The canonical formulation of this argument can be found in Putnam (1975, 73):

> The positive argument for realism is that it is the only philosophy that does not make the success of science a miracle. That terms in mature scientific theories typically refer (this formulation is due to Richard Boyd), that the theories accepted in a mature science are typically approximately true, that the same terms can refer to the same even when they occur in different theories- these statements are viewed not as necessary truths but as part of the only

scientific explanation of the success of science, and hence as part of any adequate description of science and its relation to its objects.

Folk psychology and the language of intentional cognitive states are undeniably useful for predicting and manipulating otherwise intractably complex phenomena, namely, human behavior. While granting that folk psychology suffers from many gaps and failures, it also enjoys many successes, and there must be an explanation for this success short of making it a miracle. The best explanation is that there is at least some matchup between theory and reality, even though it is certainly neither a precise nor a complete one.

Second, as Patricia Churchland writes (1986), reduction is typically *not* a one-to-one matchup from the entities in the old theory to those in the new, with direct implications from the laws of the new theory to the laws of the old. Rather, what are traditionally considered to be successful reductions usually involve a revision of the old theory to something close to it, and then a reduction of the revised older theory to the new one. Patricia Churchland discusses the transition from classical Newtonian mechanics to Einstein's special theory of relativity, which is generally considered a case of reduction (not elimination). It turns out, she argues, that there is no one-to-one matchup of the terms and laws of the old and the new theory connected by entailment relations. Rather, the laws of an analogue of classical mechanics, altered from the original theory, can be deduced from the laws of special relativity, plus limiting assumptions.

Folk psychology likely bears the above sort of relation to cognitive psychology and cognitive neuroscience. Memory, for example, is both a folk and a scientific concept. The folk concept of memory involves "laws" such as the following:

If S desires that *p* and remembers that in the past, doing *q* is likely to bring it about that *p* then, ceteris paribus, S will do *q*.

For another example, if I owe you money but am confident that you forgot that I owe you money, and if I am (counterfactually of course!) an unscrupulous individual, we can explain and predict my behavior in terms of the imprecise folk concepts of remembering and forgetting.

The scientific rendering of this concept has made it far more precise, distinguishing short-term and working memory from long-term memory, episodic from semantic, spatial versus object working memory, as well as distinguishing retrograde from anterograde amnesia. Empirical investigation has helped discover the capacity of long-term memory (apparently unlimited), the capacity of short-term memory (5-9 items), methods of increasing that capacity in short-term memory (chunking), the reliability or unreliability of recall from highly salient but emotionally charged events such as those involving violent or potentially violent crimes, etc. Further, many of these more precisely specified phenomena have been localized to particular brain structures: The hippocampus appears to be the mechanism for encoding short-term memory into a format usable for long-term storage for later retrieval, working memory seems to involve the prefrontal cortex, with spatial working memory at the dorsolateral prefrontal cortex and object working memory at the orbitofrontal cortex and more lateral areas such as the inferior convexity. In addition to localizing particular psychological functions to anatomically gross brain structures, the psychological process of consolidation, or, the conversion of short-term memories into long-term memories appears to be implemented, at least partially, by the sub-cellular, molecular process of long-term potentiation[10].

[10] Long-term potentiation (LTP) is a process whereby neuron anatomy undergoes long-term changes such as the growth or trimming of dendritic processes, which has the functional effect of strengthening or weakening synaptic connections between adjacent cells. Bickle (2003) has argued that this phenomenon is an example of his "ruthless" reductionism from psychological processes to neural and sub-neural processes.

The scientific concepts are significantly revised from the folk concepts, but clearly they are also derived from the folk concepts. The localization of particular cognitive functions to particular brain structures is promising towards the eventual reduction of cognition to neural states and events. Patricia Churchland noticed that older theories get revised before getting reduced. We should take that lesson seriously, but should also notice that the revisions of the cognitive theories do not occur in a vacuum, abstracted from advances in neuroscience. Rather they happen concurrently with the neuroscience informing the psychology and vice versa (this is essentially a description of the discipline of cognitive neuroscience). With the mutual interplay between related scientific approaches to understanding mind and cognition, the likelihood of reduction and hence, not elimination, becomes far greater. The theoretic notion of *representation* is probably analogous to short-term memory, working memory and so forth, whereas *intentionality* probably has the folk concept of memory as its analogue. We shouldn't get rid of intentionality, just revise and clarify *what it really is*, and it may just turn out to be to some particular property of brains.

Third and finally, all that I need to rescue my project from the grips of the eliminative materialist is representation in the core sense that I've outlined. I am not wedded to any view about folk psychology. It is consistent with my view that there are no propositional attitudes and hence no intentionality, yet representation is still a philosophically perplexing phenomenon in need of naturalization. It is also consistent with my view that there *are* propositional attitudes and hence there is intentionality. If this is the case then the naturalization of representation will play a role towards furthering the project of naturalizing intentionality. Reduction and elimination are endpoints on a continuum, and eliminative materialism only threatens my project if it turns out that we are on the terminus of that continuum, where not only is folk psychology false, but no remnants of it, including the

weaker notion of representation, survives the inexorable advances of neuroscience. But if that is the case then we need an answer to the no-miracle argument, which is not forthcoming.

1.3.3 Dennett, Darwin, and the Distinction between Original and Derived Intentionality

I endorse the following distinction: Some states are representations *not* in virtue of their relationship to other representational states, while other states are representational in virtue of being suitably related to distinct states that are themselves representational. My goal in this dissertation is an explanation of basic, original representations. Dennett denies that there is such a distinction.

In his (1987) article "Evolution, error, and intentionality" Dennett articulates several versions of the original/derived distinction, but they are not all equivalent, and the arguments he provides against one version do not generalize to the others. I number the different versions of the distinction to keep them clear. He begins by stating that

> #1: The doctrine of original intentionality is the claim that whereas some of our artifacts may have intentionality derived from us, we have original (or intrinsic) intentionality, utterly underived … we are Unmeant Meaners (1987, 288).

The first version of the distinction is similar, although not explicitly identical to, the version I endorse. Dennett provides a second reading:

#2: Any computer program, any robot we might design and build, no matter how strong the illusion we may create that it has become a genuine agent, could never be a truly autonomous thinker with the same sort of original intentionality we enjoy. For the time being, let us suppose that this is the doctrine of original intentionality, and see where it leads us (1987, 290).

This second reading is the distinction between a designed artifact and a human, and the conclusion reached is that no such designed artifact, no matter how complex, could have original intentionality. Dennett's argument as to why there is no distinction relies on #2, and that argument does not apply to #1.

Dennett provides a third version, which relies on whether content is unique and determinate (corresponding to original intentionality) or indeterminate (corresponding to derived):

#3: …[Dretske gives a story about how an organism could] come to establish an internal state that has a *definite, unique* function and hence functional meaning [and it is functional meaning for Dretske that determines original intentionality] (1987, 305).

Finally, the fourth version is the distinction between whether content is "real" (for original) as opposed to merely "as if" or not quite real (for derived).

#4: Why should Dretske resist the same interpretive principle in the case of natural functional meaning? Because it is not "principled" enough, in his view. It would fail to satisfy our yearning for an account of what the natural event *really* means (1987, 304).

The third and fourth versions of the original/derived distinction require a bit of interpretation as they are not stated as explicitly as the first two. However as we'll see, most of the work is done by the second reading of the distinction, which is quite explicit.

Dennett begins the article with a discussion of a vending machine that discriminates US quarters from other similar objects. We may interpret the machine using intentional language, so that when a US quarter is inserted into the machine, the machine goes into state Q, and we can interpret that state as "meaning" (note the scare quotes) "I perceive/accept a genuine US quarter now" (1987, 290). Sometimes the machine goes into state Q in response to a non-US quarter, and other times it fails to go into state Q when a quarter is inserted. Thus, it does not always go into state Q when it is "supposed to". In these cases, we can say that the machine "misperceives" or makes an "error". The ascription of intentional states and the designation of some states as veridical and others as mistaken, are relative to the context determined by the intentional states of its users (for example, the US vending company). In this sense, the machine has derived intentionality, since the intentionality of its states is only relative to the intentional states of its users.

However, imagine that the vending machine is transported to Panama. Panamanian balboas are easily distinguishable from US quarters by the design stamped on the front and back, but not by weight, shape, or thickness. The same vending machine could be, without modification, immediately used as a balboa-detector. When the machine is relocated to Panama and used as a vending machine that

accepts balboas, Dennett asks, at what point should we say that the machine no longer makes an "error" in accepting balboas? In the US, accepting a balboa would be a "mistake", but transported to Panama, accepting a quarter would be a "mistake". Is there any fact of the matter about what the states of the machine "really" mean?

There is freedom about what we should say, Dennett tells us, because it is only relative to the function of the machine that an ascription of intentional states is licensed, and the function of the machine is indeterminate.

> And given that this historical fact about its origin licenses a certain way of speaking, such a device may be characterized as … a thing whose function is to detect quarters, so that *relative to that function* we can identify both its veridical states and its errors (Dennett 1987, 292; emphasis in the original).

Dennett thinks that human intentional states are like the "intentional" states of the vending machine's coin-detector. They are relative to an interpretation, and that interpretation is licensed by functional ascriptions. The functional ascriptions come from natural selection, and, just like functional ascriptions dependent on human uses, these functions are inherently indeterminate. Sometimes, he tells us, there simply is no fact of the matter about a function and hence, an intentional content.

Consider Dennett's next thought experiment. Imagine a person who wants to preserve her body for many hundreds of years, and the only possible way of doing this is to keep it in a hibernation machine. The hibernation machine needs a constant store of energy to continue functioning. Since the

person will not be awake, she must design the machine so that it is able to cope with whatever obstacles arise in order to maintain a continuously updated energy store. One strategy is to leave the machine stationary, next to a reliable energy source. But if in the future something were to occur where that energy source or location is not viable, the machine would fail and the person would die. So the better strategy would be to give the machine flexibility in behavior, enabling it to move around so that it can pursue its "goal" of keeping its human alive. The machine would need to be able to react to its environment, to form intermediate "goals" and "strategies" for pursuing them, and, assuming that this person and her hibernation machine are not the only ones around (surely the trend will catch on), the machine should be able to "communicate" and make "alliances" with other hibernation machines, and so forth.

All of this occurs with the person inside totally unconscious and unable to control the machine or make any decisions, and hence, without any occurrent intentional states. Whatever "intentionality" we ascribe to the machine is derived from its relationship to the goals of the person who built and designed it. Dennett draws the following conclusion from this thought experiment:

> I want to draw out the most striking implication of standing firm with our first intuition: no artifact, no matter how much AI wizardry is designed into it, has anything but derived intentionality. If we cling to this view, the conclusion forced upon us is that our own intentionality is exactly like that of the robot, for the science-fiction tale I have told is not new; it is just a variation on Dawkins's (1976) vision of us … as 'survival machines' designed to prolong the futures of our selfish genes. We are artifacts, in effect, designed over the eons as survival machines for genes that cannot act swiftly and informedly in their own interests (1987, 298).

This is Dennett's argument: No artifact can have original intentionality. Humans are artifacts designed by natural selection. Therefore humans do not have original intentionality. Further, Dennett notes that surely the "intentionality" of genes is itself merely "as if" intentionality, and is not "real" intentionality[11]. So it is just interpretation all the way down, and there is no original/derived distinction. It is all derived, "as if", and indeterminate.

First, notice that Dennett's argument applies to reading #2 (and possibly #'s 3 and 4), but not to my reading or Dennett's #1. But why should we accept the original/derived distinction as that between designed artifacts and humans? We are naturalists and thus accept that humans are a part of the physical world and subject to the forces of natural selection. Hence it is more in keeping with this view to argue that, since natural selection has "designed" physical systems of sufficient complexity to exhibit "real" original intentionality, so it must follow (at least in principle) that other systems of sufficient complexity could be designed that also exhibit original intentionality. Thus, we should not accept the first premise of his argument, which is that no artifact can have original intentionality. Instead of concluding that we do not have original intentionality, we should take the more plausible line and conclude that Dennett's reading of the original/derived distinction is wrong, or, that artifacts can have original intentionality.

Second, this argument does not apply to the original/derived distinction that I have endorsed. The distinction is a metaphysical one of ontological dependence. Dennett's argument is entirely orthogonal to this reading of the distinction and hence does not show that there is no such distinction.

[11] "This vision of things, while it provides a satisfying answer to the question of whence came our own intentionality, does seem to leave us with an embarrassment, for it derives our own intentionality from entities – genes – whose intentionality is surely a paradigm case of mere *as if* intentionality" (Dennett 1987, 298-299; emphasis in the original).

There is a different argument we might take from Dennett's discussion, which goes like this. The ascription of intentionality is licensed only relative to the ascription of a function. Hence, the ascription of intentional states (and thus a particular content) is relative to some function. But function ascription is indeterminate, and thus content is indeterminate. The original/derived distinction is that between determinate and indeterminate content, and since all content is indeterminate, it follows that there is no original/derived distinction.

This argument fails for at least three reasons. First, it only applies to Dennett's #3. It does not apply to his first reading or to the construal that I endorse. Like the previous argument, it does not even address the distinction that I endorse. Second, we have no reason to assume that original intentionality must have determinate content. I will return to this point below, in this subsection.

Third, Dennett vacillates between whether *ascriptions* of content are indeterminate, or *content itself* is indeterminate. These are two very different issues. The first is an epistemological issue about how we can know what the content of some intentional or representational state is, while the second is a metaphysical claim about what the content of an intentional or representational state is.

Dennett's discussion of the original/derived distinction serves to confuse the issue more than to clarify it, as he elides several important distinctions and draws several unwarranted implications from his elision. In addition to conflating the above four divergent views, he discusses the notion of privileged access and its purported relation to the distinction in question as well as the (purported) logical relations between the doctrines of natural selection and the original/derived distinction.

Pressing the original/derived distinction, he claims, implies denying that we are artifacts designed through the process of natural selection[12]. Further, accepting natural selection, he claims, implies denying that we have privileged access to the contents of our own thoughts, but further, it also implies that "*there are no such deeper facts* [that fix the meanings of our thoughts.] Sometimes functional interpretation is obvious, but when it is not … when more than one interpretation is well supported – there is no fact of the matter" (1987, 300).

None of this follows. Pressing the original/derived distinction only implies denying the "artifact view" of humans if you accept *both* Dennett's #2 *and* the view that the artifacts adverted to in Dennett's #2 (computers, robots) are *relevantly the same kind* of artifact as humans are in the artifact view of humans. But we should not accept either conjunct. One crucial difference between the two kinds of artifacts, pointed out by Dennett, is that our artifacts are designed by conscious designers with intentions, while we (qua artifacts) are "designed" (note the scare quotes) by blind, unthinking, unconscious forces of natural selection.

Second, accepting natural selection does not imply that there are no deeper facts that fix a determinate content for our thoughts. This only follows if you (i) elide Dennett's #'s 2 and 3 , (ii) argue that natural selection implies that humans are relevantly the same kind of artifact as computers and robots, hence do not have original intentionality, and then (iii) from conflating #2 and #3, conclude that since humans don't have original intentionality, the content of their thoughts must be indeterminate, thus there are no deeper facts of the matter. But I've already argued that #2 and #3 are distinct so (i) is not legitimate, and further (ii) is unsupported and there is reason to believe that they are not the same

[12] "The idea that we are artifacts designed by natural selection is both compelling and familiar; some would go so far as to say that it is beyond serious controversy. Why then, it is [sic] resisted not just by Creationists, but also (rather subliminally) by the likes of Fodor, Searle, Dretske, Burge, and Kripke?" (1987, 300). Dennett exegesis is needed here, but what the above named authors all have in common is their pressing of the distinction. Hence, I claim that Dennett claims that pressing the distinction implies denying that we are artifacts designed through natural selection.

kind of “artifact”. Further, I repeat that both #’s 2 and 3 are distinct from the reading of the distinction that I am interested in.

Another argument on which Dennett might be basing his conclusion that natural selection implies that there are no deeper facts about content is that facts about meaning or content are fixed by facts about functional interpretation. But that just assumes a particular brand of teleosemantics.

With respect to privileged access and the relation between naturalism, indeterminacy, and natural selection, he says

> Either you must abandon meaning rationalism- the idea that you [have privileged access to the contents of your own thoughts] – or you must abandon the naturalism that insists that you are, after all, just a product of natural selection, whose intentionality is thus derivative and hence potentially indeterminate (1987, 313) … You can’t have realism about meanings without realism about functions (1987, 321).

Accepting naturalism does imply accepting the basic framework of the theory of natural selection, but this does not imply that all intentionality is derivative, for the numerous reasons I’ve given above. Dennett sets up a false dichotomy between the doctrines of privileged access and a confused mishmash of several different theses that he conflates and puts all under the heading of the “original/derived distinction”.

The topics of indeterminacy and “real” versus “as if” content come up frequently in Dennett’s discussions, and are in need of clarification. The concept of indeterminate content admits of at least

two renderings. The first is that grounded in the claim that "there are no deeper facts", or "there is no fact of the matter" about what the content is. The second is the claim that there is no *unique* content for a given representation or intentional state.

The first rendering of indeterminacy seemingly implies anti-realism. That there is no fact of the matter about what the content of a given representation is directly implies that the state in question does not have a content (and hence is neither a representation nor an intentional state). So for example if we were to inquire: "what is the content of state X?", and the accurate response is, "there is no fact of the matter about what the content of state X is", or, "there are no deeper facts which fix the content of state X", then state X does not have a content. The instrumentalist's reply is that, while there is no "deeper fact" about what the content is, still, it is useful and predictive to treat the system in question as if it had intentional states with such and such contents. Further, adopting the intentional stance (Dennett 1987) allows us to see *real patterns* that could not be seen from any other perspective. Dennett insists that his view is "a *sort* of realism" (1987, 37). But he goes further to say that

> These patterns are objective – they are *there* to be detected – but from our point of view they are not *out there* entirely independent of us, since they are patterns composed partly of our own "subjective" reactions to what is out there (Dennett 1987, 39).

So on the one hand, this rendering of indeterminate content seems quite obviously to imply anti-realism about representations or intentional states, while on the other hand, the claim that adopting the intentional stance is merely a way of looking at systems in the universe that allows us to see patterns that are really (objectively) there seems to imply some version of realism. The

instrumentalist position is an unstable one: if there are real objective patterns "out there" to be discovered and the intentional stance allows us to see those patterns while providing explanatory, predictive and manipulative utility, then the no-miracle argument applies. There must be some explanation for the success of adopting the intentional stance, and the only explanation is that adverting to beliefs and desires works because there *are* beliefs and desires. But this is at odds with the straightforward claim that (at least sometimes) there are no deeper facts, or there is no fact of the matter about content and hence, there is no content and the state in question is not representational. But if we are not to accept the obvious transition to anti-realism, what then?

A second reading of the indeterminacy claim is that there is no *unique* content for any given intentional state. In this case, there is no (singular) *fact* of the matter because the state in question has more than one content or is about more than just one state of affairs. Rather than claiming that there being no deeper fact of the matter implies anti-realism, we can maintain the position that sometimes content is indeterminate in the sense that sometimes intentional or representational states have *more than one* content.

Another way of thinking of this is as follows. A representational state is either about something or it is not. If it is not, then it has no content and hence is not a representational state[13]. But if it is a representational state, then it must have *some* content or other. If Dennett is right that sometimes intentional states are indeterminate with respect to what they are about, then, in order for them to *be* representational, the concept of indeterminacy cannot imply anti-realism about content. So we are left with the second reading, which claims that there is no *unique* content. But this is entirely consistent with the claim that some states are representational not in virtue of their relationship to some other

[13] This will need to be qualified later on, but it's going to take some set up work to characterize that qualification. Rather than confuse the issue I'll just ignore it for now.

state that is representational, while others are. So the (possibility of the) indeterminacy of content does not imply that there is no original/derived distinction.

It is important to keep in mind here the distinction between representation and intentionality. Dennett's arguments pertain to intentionality, not representation as I've articulated it. Further, he doesn't take intentional states to be discrete states (of a person's brain or otherwise) (Dennett 1982), but rather, underlying states which explain or systematize whole patterns of behavior discernible only from the intentional stance. So the apparent anti-realism about intentionality need not bother us any more than Churchland's rejection of the ontology of folk psychology. I've argued here that even if representations, as discrete states of an organism's central nervous system, are sometimes indeterminate in what they are about, this need not threaten the original/derived distinction, and hence, need not threaten my project.

1.3.4 Horgan on Folk Psychology and Cognitive Science

In his (1992), Horgan reviews three books (Cummins 1989; Baker 1987; Garfield 1988) that take up questions about the philosophical foundations of computational cognitive science. The major thesis of Horgan's paper is as follows. All three authors under discussion have accepted, but need not and should not, the following Fodorian thesis:

> (F. 15) If there is a place for intentional categories in a physicalistic view of the world, and if a physicalistic view of the world is correct, then the intentional can be "naturalized", in the sense

> that there are tractable sufficient conditions, formulable in non-intentional and nonsemantic vocabulary, for a physical system to have intentional states[14] (Horgan 1992, 457).

Because of this, Baker wrongly rejects physicalism because she despairs of finding those tractable conditions, Garfield misconstrues the nature of the project of the naturalization of intentionality, and Cummins, like Fodor, is on a wild goose chase for tractably specifiable conditions that are likely not to be found. Prima facie, it would seem that, according to Horgan, I am also on a wild goose chase.

Horgan argues that it may also be the case that physicalism is true in that intentionality is not *sui generis* or irreducible, yet there is no way to tractably specify the conditions of supervenience. Or in other words, naturalism, broadly understood, can be true, even though the project of "naturalizing intentionality" as construed by Fodor et al. is doomed to fail.

> In short, it might be that the search for tractably specifiable, cognitively surveyable, nonintentional and nonsemantic sufficient conditions for intentionality is utterly hopeless – and yet that the intentional supervenes on the nonintentional nonetheless (Horgan 1992, 461).

Horgan reaches this conclusion through an analysis of the relationship between computational states, propositional attitudes, and brain states, and with the help of a new multiple realizability argument. Two theses crucial to Fodor's view are:

[14] "Roughly, a tractable specification is a relatively compact, relatively non-baroque, nondisjunctive, cognitively surveyable, formulation of *sufficient conditions* (for some philosophers, sufficient *and necessary* conditions)" (Horgan 1992, 453).

(F.1) The computational conception of the mind is correct (Horgan 1992, 450).

(F. 9) An adequate computational cognitive science would posit beliefs, desires, and other PA's [propositional attitudes] (1992, 452).

Starting from a realist position about the propositional attitudes, as well as the classical computational theory of cognition, we get the following view of mentality. "(F.3) Mental states [such as propositional attitudes] and processes are type identical to computational states and processes" (1992, 450), while computational states and processes are *realized* by (and multiply realizable by) physical systems. The computational view is a theory of the *identity conditions* of mental states, but Horgan suggests instead that computational states should themselves be thought of as *realizer states* for mental states, just as physical states are realizer states for computational states. He borrows the term 'psychotectonic' from McGinn (1989, 71) to describe the realization relation of mental states by computational states. *Psychotectonic realization* is, for Horgan, the realization of the functional architecture of cognitive states and processes by computational states and processes[15]. "The core claim, then, is that mental states are psychotectonically realized by certain functional/computational states, which in turn are physically realized by certain neurobiological states" (Horgan 1992, 454-455).

Horgan uses a new multiple realizability argument, which parallels the standard multiple realizability argument against the type identity of mental states with brain states, to argue for this core claim. The argument goes like this. Even given the computational theory of mind and some version of

[15] "Borrowing from Colin McGinn (1989, 171) the term 'psychotectonics' – an apt name for scientific theorizing about cognitive functional architecture – I will call the relevant relation *psychotectonic realization*" (Horgan 1992, 454).

intentional realism, “the possibility remains open that in different kinds of creatures (say, humans and Martians), the same belief states, with the same contents, are differently realized psychotectonically” (1992, 455). He enumerates some alternative ways that belief states might be realized by computational states. Perhaps Martian mentalese also uses a system of language-like representations, even though the language is different. Or, perhaps the inner language is the same, but the mental states are realized by different computational relations (since in general, distinct algorithms can compute the same function), and finally, perhaps there are creatures whose mental states are realized by computational relations to systems of mental representations that are not language-like. Horgan claims that it would be chauvinistic to identify mental states with their computational realizers in humans, just as it is chauvinistic to identify them with their physical realizers in humans (see Horgan 1992, 455-456).

Given the thesis that computational states realize but are not identical to mental states, the question arises of how mental states are realized. Horgan posits two potential kinds of psychotectonic realization: direct and indirect. Direct realization occurs when the states that realize propositional attitudes are natural kinds, while indirect realization occurs when they are not: “although PA’s would indeed be psychotectonically realized by certain states countenanced by scientific theory, these realizing states would be quite baroque and complex, rather than being scientific natural kinds” (1992, 459). To make this possibility more vivid, Horgan presents the variety of clothing concepts as an analogy (citing Cummins and Schwarz 1988, 49). *Being a hat, being a scarf*, etc., are not scientific natural kinds, yet these are real physical objects. Linguistic competence with the use of clothing terminology involves the tacit understanding of a large set of systematic ceteris paribus generalizations, and legitimate causal explanations can be given which advert to clothing concepts. However, even though the extensions of clothing concepts are not scientific natural kinds, they are nonetheless realized (indirectly) by physical

objects whose physics-level characterization would be enormously complex. The propositional attitudes may be just like that, in which case the project of naturalizing intentionality is just as doomed as would be the project of naturalizing clothing.

Horgan uses his new multiple realizability argument to support his thesis about the psychotectonic realization of intentional states by computational states, and then provides an analysis of the different kinds of psychotectonic realization we should expect. Given the strong requirements of necessary and sufficient (or even just sufficient) conditions, it is more likely that the reduction of intentional states to computational states, and then again computational states to physical states, will be the indirect kind. The analogy to clothing concepts makes this possibility all the more vivid, and thus we should not expect Fodor's project, or my project, to succeed. The whole chain of reasoning begins with his second-order multiple realizability argument, but that doesn't work. However, Horgan has a much simpler point to make, which does not require the complex machinery of indirect psychotectonic realization. That simpler point is useful in further clarifying the sort of project those of us interested in the naturalistic foundations of mind should engage. But first I will address his argument as stated.

To address Horgan's new, second-order multiple realizability argument, it is helpful to take a brief look at the standard (first-order[16]) multiple realizability argument, and its motivation. Putnam (1960, 1975), Armstrong (1968, 1981), and Lewis (1966) were important early architects of the functionalist conception of mind. The basic idea is very simple. Behaviorism was wrong in denying the existence of inner states which mediate perceptual inputs and behavioral outputs. But behaviorism was also onto something: there seems to be a deep conceptual connection between mental states and behavior. Hence functionalism, the view that mental states are states which play a particular functional

[16] I use 'first-order multiple realizability' to refer to the multiple realizability of mental states (or computational states, depending on context) by physical states. I use 'second-order multiple realizability' to refer to Horgan's claim that mental states can be multiply realized by computational states, which can themselves be (first-order) multiply realized by physical states.

role in a causal economy of inputs, parallel mental states, and behavioral outputs, was born. At its basic level, functionalism is a conceptual analysis of what it is to be a mental state. With the analysis of mental states as states that play or are apt to play a particular causal role, we have the following corollary: it is possible that very different physical (or even non-physical) systems can realize states that play the particular causal roles that mental states are thought to play. Hence multiple realizability.

In addition to the construction of functionalism as a philosophical theory of mind, important advances in the theory of computation and computer science were made, especially by Turing (1950). Most relevant to the present discussion, what Turing's work[17] led to was the realization that we could build machines whose states could be defined in purely syntactic/formal ways (that is, without reference to anything outside of it in the way that we define representational states in terms of their contents), yet those states could implement semantic relations (such as truth-preservation). In other words, it is possible to build machines whose inner states operate on and are describable by purely mechanical (hence physical) principles, but at the same time implement semantic properties and respect semantic relations. Furthermore, different physical systems could realize the same abstract functional architecture, and the hence same computational states. This wedding of syntax to semantics, along with the multiple realizability of formal/syntactical relations by different physical systems, lent much support to the multiple realizability claim from the functionalist theory of mind.

Thus, given the conceptual analysis of mental states as states apt to play a particular causal role, coupled with the success of computer science in demonstrating the possibility of physical systems that (i) play something like the relevant causal roles postulated by philosophers as well as (ii) apparently implementing semantic properties and relations solely in virtue of formal/syntactic properties, it is

[17] And von Neumann's, and many others. See Haugeland (1985) for an excellent introduction to artificial intelligence.

reasonable to conclude that computational states and processes are identical to mental states and processes. Hence, the computational theory of mind was born. (I am not endorsing the computational theory of mind, nor its corollary of multiple realizability. Rather, I am trying to clarify the motivation behind the standard multiple realizability claim to see if corresponding motivations exist for second-order multiple realizability.)

Horgan models his second-order multiple realizability thesis on the standard multiple realizability thesis. But what is the motivation, which has a parallel to the first-order level, for ascending to the second level? The philosophical claims about functionalism and the nature of mental states which imply multiple realizability do not have a parallel at the second level. The discovery from computer science that different physical systems can instantiate the same formal/syntactic properties also lacks a parallel at the second level.

Horgan claims that "the possibility remains open that in different kinds of creatures … the same belief states, with the same contents, are differently realized psychotectonically" (Horgan 1992, 455). Horgan assumes that "psychotectonic realization" of mental states by computational states is in fact a possible phenomenon, and further, that it is possible that mental states can be differently psychotectonically realized. But whether or not it is even possible is just what is at issue between Horgan and the classical computationalists such as Fodor. The computationalists, borrowing from their functionalist predecessors, claim that thought/thinking/cognition *just is* a species of computation. That is, mental states and processes are not merely modeled by or realized by computational states and processes, but rather mental states and processes *are* computational states and processes. The assertion that it is possible that the same mental state can be differently psychotectonically realized begs the question against the computationalists. If second-order multiple realizability is true, then the computationalist's identity claim is false, but no argument has been given for second-order multiple

realizability. Additionally, Horgan uses second-order multiple realizability to argue for the alternative thesis that mental states are psychotectonically realized by computational states. But this also begs the question: if his alternative thesis is true, then his second-order multiple realizability claim would follow. But he doesn't get to make the second-order multiple realizability claim about how mental states are psychotectonically realized, unless he has independent reason to believe that there is such a thing as psychotectonic realization, as opposed to the computationalist's identity claim. So his multiple realizability argument for psychotectonic realization begs the question.

Horgan does present an argument of sorts when he details the different ways that he thinks mental states can be psychotectonically realized by computational states. The second possibility he lists is that

> perhaps … mental states are psychotectonically realized in Martians via different *computational relations* than in humans. (After all, in general various different algorithms can compute a given (computable) function; accordingly various different computational relations to internal representations could subserve the same transition function over these representations) (Horgan 1992, 455)

The argument is in the parentheses. In general different algorithms can compute the same transition (that is, input-output) function, so different computational states can implement the same functional states, and if functional states = mental states, then Horgan's second-order multiple realizability would follow. However, this argument depends on an implicit premise on how to individuate both computational and functional states, which is not supported.

For this argument to work, we have to individuate functional states coarsely, in terms of their input-output states, so that two state tokens that are input-output equivalent would be considered functionally type identical. We also have to individuate computational states at a much finer grain than this, where input-output equivalence is not sufficient for computational type identity, but a further algorithmic equivalence is needed[18]. It would then follow that different computational (algorithmic) states can instantiate the same mental (functional/input-output) state. But why is that the right level of fineness of grain for individuating both functional and computational states?

The basic functionalist claim is only that mental states are states that play a particular role. How to define that role is unclear. One way is (as Horgan apparently suggests) solely in terms of inputs and outputs. So if two particular states each always produce some output state *O* whenever confronted with the input state *I*, then we can say that they are input-output equivalent and hence the same type of state. But another option is to individuate *roles* more finely, in terms of the algorithm or method, or intermediary states that they go through in order to produce *O* in response to *I*. Then if we have two (functional) states *X* and *Y*, both of which produce *O* in response to *I*, but which go about doing so via different intermediary states (hence, they implement different algorithms for computing the function), then we should say that they are different types of states, because they play *different* functional roles. If we individuate mental states by their roles, then we would say that they are different mental states. Thus, the observation that different algorithms can instantiate the same input-output function is not telling towards whether mental states are identical to or psychotectonically realized by computational

[18] Pylyshyn denotes these different kinds of type identity "weak" and "strong equivalence", with weak equivalence being input-output equivalence, and strong equivalence being identity of algorithms, or the completely specified procedure that takes the system from the input state to the output state (see Pylyshyn 1984, 87-106 and especially 88-89).

states. That depends on how you individuate both computational states and mental states[19]. No reason has been provided for why we ought to individuate cognitive states at the coarse grain while individuating computational states at a much finer grain.

I see no reason to follow Horgan in his ascent to second-order multiple realizability. His argument only superficially parallels the standard first-order multiple realizability arguments against type-type psychophysical identity claims, and the underlying motivation for the standard claim is not shared by the second-order claim. Further, Horgan's multiple realizability argument for psychotectonic realization begs the question against those who claim that mental states are identical to computational states. Finally, his argument based on the multiple algorithms that can instantiate computable functions is not decisive because it depends on a further unsupported assumption about how to individuate both computational and functional or mental states[20].

Horgan's multiple realizability argument for psychotectonic realization doesn't work, and that is the crucial premise that leads to his further claim that it is likely that the kind of psychotectonic realization that will occur is the indirect one, which in turn leads to his claim that we should not expect there to be tractable conditions for the reduction of the intentional to the physical. However, he has a much simpler route for getting to this intractability claim, which does not need the complicated and ultimately unsuccessful machinery of second-order multiple realizability and indirect psychotectonic realization. He argues for this in his (1994), to which I now turn.

An underlying assumption of the "naturalizing intentionality" project is this:

[19] The option considered above is to individuate functional and hence mental states at the same fine-grained level as we do computational (algorithmic) states. But another is to individuate computational states coarsely, such that input-output equivalence is sufficient for type identity of computational states. Pylyshyn (1984, 89) advocates individuating both computational and functional or cognitive states at the fine-grained algorithmic level.

[20] I have argued that Horgan's second-order multiple realizability should not be accepted. However, I should make it clear that I also do not take the standard arguments for multiple realizability as decisive.

> Either (i) there are tractably specifiable non-intentional, non-semantic, sufficient conditions (or sufficient and necessary conditions) for something's being a mental representation with a specific representational content, or (ii) mental content is among the ultimate, fundamental, and unexplainable properties of things (Horgan 1994, 309).

Horgan argues that this assumption is so deeply ingrained that it is not even noticed (p. 309). However, it is a false dichotomy. A third possibility is that intentional properties supervene on physico-chemical properties yet the supervenience base is baroque and intractably complex, thus dooming the project of finding them[21]. His reasons are as follows.

First, only one aspect of the "non-basic" status of intentionality is its supervenience on physical states and properties. But supervenience does not imply that there are cognitively surveyable, tractable conditions. Hence, supervenience does not presuppose these tractable conditions. Second, he provides an inductive argument on past failures: not only are there usually counterexamples to the sorts of reductive analyses provided by Fodor et al. but we should also notice the failure of giving reductive necessary and sufficient conditions for other concepts. Horgan appeals to the prototype view of human

[21] Here is a long quote from (Horgan 1994, 309): "Perhaps, for instance, the supervenience base for a token thought … generally involves a good-sized chunk of space-time extending well beyond the cognizer's own body and well beyond the time at which the token thought occurs; perhaps it involves a rather gargantuan number of physico-chemical goings-on within that extended spatio-temporal region; and perhaps there isn't any simple way to describe, in non-intentional and non-semantic vocabulary, all the *relevant* aspects of this hugely complex supervenience base. Perhaps, in addition, the supervenience of the intentional on the non-intentional is largely a holistic matter – with the intentionality of thoughts, utterances, and inscriptions supervening not individually (one token at a time), but rather collectively, as part of the correct global intentional interpretation of a cognizer—or perhaps of the cognizer's whole community or whole species."

concepts (citing Rosch 1973, 1975, 1978) to argue that no concept admits of clearly delineable necessary and sufficient conditions so we should not expect the concept of intentionality to admit of them either.

Horgan concludes that it is time to drop the project of providing tractable sufficient conditions for intentionality, and instead focus on the project of understanding what sorts of supervenience relations and explanations are acceptable from within the naturalistic framework, and whether intentionality really is susceptible to those explanations. Horgan's claim that the naturalizing intentionality project has an unstated assumption of a false dichotomy does not depend on his further views about second-order multiple realizability.

His point is well-taken. There is indeed this assumption underlying the project as conceived by Fodor et al., and we have no reason to take that assumption for granted. Rather than despairing of the project and focusing on understanding which supervenience relations are naturalistically acceptable, there is an important lesson to be learned here, both about the nature of the project we ought to engage and the methodology we ought to use.

The project of giving sufficient conditions for intentionality just is the project of giving a reductive analysis of intentionality, and the rules that Fodor et al. have accepted is that their conditions cannot make use of intentional or semantic terms. The goal is to understand what intentionality *is*. In personal communication, Devitt has argued that I must make a distinction between what he calls the *constitution question* and the *implementation question*. The constitution question is that of understanding the nature of representation, or what representation is, while the implementation question involves understanding how some particular physical system implements or realizes representation[22]. One could not approach the latter scientific question without some handle on the

[22] In print Devitt (1996) puts the point as follows. The fundamental task for semantics is this: "The 'basic' semantic task is to say what meanings *are*, to explain their *natures*" (1996, 54). This is analogous to the constitution

former and explanatorily prior philosophical question of constitution. This sort of sentiment is also expressed by Horgan when he says

> just as it would be chauvinistic to identify mental states with the neurobiological states that happen to *physically* realize them in humans, it would also be chauvinistic … to identify mental states with the syntactic/computational states that, according to computational cognitive science, happen to *psychotectonically* realize them in humans (Horgan 1992, 456, emphases in the original).

The key phrase in this quote is 'happen to': the neurobiological states that "happen to" realize mental states in humans, in virtue of being referred to as such, are thereby rendered inconsequential or unimportant to understanding what mental states *are.* Fodor ignores questions of implementation as inconsequential to the project of understanding and naturalizing intentionality[23]. There is of course a large literature on multiple realizability, the autonomy of the special sciences, and the appropriate level of abstraction for explaining intentionality and other properties of mind. I am not dismissing that

---

question. While discussing the Churchlands' eliminativism, Devitt argues that, to make their case, the Churchlands would have to (among other things) show that physical states of the brain do not have the properties or structure required of realism about representational states. "Yet, the Churchlands insist, current research on the brain shows no sign of representational states with the sentencelike structures that [realism about representations] requires" (1996, 254). That representations must have a sentencelike structure is a partial answer to the constitution question. Showing that there aren't any of those things is an answer to the implementation question. Notice that the constitution question is conceptually prior to the implementation question.

[23] Here is a typical statement of the Fodorian view on the relevance of neuroscience to "serious" philosophy, from his (1990, 125): "For simplicity, I assume that what God sees when he looks in your head is a lot of light bulbs, each with a letter on it … A mental-state type is specified by saying which bulbs are on in your head when you are in the state. A token of a mental-state type is a region of space time in which the corresponding array of bulbs is lit. This is, I imagine, as close to neurological plausibility as it is ever necessary to come for serious philosophical purposes."

literature in what follows, but rather suggesting a methodological strategy based partly on Horgan's discussion above.

Horgan's observation that there is an unstated and unjustified premise implicit in the "naturalizing intentionality" project is suggestive of both the nature and methodology of this project. The nature of that project appears to be to answer the constitution question, and the methodology is to ignore implementation as irrelevant. Horgan concludes that the project should be re-conceived, as an analysis of supervenience relations and naturalism. However he also apparently accepts the methodology: he thinks it would be chauvinistic to identify mental states with either the computational or neurobiological states that "happen to" realize them in humans, suggesting that an analysis of actual implementing states would be provincial, and would not get at the "deep" question of what mental states "really *are*".

I suggest rather that it is both the project *and the methodology* that is in need of re-conception, as they go hand in hand. One obvious and undeniable starting point that I absolutely take for granted is that, whatever other things (space aliens, robots, etc.) *possibly* have minds and mental states, humans *actually do*. There is an intimate relationship, whatever it is, between the central nervous system and mental states. If we want to understand one aspect of mind, namely, representation, and further want to understand how minds fit in with the rest of the physical world, it is *methodologically naïve* to assume that it is alright to ignore the brain[24]. It follows from this that it is methodologically naïve to assume that it is alright to ignore questions of implementation.

[24] I borrow the phrase "methodologically naïve" from (Eliasmith 2000). Eliasmith answers Fodor's (1999) rhetorical question, "why, why, does everyone go on so about the brain?" with this (Eliasmith 2000, 5): "It seems rather obvious why 'everyone' interested in mental function goes on so about brains: brains are the only agreed upon instances of physical systems exhibiting mental function. Methodologically speaking, if we get a good theory about how brains perform the mental functions they do, we have *at the very least* a partial theory of how physical things give rise to mental things (or realize mental relations). Such a partial theory would be a great improvement over

The methodological strategy that I suggest, and which I adopt in this dissertation, is a dual one of approaching the problem of representation from both the constitution perspective and the implementation perspective. Incremental advances from each side both inform, and refine the questions asked, on the other side. Often in scientific inquiry the phenomenon of interest is only vaguely or generally understood, and through empirical investigation the theoretical underpinning of what was being investigated in the first place becomes clearer or more developed. Or, sometimes we are able to recognize a thing for what it is, yet do not understand its underlying nature. For example, we can recognize water by perceiving its superficial properties. Nonetheless, to understand its underlying nature, to realize that water is constituted by H_2O molecules, we need the tools of modern chemistry.

In this case, the phenomenon of interest is representation. Unfortunately, the situation is complicated right from the start, because it isn't clear what those superficial properties are by which we might recognize a thing as a representation. I have given a preliminary characterization of representation in terms of aboutness, the possibility of error, and causal efficacy, or even more broadly, in terms of surrogacy or pointing. However, if I point to a brain and tell you, "this piece right over here has aboutness", I would expect you to rightly claim that I've probably begged the question. That's not my strategy.

Rather, in chapters 7 and 8 I describe some recent experimental results involving single-cell intracortical recordings taken from a monkey engaging in a sensory discrimination task. I make no general claims about how to recognize representations in the absence of a theory. However, I will argue that, at least in this case, I have identified some candidate neural vehicles of representation. An analysis of those vehicles can be used to refine our theoretical understanding of representation (i.e. the

---

what is currently on offer, even if it is only partial. And, of course, there is always the prospect that such a theory can be generalized to cover more than brains: we *can't* rule out this possibility without having seen such a theory to start with. These, I take it, are good reasons for thinking that knowing neuroscience will help unravel some of the mysteries of our mental lives."

constitution question), as well as refining and providing stronger constraints on the considerations we use to identify the states that implement it.

This back-and-forth, incremental approach is reminiscent of Rawls' reflective equilibrium in moral theorizing and Quine's metaphor of Neurath's boat in justifying our practices of scientific reasoning. We identify some brain states that may reasonably be assumed to be representational (I have not yet provided any constraints on how to do this), analyze how they go about physically implementing representation, and then sharpen our conception of what it is to be a representation. With this sharpened conception in hand, we reassess empirical investigation, and perhaps sharpen our conception of representation further.

By confronting our task with the dual-approach strategy, we can take Horgan's observation into account. We do *not* assume that there are tractable, cognitively surveyable sufficient (or necessary and sufficient) conditions for "Mental state *M* has content *C*". There is, in fact, no need to assume that representation is entirely a unitary phenomenon amenable to this kind of a reductive conceptual analysis. However, through back-and-forth analyses of both the theoretical conception of representation and the implementation of it in physical systems, we approach ever closer to an understanding of representation (and hence mind) and its place in the physical world.

A final note on methodology: The commonsense concept of representation was a springboard that philosophers have used to get at the technical philosophical concept of intentionality. Other researchers have, more implicitly, relied on the commonsense notions of pointing or surrogacy, as the foundation for their various technical concepts of representation (i.e., symbolic data structures, activation vectors, and neural detectors and neuron coding schemes). Since there is this common core to both the commonsense concept as well as each of the technical concepts, I make the assumption that

there is something in the world to which that common core refers. Yet, we don't have a clear understanding of what that thing is, and in particular, we don't understand how physical or biological processes give rise to that thing. The goal of my dissertation is to explain what that thing is, and to do so in a way that is consistent with naturalism. This is ultimately both a conceptual and an empirical question. As a result, the methodology should reflect that.

The methodology that I use in chapters 1-6 is the standard philosophical methodology, and I intend the conceptual work of these chapters to stand alone and on its own merit. Nonetheless, I also contend that conceptual work must be tempered with empirical investigation, just as empirical work is necessarily grounded in some or other conceptual framework. However, the dual-approach methodology will not become prominent until chapter 7, at which point I will provide a more detailed description and defense of it.

1.4 The Plan

At bottom there are only a handful of proposed solutions to the problems of naturalizing intentionality and representation, and various combinations thereof. These include causal history, counterfactual causation/information carrying, resemblance/similarity/isomorphism, teleology, and functional/causal/computational-role theories. All versions of role semantics are holist theories, in the sense that the meaning or content of any term or state is determined by that of every other term or state in some system. Additionally, there is a debate about whether meaning or content-determining factors are external to the cognitive agent (Putnam 1975). In this dissertation I presuppose externalism and I presuppose that the sense of holism mentioned above is false. If the theory based on those

assumptions is successful, then I have an argument for them. If not, I might have to revisit my assumptions.

Here is how I make my case. In chapters 2-4 I critically discuss, respectively, Fred Dretske's, Jerry Fodor's, and Ruth Millikan's theories of representation or intentionality. The critical discussion in these chapters builds the groundwork for my theory, the *structural preservation theory of original representation*, which I develop in chapter 5 and defend from objections in chapter 6.

In 7 and 8 I use the strategy outlined above. In chapter 7 I argue that I have identified physical vehicles of representation without presupposing any substantive theory of representation. As a result, structural preservation theory is amenable to empirical test, and I argue that I have found empirical confirmation of the theory. In chapter 8, I *assume* structural preservation theory, and argue that certain brain states are representations on the grounds that my theory implies this. Then I use analysis of representational vehicles to illustrate, refine, and clarify the theory, but not to confirm it. I conclude this chapter with a discussion of how my theory contributes to the traditional set of philosophical concerns in this area, and how it can be "scaled up" to explain less basic kinds of representations. This concludes the main body of the dissertation.

Finally I have three appendices. It is commonly assumed that concepts from measurement theory are immediately applicable to a theory of representation. They are not. In Appendix A I take some results from measurement theory and use them to connect *two* empirical relational systems, for various different cases, and I outline the implications of various assumptions we might make about empirical relational systems. In Appendix B I provide a detailed literature review of the neuroscience literature from which I have drawn my examples. This is an important task because, if my theory works, then the brain states discussed in this literature are vehicles of original representation. An

understanding of the constitution *and implementation* of representation are equally important components of an adequate understanding of mind in the physical world. Finally in Appendix C I review an alternate method of typing biological relational systems based on discrimination thresholds, and show how structural preservation theory can account for this.

Chapter 2: Information and Representation

2.0 Introduction

The overarching goal of chapters 2-4 is two-fold. First, these chapters constitute the negative side of my project, where I argue that previous work on this problem has been unsuccessful, thus making room for my contribution. Second, I construct some of the foundation upon which my theory will be built, particularly in chapter 4. I approach this dual aim through critical analysis of the work of Dretske, Fodor, and Millikan on intentionality.

In this chapter I argue the following. Dretske's work on information makes an important contribution to the literature. However, despite its advertisement as "an objective commodity, something whose generation, transmission, and reception do not require or in any way presuppose interpretive processes" (Dretske 1981, vii), the notion of information drawn from Shannon and Weaver's work in communication theory (Shannon and Weaver 1949) is not objective in the sense needed for a reductive theory of representation, and so cannot be the basis for an adequate naturalistic theory. Information is very widely discussed in this literature, but the philosophical assumptions underlying this family of concepts are seldom articulated. Much confusion arises because of this, and it's important to draw apart several different senses of the word 'information'.

2.1 Information Theory

"In the beginning there was information. The word came later. The transition was achieved by the development of organisms with the capacity for selectively exploiting this information in order to survive and perpetuate their kind" (Dretske 1981, vii). Thus Dretske opens the preface to his seminal work, *Knowledge and the Flow of Information* (henceforth *KFI*). The basic thesis of the book is that knowledge is information-caused belief, and a necessary subsidiary is that information, properly understood, is an *objective*, and hence interpretation and mind-independent commodity. From simple organisms, qua physical information-processing systems, evolve genuine cognitive agents with the whole host of folk psychological, intentional states. Meaning, belief, and knowledge are all constructed out of information.

In his later work (1986, 1988, 1995, 2002) Dretske retains this thesis on information, but focuses more closely on a teleological component to representation. A representational system, for Dretske, is a system that has the function of carrying information. He constructs an intricate taxonomy involving information-carrying, different kinds of representation, different kinds of functions, and different levels of intentionality. But the basic idea is simple: to represent, a system or state must carry information, and must have the function of doing so. If the analysis of information fails as I claim, then so does the rest of the theory.

Information, in the engineering sense[25], is a measure of the reduction of possibilities. In the generation, transmission, and receipt of information, there is a *source*, a *channel* over which information is transmitted, and a *receiver*. At the source, we distinguish individual events that generate information

[25] I will henceforth omit 'in the engineering sense'. From now on when I use 'information', I always intend it in the technical sense from engineering unless otherwise stated. It should be noted that there is more than one technical sense of 'information'. Later in this chapter I will draw them apart.

from the average amount of information the source can generate. The information generated by the occurrence of a single event from a set of possible events is called the event's *surprisal value*. The average of surprisal values, weighted according to their respective probabilities, constitutes the *average information,* or *entropy.* Similarly, at the receiver, we distinguish information in a particular event from the average amount of information at the receiver. *Noise* is information that makes its way to the receiver, but which was not generated at the source, and *equivocation* is information generated at the source, but which does not make its way to the receiver. A perfect communication channel is one over which no noise is generated, no equivocation is created, and hence no information is lost. The quantity of most interest is *mutual information*[26], a quantitative measurement of the degree of covariation, or dependence, between events at the source and events at the receiver.

It is widely assumed that information from engineering theory is objective, and the task for philosophy is to figure out how to use that notion to develop a *semantic* theory of information, or a theory of informational *content*, and apply that to the problems of intentionality and representation. However, the information discussed by engineers is not objective. To make that argument it's important to give a fair exposition of the engineers' concepts and so I'll briefly introduce the equations that define them.

'$N(s_i)$' refers to the noise associated with, or relative to, a single event at the source, *s*. Each r_k is an event at the receiver. Recall that $\log\frac{1}{x} = -\log x$. Every probability lies between 0 and 1, so the logarithm of any probability is negative. To make entropy a positive quantity, the logarithm of the reciprocal of the probability is taken. This is the reason for the negative sign in front of equations (1) and (3). All logarithms are to base 2.

[26] Dretske does not so name this quantity. He says, "$I_s(r)$ [that is, mutual information] is a measure of the amount of dependency between *s* [the source] and *r* [the receiver]" (p. 16). 'Mutual information' is the standard term for this quantity. See (Cover and Thomas 2006).

$$N(s_i) = -\sum_{k=1}^{m}[p(r_k|s_i)\log p(r_k|s_i)] \quad (1).$$

Average noise $N(s)$ is the weighted sum,

$$N(s) = \sum_{i=1}^{n} p(s_i) \cdot N(s_i) \quad (2).$$

Equivocation is defined similarly. Instead of taking the probability and logarithm of $(r|s)$, we take those of $(s|r)$:

$$E(r_i) = -\sum_{k=1}^{m}[p(s_k|r_i)\log p(s_k|r_i)] \quad (3),$$

with average equivocation as

$$E(r) = \sum_{i=1}^{n} p(r_i) \cdot E(r_i) \quad (4).$$

The information associated with a single event, or the event's surprisal value, is

$$I(s_i) = \log \frac{1}{p(s_i)} \qquad (5).$$

This defines the surprisal value of the occurrence of one event out of a set of events at the source, each defined by a probability distribution function. The entropy at the source is the weighted sum:

$$I(s) = \sum_{i=1}^{n} p(s_i) \cdot I(s_i) \qquad (6).$$

$I(r)$ is defined similarly (replace 's_i' with 'r_i'). Mutual information is

$$I_s(r) = I(r) - N(s) \qquad (7),$$

or alternatively in terms of equivocation

$$I_s(r) = I(s) - E(r) \qquad (8).$$

I've followed Dretske's discussion of the concept from his 1981, which is equivalent to expositions in a standard textbook, so we can be confident that we are not dealing with an

oversimplification[27]. More substantially, notice the conditional probabilities in the definitions of noise and equivocation. This is where the action is, because this is what provides the connection between what happens at the source with what happens at the receiver.

The point to stress is that, as a quantitative measurement of statistical dependence or covariation between events, mutual information *cannot be measured* without a numerical assignment of probabilities to those events, or to the associated conditional probabilities.

Finally, Dretske doubly dissociates causation from information: information can be carried without causation and an effect need not carry information about its cause. Imagine that a source, *A*, affects two causally independent receivers, *B* and *C*. A television studio transmitting to two different television sets is an example. There is no causal link between *B* and *C*, but the states of *B* and *C* carry information about each other, because of their having a common cause. There may be a perfectly noiseless channel between the two, where the conditional probability of b_i on c_i is 1.

Any randomly caused event yields an example of causation without information. Here is one of Dretske's. Suppose that a boy cries 'wolf' randomly. Assume that on one occasion a wolf does appear, and causes the boy's 'wolf' cry. In that circumstance, the probability of the cry conditional on the presence of the wolf is equal to the probability of the presence of a wolf, but the 'wolf' cry carries no

---

[27] Cover and Thomas (1991) define mutual information $I(X;Y)$ as

$$I(X;Y) = H(X) - H(X|Y).$$

$H(X)$ is entropy (that is, $I(r)$). $H(X|Y)$ is what's known as *conditional entropy*, defined as (Cover and Thomas 1991, 16)

$$H(X|Y) = -\sum_{y\in Y}\sum_{x\in X} p(y,x)\log p((x|y).$$

To see that Dretske's $I_s(r)$ is the same as Cover & Thomas' $I(X;Y)$, note that the joint probability distribution $p(s,r) = p(s)p(r|s)$, thus making the two definitions notational variants.

information on the presence of a wolf. The cry is, in effect, all noise. Since the presence of the wolf caused the cry, there is causation without information.

Mutual information does not express a historical story about what *did* happen. What matters is a counterfactual story about what *would* happen. Dretske says:

> Questions about the flow of information are, for the most part, left unanswered by meticulous descriptions of the causal processes at work in the transmission of a signal … Knowing what caused the neural discharge [which is purported to carry information] is not enough; one must know something about the *possible* antecedents, causal or otherwise, of this event if one is to determine its informational measure (1981, 33-34).

2.2 Informational Content

I now turn to Dretske's use of this apparatus to provide a theory of *informational content*. There are three constraints that Dretske thinks that theory should satisfy. For *r* to carry the information that *s* is *F*, it needs to be the case that:

(A) The signal carries as much information about *s* as would be generated by *s*'s being *F* (1981, 63),

(B) *s* is *F* (1981, 64), and

(C) The quantity of information the signal carries about *s* is (or includes) that quantity generated by *s*'s being *F* (and not, say, *s*'s being *G*) (1981, 64).

He justifies the first constraint by appealing to the *Xerox principle*: "If *A* carries the information that *B*, and *B* carries the information that *C*, then *A* carries the information that *C*" (1981, 57). This is necessary, Dretske argues, for there to be a chain of communication, or a flow of information from one source to others. But if the Xerox principle is true "then the *amount* of information that the signal carries about *s* must be equal to the amount of information generated by *s*'s being *F*. If *s*'s being *F* generates 3 bits of information, no signal that carries only 2 bits of information about *s* can possibly carry the information that *s*is *F*" (1981, 58).

The second constraint is justified by appealing to the relationship in ordinary parlance between information and knowledge or truth. Providing someone information is a means to their gaining knowledge, which requires truth. Hence, a theory of informational content must make it the case that in order for *r* to carry the information that *s* is *F* it must be the case that *s* is *F*. Finally the third constraint is meant to provide a connection between the amount of information generated by *s*'s being *F* and *r*'s carrying the "right" information about *s*. In other words, just because *r* carries, say, 3 bits of information (the same amount generated by *s*'s being *F*), it does not follow that those three bits have anything to do with *s*'s being *F*, rather than with *s*'s being *G*.

Dretske provides the following definition of informational content, where *k* is the receiver's antecedent knowledge about the possibilities at the source:

> *Informational content*: A signal *r* carries the information that *s* is *F* = The conditional probability of *s*'s being *F*, given *r* (and *k*), is 1 (but, given *k* alone, less than 1) (1981, 65).

Dretske claims that this is a recursive definition, from which *k* can eventually be eliminated. Thus, it is supposed to define an objective commodity. This definition supposedly satisfies the three conditions, since if the conditional probability is 1 then the equivocation is 0 and the signal must carry at least as much information about *s* as *s*'s being *F* generates. Further, with a conditional probability of 1 then *s* must be *F*. Finally Dretske claims that this definition satisfies condition C, but since condition C is itself unclear (what could it be for a quantity to "include" another quantity?) it is unclear how this definition does or does not satisfy it[28].

Given a theory of informational content, Dretske proceeds to characterize different kinds of representation, different levels of intentionality, and the nature of knowledge, belief, and misrepresentation in these terms. However, we need to clarify the use of 'information', since we now have at least five different senses of it.

The first sense of 'information' is the colloquial sense, which involves meaning, content, or data, as well as knowledge. It is not a precise concept. The second sense is average information or entropy, the probabilistically weighted sum of the values of a monotonic function over possibilities at a source. This is a well-defined concept that has little to do with meaning or content. It is well-defined because, *given* certain assumptions about probability, it can be calculated precisely. However, absent antecedent assumptions that certain counterfactuals are well-defined, that prior and conditional probabilities exist, and that nomological possibility admits of probabilistic quantification, the concept of average information is ill-defined.

[28] Dretske remarks: "And condition (C) is satisfied because whatever *other* quantities of information the signal may carry about *s*, our definition assures us that the signal includes the *right* quantity (the quantity associated with *s*'s being *F*) in virtue of excluding just those situations that motivated the imposition of this requirement" (1981, 66-67). But this remark only amounts to the claim that his definition of informational content satisfies condition (C). It does not help to explicate what that amounts to, or how it does so.

The third sense of 'information', closely related to average information, is suprisal value. This is well-defined as well: it is a monotonic function of the possibilities at a source. Average information is the probabilistically weighted sum of surprisal values. It is important to recognize this as another sense of 'information' because it is frequently used in Dretske's expositions. For example, he says that a signal must carry as much information about *s* as *s*'s being *F* generates. But the generation of information by *s*'s being *F* is the surprisal value of one particular event out of the set of possible events at the source. It is not the average information.

The fourth sense of 'information' is mutual information. Given the same assumptions necessary for defining average information, mutual information is also well-defined. It measures the statistical dependence between the probability distributions of two random variables. Mutual information, like average information, has little to do with information in the colloquial or semantic sense.

The fifth sense of 'information' is informational *content*. At this point we have an explicit attempt to bridge the technical with the colloquial concepts. Dretske uses the technical concepts of average and mutual information to define informational content. It takes into account some aspects of the colloquial use of 'information', such as its connection to knowledge and truth, the transmission of meaningful information through multiple sources and receivers (this is the Xerox principle), and the connection between the carrying of information by a signal and that which the signal carries information about (this is his condition C). Dretske's informational content relies on the same assumptions about counterfactuals and conditional probability that the technical concepts do.

Finally, the sixth sense of 'information' is more or less Grice's "natural meaning" (Grice 1957). By contrast to "nonnatural meaning", which is the sort of meaning that we associate with the use of conventional signs (most obviously in writing or speaking), Grice argues that spots mean measles in a

different sense. In this use of 'meaning', "X means that *p*" implies that *p*, but does not imply that anyone conventionally/linguistically *means p* by 'X'. In his later writings, Dretske briefly describes the carrying of information in terms of Grice's natural meaning, using examples such as that smoke means (or carries information that there is a) fire, tracks in the sand mean (or carry the information) that a bird has walked by, and tree rings carry information about the age of a tree. As with Grice's natural meaning, 'X means or carries information on Y' implies that Y. There is no such thing (in this sense) as "mis-information". Clearly Dretske intends the sixth sense of 'information' to be equivalent to the fifth sense, of informational content.

2.3 Semantic Content and Belief

Dretske's theory of the content of belief builds on informational content. First, some information is *nested* in other information: "The information that *t* is *G* is nested in *s*'s being *F* = *s*'s being *F* carries the information that *t* is *G*" (p. 71). For example, if a signal carries the information that *s* is a square, it also carries the information that *s* is a rectangle, since that information is analytically nested in *s*'s being a square. For another example, if a signal carries the information that *s* is rapidly freezing water, it also carries the information that *s* is expanding[29]. As a result of nesting, informational content is not specific enough to qualify as intentional or belief content, so Dretske introduces *semantic content*.

[29] Analytic and nomic nesting, which are respectively exemplified in the text, are the only two kinds of nesting relationship, on Dretske's theory. Thus, it is not enough that all *F*s are *G*s for a signal carrying the information that *s* is *F* to also carry the information that *s* is *G*. Rather, that all *F*s are *G*s must be either analytic or grounded in a natural law.

The semantic content of a signal is the most specific piece of information carried by that signal, or, it is the informational content that is not nested in any other informational content[30].

Semantic content is still a kind of informational content, and there is no such thing as misinformation because information cannot be false. However, beliefs can be. To understand the content of *beliefs*, we need semantic content plus something else.

Dretske's solution is that symbol tokens have meaning in virtue of the type of which they are a token. This way, each token has meaning, but can be false, or tokened inappropriately, as well. The key is to explain how symbol types become the types that they are, so that they carry the particular meaning that they do. He does this in terms of teleofunction: the tokens of a particular type have the function of carrying a particular piece of information. Given that they have this function, each token *means* whatever information it has the function of carrying. Semantic content for beliefs, which Dretske calls *meaning*, is thus semantic content plus the function of carrying that information[31].

A signal gets the function of carrying a particular piece of information (say, that an *F* is present) when the organism develops a mechanism that is selectively sensitive to *F*ness. In this way, the organism learns what *F*s are by developing a structure type that is selectively sensitive to *F*. Later instantiations of that structure type then mean that something is *F*, that an *F* is present, even if they do not carry that information (hence, even if they are false). "In short, the structure type acquires its

---

[30] Dretske explicates *semantic content* in terms of *digital content*, where semantic content is the information that a signal carries in completely digitalized form:

> S carries the information that *t* is *F* in digital form if and only if that is the *most specific* piece of information about *t* that *S* carries (1981, 177) … More technically, *S* carries the information that *t* is *F* in digital form if and only if (1) *S* carries the information that *t* is *F*, and (2) there is no other piece of information, *t* is *K*, which is such that the information that *t* is *F* is nested in *t's* being *K*, but not vice versa (1981, 260, n. 7).

[31] A reminder of Dretske's taxonomy might be helpful here. We begin with the various forms of information from engineering theory, from which the concept of *informational content* is developed. *Semantic content* is completely digitalized informational content, and *meaning* (or *belief content*, or *intentional content*) is semantic content that a state has the teleofunction of carrying.

meaning from the sort of information that led to its development as a cognitive structure" (Dretske 1981, 193).

To qualify as a belief, however, a semantic structure must not only be a token of a semantic type but it must also contribute to the control of the organism's behavior. When a semantic structure partially determines output, or when the content that it carries is stored for possible future use (as in memory), then it is a cognitive structure, and is (or can be) a belief.

2.4 Critical Analysis: *Knowledge and the Flow of Information*

Dretske attempts a naturalistic reduction of the content of beliefs, which is what I have described in chapter 1 as the project of naturalizing intentionality. While this is distinct from my goal of explaining *representation*, it is nonetheless crucially relevant to my project. His account has two main components: semantic content and teleofunction. For his analysis to work, both components must be defensible.

Most of the published criticisms of this stage of Dretske's work focus on the teleological component. I will briefly review those criticisms, and then develop a further critique of the first component.

2.4.1 The Learning Period and Idealization

One well-known objection is that Dretske's theory cannot handle error or misrepresentation. Since his theory of the semantic content of belief is essentially a covariation theory, a belief will represent or have as its content whatever it covaries with, or whatever it *best* or most closely covaries with. But for error to occur, the covariance conditions that determine content cannot also be the

conditions that determine truth. Somehow, the conditions for truth and for content have to come apart, and covariance alone does not provide the resources to separate them. Dretske is aware of this, as he recognizes that informational relations do not allow for error. To solve this problem, he introduces the concept of a learning period, which is an idealized situation in which information is always carried and by definition there are no mistakes in recognizing the information (or, decoding the message). As the organism learns to recognize, or develop structures that are selectively sensitive to some particular features of the environment, those semantic structures get marshaled for use in the control of the organism's behavior. In virtue of these selectively sensitive mechanisms becoming part of the organism's control system, those mechanisms develop the *function* of carrying the information that they do. When the learning period ends, tokens of that type have the intentional content that they do in virtue of being tokens of that type. During the learning period, there is, and can be, no error. After the learning period, tokens of that type can be tokened inappropriately or in error.

The usual objection is that there is no principled distinction between the periods before and after learning (Loewer 1983; Fodor 1984; Cummins 1989). Cummins notes that organisms can be, and are, in a state of perpetual learning, conceptual change, and behavioral modification. Further, the assumption of an idealized learning period, in which the perceptual environment is always optimal and no mistakes are made, is empirically implausible (Cummins 1989). These considerations suggest that there is no such idealized learning period, and so it cannot be used to ground conditions for content in the absence of truth conditions.

A second objection is that the theory would not work even with a principled distinction marking the end of the learning period (Fodor 1984). What matters for covariation theories, and for Dretske's information-based theory in particular, are the counterfactuals, not actual history. Suppose that *R* represents *S*. During the learning period, only *S*s covaried with *R*s, and *R*s came to be part of the

organism's control system because they carried information about *S*. Once the learning period has ended, error is possible, so assume that, after the learning period, a *T* caused *R* to token. On Dretske's theory this token is a misrepresentation since *R* does not have the function of carrying information about *T*. But that doesn't work: as Dretske states, the counterfactuals are what matter, not the actual history. Even though in the actual learning period no *T*s caused *R*s to token, if a *T* caused an *R* to token after the learning period, then we should say that a *T* would have caused an *R* to token during the learning period, had it occurred then. Then the content of *R* is not *S*, but is the disjunctive content (*S* or *T*), and hence, *T*'s tokening of *R* after the learning period is not in fact an occurrence of error, but accurate representation.

According to Cummins, the only way for a covariance theorist to handle error is by appealing to idealization. Essentially, the idea is that under ideal conditions, *R* occurs when, only when, and because it is caused by *S*, and hence represents *S*. Misrepresentation occurs as a result of non-ideal conditions. Further, there are only two ways of fleshing out the 'ideal conditions' clause: in terms of proper functioning versus malfunctioning, and in terms of optimal perceptual conditions. On the first option, we idealize away from malfunctioning so that *R* is said to represent *S* in a system so long as the system is not malfunctioning. But sometimes misrepresentation occurs as a result of *proper* functioning. Many of the well-known perceptual illusions occur because, while the organism's perceptual system is functioning properly, it is placed in nonstandard or "weird" perceptual environments[32]. So this can't be right.

The other option is to idealize away from nonstandard environments. Then *R* represents *S* in a system iff, were the system functioning properly and the circumstances ideal, then *R* would occur when,

[32] The Ames room is one example of this: it is a room whose corners are not square, but when objects are placed in different locations in the room, the appearance of their relative sizes becomes distorted. The illusion is not due to malfunctioning of the visual system, but to the "nonstandard" or "weird" perceptual environment. See (Cummins 1989, 41-42).

only when, and because *S* is present. According to Cummins, this proposed solution is incompatible with an empirical assumption of the computational theory of cognition. This theory assumes that systems get into states covarying with states of its environment *because* of its representational resources, which include a great deal of stored knowledge. Hence the system infers what the distal source of its percepts are not only from the information available in the system due to what it was caused by, but also by stored information (first, nontechnical sense of 'information' here). To flesh out what the ideal circumstances are from which we are defining representation, we need to have not only proper lighting, etc., but also the right stored knowledge. And this means having the right representations with the right content. "And *that* means we cannot fill out [this suggestion] without making liberal use of the very notion that [it] is supposed to explain" (Cummins 1989, 45).

In his later work Dretske more carefully fleshes out the teleological component of his theory, but never significantly revises the informational component. Rather, in all later works he relies on his exposition as set forth in *KFI*. It is not however without its own problems.

2.4.2 The Objectivity of Information

To play a significant role in the grounding of representation or intentionality as a feature of the natural world, information must be objective, in the sense that its existence and nature are mind-independent. As Dretske notes (and I quote again for its importance), it must be the sort of thing that is "an objective commodity, something whose generation, transmission and reception do not require or in any way presuppose interpretive processes" (Dretske 1981, vii). I argue in this section that neither entropy, nor mutual information, nor Dretske's informational content are indeed objective in this sense. I make this argument from two angles. First, I discuss the background conditions that define

probabilities, and second, I ignore probability and consider the distinction between a signal and a channel.

2.4.2.1 Background Conditions Determine Probabilities

An object, event, process, or property is objective if its existence and nature are independent of (or do not rely on or presuppose) minds, cognitive states, or interpretive processes. This general way of thinking about objectivity is consistent with Dretske's use of the concept. A paradigmatic example of an objective property is length or height: the Empire State Building is 448.7 meters tall. The building has this property independent of any minds knowing this, agreeing it to be so, or interpreting it as such. However, there is a certain level of arbitrariness or social/pragmatic influence to even a paradigmatically objective property such as length: it is arbitrary that *that length* is called a "meter". But, *given that that length is one meter*, it is neither arbitrary nor a function of social convention that the Empire State Building stands 448.7 meters tall. While there are subjective factors at work in the choice of the meter as the unit of length, given the unit, it is neither arbitrary nor subjective that the building is so many units tall. By contrast, the kind of non-objectivity that would be threatening to a reductive account of mind or mental representation, and that Dretske rightly seeks to avoid, would be an explanation of representation or intentionality that directly presupposes cognitive, intentional, or interpretive states, such as Dennett's intentional stance.

The problem is in the assumptions that lie behind assignments of probability to certain events, with respect to both independent and conditional probabilities. A particular event or process *having the probability that it does* is dependent on judgments of relevance, and thus, event *e* having probability *p* is ineliminably infected with subjective, interpretive, cognitive processes, even though this is not immediately obvious. To clarify, I do not merely claim that our ability to *know* that an event's

probability is *p* is dependent on relevancy judgments. Rather, I claim that an event's *having* probability *p* is relative to judgments of relevance.

I am not attempting to arbitrate between the objectivist/frequency and the subjectivist/Bayesian interpretations of probability. For Dretske's theory of informational content to serve its purpose, we must assume an objective interpretation of probability. If we assume a subjective interpretation then the probabilities upon which entropy, mutual information, and informational content are dependent, are not objective commodities and would render a reductive account of mental representation or intentionality based on them viciously circular. I grant Dretske the objective interpretation of probability. However, this does not solve all problems.

The objectivist interpretation of probability can only be assumed relative to some background conditions with respect to what is held stable and what is not. But those background conditions are themselves subject to, and determined by, judgments of relevance. For example, whenever you toss a fair coin the probability of getting heads is ½. The frequency interpretation is that, over very many repetitions of this, the average value of the function describing the outcome will converge to ½. But the notion of a "fair coin" is just a shorthand way of setting background conditions: we assume that the coin's weight is evenly distributed, that the wind does not blow in a fashion that favors the coin's falling on one side over the other, that quantum fluctuations do not occur that cause the shape of the coin to alter, that gravitational forces remain constant, that wormholes and black holes and dark matter fluctuations do not alter "normal" physical processes, that the moon does not alter its position with respect the earth, that no massive meteor or asteroid begins to circle the earth, that no malevolent demons affect the outcome, etc. The list could go on, but the idea of it being a fair coin is to set up these background conditions. What makes these various scenarios part of the background is that they are *not relevant*. To specify the probability of the outcome of coin-flipping, we must decree these

scenarios irrelevant. But what is and is not relevant to any given situation is essentially a judgment. Further, it is important to emphasize that changing the background conditions which define what is and what is not relevant, *changes the frequency, and hence the probability of particular events*. Thus, different judgments about relevance results in the same event having a different probability, thus making specific probabilities themselves relative to cognitive states or interpretive processes.

Much of contemporary scientific research relies heavily on statistical measures, themselves based on more fundamental assumptions about probability. I am not suggesting a problem for the scientific use of that methodology, nor any general problem about the realist interpretation of scientific results and practice. There is no reason to believe that those (possibly inexpressible) relevancy judgments used to set up the background conditions upon which frequency based probability assignments are made are not themselves (at least sometimes) responding to real, objective regularities in the world. If they are, this justifies and validates claims of knowledge acquisition due to scientific results which are ultimately based on these relevancy judgments. The problem is not the scientific use of objectively-interpreted probability; it is this particular philosophical use.

Dretske endorses something like an "engineer's ideal"[33], where in order to understand something, you have to know (at least in an abstract and idealized sort of way) how one would go about building one. Fletcher (2008, 13) takes this idea and makes it more precise by endorsing what he calls the *Explicitness Condition*, which says that if a standard can be used as a basis for making objective claims about reality, that standard can be made explicit. At least ideally, it cannot be left implicit. The idea behind this is that if we cannot explicitly state what the criteria are by which we make purportedly

[33] "All I intend by my provocative claim [that if you can't make one, you don't know how it works] is that philosophical naturalism is motivated by a constructivist model of understanding. It embodies something like an engineer's ideal, a designer's vision, of what it takes to really understand how something works … If you want to know what intelligence is, you need a recipe for creating it out of parts you already understand" (Dretske 2002, 491).

objective judgments then those judgments must be subjective and interpretation-dependent[34]. For two extreme examples, we can make explicit what the criterion is for making length judgments, by pointing to the canonical meter stick as a standard against which to make length judgments, but we cannot make explicit what makes a painting beautiful by pointing to the Mona Lisa. While I do not argue that this causes a problem for the statistical inferences made in the sciences, it does cause a problem if we seek to identify representation or intentionality with information.

To put it in Fletcher's terms, the background conditions that define frequency-based interpretations of objective probability do not and cannot satisfy the Explicitness Condition, because the list of relevancy judgments cannot be made explicit. To put it in Dretske's terms, if we use notions of probability to underwrite our theory of representation or intentionality then we have not satisfied the engineer's ideal, because we have not provided a blueprint or instruction manual for making intelligence that does not itself use intelligence as one of its building materials.

2.4.2.2 The Distinction between Signal and Channel

Both Fletcher and Loewer (1983) argue that probability per se is not a critical part of Dretske's theory, and that the problem does not really lie in the invocation of probability. Since Dretske only makes use of two probabilities (1 and less than 1), they say, we can effectively ignore that part of his theory and focus on the nomic covariation. In one sense this is incorrect. A probability assignment of 1 is just as specific a probability assignment as that of any other number, and depends just as squarely on the problematic background conditions. Hence, precise, quantified probability assignments are crucial to Dretske's theory and any problems attendant on their use plague Dretske's theory. In another sense

[34] "The idea behind the Explicitness Condition is that if a standard is to support [the making of objective claims], its application can neither depend upon nor leave room for subjective interpretation. If it is in fact interpretation-independent, it ought to be amenable to explicit expression" (Fletcher 2008, 13).

Fletcher and Loewer are correct, since the probabilities are not what cause the problem, the background conditions do. What follows is a second way of approaching this that does not rely on probability.

In his précis of *KFI*, Dretske discusses the advantages and disadvantages of setting the requisite conditional probabilities to one. An apparent disadvantage, which he argues can be avoided, is that requiring a conditional probability of one leads to the outcome that little or no information is ever transmitted, because there is always some noise or equivocation, and hence the conditional probabilities are almost never one. He introduces his reply with this: "These concerns are addressed in chapter 5, a chapter that will prove tedious to almost everyone but devoted epistemologists (i.e., those who take skepticism seriously)" (Dretske 1983, reprinted in Bernecker and Dretske 2000, 112). While the concerns that he addresses are relevant to epistemological debates about skepticism, in the context of this discussion, they stem from more fundamental metaphysical questions about the nature of the world of which knowledge is purported. Hence, they cannot be so easily dismissed as abstruse trivia relevant only to professional epistemologists.

Recall that for information transmission there must be a source, a channel over which a signal is transmitted, and a receiver. There is a distinction between the signal which carries the message, and the channel which supports the signal. For mutual information to exist, the signal (or the state of the signal) needs to depend, in a lawful way, on conditions at the source. However, the state of the signal also depends on channel conditions. To differentiate the channel from the signal, Dretske says that

> The channel of communication = that set of existing conditions (on which the signal depends) that either (1) generate no (relevant) information, or (2) generate only redundant information (from the point of view of the receiver) (1981, 115).

He describes it in the précis as follows. "The framework of fixed, stable, enduring conditions within which one reckons the flow of information is what I call 'channel conditions'" (Dretske 1983, 113 in the reprint).

To illustrate this, Dretske (1981, 111-123) asks us to consider a voltmeter attached to a resistor in an electric circuit. When the device is working properly, the position of the pointer carries information about the voltage drop across the resistor. Assume that the voltage drop across the resistor is 7 volts. When attached to the voltmeter, this difference in voltage generates a flow of current through the wires in the voltmeter, which creates a magnetic field, which in turn exerts a torque on an armature, which, attached to a restraining spring, turns a pointer along the calibrated face of the instrument. If everything is working properly and the instrument is correctly calibrated, the 7 volt difference at the resistor will result in the pointer resting on the part of the face of the instrument that reads '7'. The signal which carries information about the voltage difference at the resistor is the position of the pointer. There exists mutual information between the state of the pointer and the voltage difference at the resistor, because there is a nomic covariance between these two states of affairs: as the voltage difference changes, the pointer changes its position. The position of the pointer (the signal) depends on the difference in voltage (the source). However, the position of the pointer depends not only on conditions at the source, but also on many other conditions, such as the current flowing through the wires, the magnetic field's creating a particular torque which acts against a restraining spring, and so forth. These are the channel conditions, while the voltage drop at the resistor is the source, and the position of the pointer is the signal.

The skeptic will claim that, unless it is known on independent grounds that all of the channel conditions are reliable, information about the voltage at the source cannot be transmitted by the position of the pointer. Dretske replies that the skeptic has made a correct observation, which is that the carrying of information by a signal always depends on a stable set of background channel conditions. However, he does not accept the skeptic's recommendation, which is to claim that the signal does not in fact carry information. Rather, the signal does carry information so long as the channel conditions are *in fact* reliable, regardless of whether anyone knows this to be the case. What makes the channel a channel (and not a source of equivocation or information) is its reliability, and whether or not it is reliable is independent of whether anyone knows it to be so.

Consider the voltmeter again. If the resistance in the leads (on which the amount of current flowing through the wire depends) varied from moment to moment, the position of the pointer would also vary, but without a change in the voltage supposedly being measured. Similarly, if the elasticity of the spring varied from moment to moment, then the angular torque produced by the magnetic field would cause the position of the pointer to differ, without its depending on corresponding differences in the voltage difference across the resistor. However, as Dretske says, "that the leads are *not* changing their resistance … is not a condition that generates information [and hence is a channel condition] because there are no (*relevant*) possible alternatives" (1981, 15, I have emphasized 'relevant'). It does not matter whether we *know* that the instrument is reliable, that the channel conditions are stable, and hence, that the pointer is not stuck, the spring has not lost its elasticity, that the wires and the leads are not corroded, etc. What matters for the transmission of information is that these things *in fact* are the case. What makes channel conditions reliable is not whether anyone knows that they are, but rather, whether those channel conditions have relevant or genuine alternative states. When they do, then they generate information and hence are not stable channel conditions. In the example Dretske asks us to

consider, the elasticity of the spring, the resistance at the leads, etc., do not have *genuine* alternative possible states. Thus, the carrying of information by a signal over a channel depends on the distinction between genuine or relevant alternative possible states, and non-genuine or irrelevant possible states.

For Dretske, the distinction between relevant and irrelevant is a matter of degree, and depends on the social or pragmatic interests and purposes of the cognitive agents for whom the information is being transmitted and used. Dretske asks us to consider: Does a person gain new information by going back to the front door to check that it is still locked twenty seconds after she has locked it? What about twenty minutes or twenty years? When there is no relevant possibility other than the door's remaining locked, no new information is gained. When its becoming unlocked becomes a relevant possibility, the probabilities change, and information is gained by finding that it is still locked. Just when this difference between the door's being unlocked becomes a relevant possibility is a matter of degree:

> Whether or not a signal carries a piece of information depends on what the channel is between source and receiver, and the question of whether an existing condition is stable or permanent *enough* to qualify as part of the channel, as a condition which itself generates no (new) information, is a question of degree, a question about which people (given their differing interests and purposes) can reasonably disagree, a question that may not have an objectively correct answer. When a possibility becomes a *relevant* possibility is an issue that is, in part at least, responsive to the interests, purposes, and, yes, values of those with a stake in the communication process (Dretske 1981, 132-133, emphases in the original).

Dretske holds a reliabilist, contextualist, and absolutist view of knowledge, and he identifies knowledge with information-caused belief. If there is a reliable connection between a state of the environment and a person's belief, then, regardless of whether that person or anyone else knows that the connection is reliable, we should say that the person knows whatever it is she believes. Second, he holds that whether or not someone knows that *p* in part depends on context-dependent, interest-relative factors such as the importance or salience of knowing that *p*, the knowledge that everyone else has or presupposes, etc. Third, he takes propositional knowledge to be absolute in the sense that I either do or do not know that *p*, and it makes no sense to say that, if I know that *p*, you could know that *p* better than me.

Since he discusses the difference between relevant and irrelevant alternative possibilities in the context of a discussion of information and its relation to *knowledge*, he argues that these considerations do not cause a problem, and in fact can be used to *explain* the above three properties that he takes knowledge to have. Specifically, his reliabilism with respect to knowledge comes from the reliability or unreliability of channel conditions, which do not depend on whether anyone knows them to be reliable. The contextual yet absolute nature of knowledge, he argues, is explained by the contextual yet absolute nature of information. Whether or not something is a channel, and hence, whether information gets transmitted, is subject to pragmatic, context-sensitive factors. However, *relative to* these factors, information can only be transmitted if there is a conditional probability of one. By analogy, whether or not something is empty is absolute: It is either empty or it is not. However, what *counts as a thing* for purposes of determining the emptiness of a container depends on context. Air and dust particles do not count as things when determining whether my pockets are empty, but they do when determining if a vacuum chamber is empty. Once it is specified that dust particles do not count, it is an objective matter of fact that my pockets are empty. Similarly, whether or not something counts as a relevant possibility,

and hence, whether information gets transmitted, depends on the context. Dretske asks, "How can knowledge be identified with information-caused belief if knowledge has this social, relative, pragmatic dimension to it while information does not?", and answers, "The answer, of course, is that information has this very same character" (1981, 132).

Identifying knowledge with information-caused belief might explain why knowledge has these characteristics. However, while an analysis of knowledge is an important topic, it is a distinct explanatory enterprise from the naturalizing intentionality project. One criterion for a satisfactory explanation of intentionality or representation is that it does not ineliminably advert to cognitive, intentional, or representational states. In that regard, any analysis of informational content, entropy, or mutual information that depends on the distinction between relevant and irrelevant possibilities, which itself depends on "social, relative, pragmatic" factors, infects all three of these items with cognitive, intentional states, and hence they are not useful for a naturalistic reduction of mind[35].

It is important to recognize that this critique applies whether you take the requisite conditional probabilities to be one, as Dretske does, or less than one[36]. Either way, a signal's carrying information depends on the distinction between channel and signal, or between the signal's dependence on channel conditions, and the signal's dependence on source conditions. That distinction cannot be upheld without recourse to "stable, enduring conditions", which reduces to whether there are relevant or genuine alternatives to certain conditions, which itself is relative to a cognitive judgment.

[35] Parenthetically, the analysis of knowledge doesn't work either. Since knowledge is identified with information-caused belief, it thus awaits an analysis of belief. The analysis of belief is given in information-theoretic terms, but for the reasons argued above, does not work, so neither does the analysis of knowledge.

[36] For example, Millikan's (2004) *local information* uses statistical correlations of less than one, but falls prey to a similar critique, which I discuss in 4.4.2. Similarly, Eliasmith (2000) and Usher (2001) use probabilities of less than one to define something like informational content, but each of their theories have the same problem.

Information was touted as an objective and measurable commodity. But a signal must carry information over a channel, and for a channel to exist it must be stable; no possible alteration can be relevant. But having no relevant alternative possibilities is relative to pragmatic, contextual, value-laden human interests; hence, information is not an objective commodity.

Notice the connection between the stable and enduring conditions that define a channel, and the background assumptions necessary to define probability density functions for random variables. In both cases, assumptions are made about what is possible and what is not, and those assumptions are based on whether certain possibilities are *relevant* for the purposes of the probability assignments or for defining channel conditions for information-carrying. They are, I submit, at bottom the same thing. In the same way that the frequency-based interpretation of probability depends on implicit relevancy judgments, the definition of channel conditions and hence the carrying of information, depend on implicit relevancy judgments. We have two different routes to the same conclusion[37].

2.4.2.3 Objections and Clarifications

Objection 1[38]: You have conflated the conditions for information transmission (i.e., stable channel conditions or stable background conditions defining probabilities) with our judgments that those conditions have been met.

[37] Fletcher (2008) provides a third route to the same conclusion. In this paper he argues that in Dretske's admission that information-carrying, since it is based on relevant possibilities, is a question of degree, Dretske "is admitting too little" (Fletcher 2008, 9). The problem is not merely determining a linear metric for the remoteness of possibilities (as Dretske seems to be claiming), but rather, determining the remoteness metric itself. Questions about the relevance of possibilities are *modally complex*: they depend on overall similarity relations between the actual world and other possible worlds which, notoriously, are systematically resistant to explicit formulation or precise characterization. Hence, the characterization of possibilities as relevant or not depends on implicit judgments of overall similarity, and thus "[Dretske] has got things exactly backwards: It is *information* that depends on *minds*, not the other way around" (2008, 19).

[38] Objections 1 and 2 are Michael Levin's, from comments on this chapter.

Reply: I acknowledge the distinction between a channel's being reliable and anyone knowing this to be the case. However, as Dretske remarks, a channel's being reliable is one and the same as that channel's not having *relevant* alternative states. Similarly, whether an alternative possibility is part of the background conditions that define probabilities is one and the same as whether that possibility is a *relevant* or *genuine* possibility. Neither of these are epistemological judgments about whether certain conditions have been satisfied.

The underlying problem is that the difference between channel and signal is *precisely* the difference between the judgment that possibility *P* is and is not relevant. For example, if leads attaching the wire to the terminal in the voltmeter are corroded, the pointer will change position in the absence of a change in voltage. Whether the pointer's position is a signal is dependent on whether those leads' being corroded is a relevant possibility. Whether the pointer's position *is a signal* is not dependent on whether they are corroded, nor is it dependent on whether anyone knows or believes that they are corroded. Rather, it depends on whether the corrosion of those leads is a relevant possibility. If it is, the position of the pointer does not carry information on voltage, and if it isn't a relevant possibility, then the position of the pointer does carry information on voltage[39]. But *relevance* is an intentional notion through-and-through, and hence, so is information-carrying.

Alternatively, the difference between event *e* having probability p_1 or p_2 is *precisely* the difference between the judgment that possibility *P* is and is not relevant. For example, I'll discuss an experimental paradigm in which vibrating tactile frequencies are applied to the fingertip of a monkey while single-cell intracortical recordings are made. The researchers sought to quantify the amount of

[39] I didn't want to muddle up the point in the text, but to be more careful, if the leads' being corroded is a relevant possibility, then the position can still be a signal but would carry less information on the voltage. Average information is the weighted sum of the log of the number of possibilities at the source. By including more scenarios in the set of possibilities at the source (i.e., the set of *relevant* possibilities), we lessen the amount of information any signal can carry. This doesn't change the point in the text, which is that information-carrying is relative to a cognitive judgment of relevance and thus is not an objective quantity.

information carried by different brain states, such as the firing rate of different neurons. To do this, they measured mutual information. To make that calculation, they used the following definition of mutual information, which is equivalent to equations (5) and (6) (Salinas et al. 2000, 5505), where 's' and 'r' stand for *stimulus* and *response*:

$$I = \sum p(s)p(r)p(r|s) \log \frac{p(r|s)}{p(r)}.$$

The researchers used eight stimulus frequencies, and designed the experiment so that the occurrence of each in any trial was equiprobable. As a result of the experimental design, $p(s) = 1/8$. However, is it a relevant possibility that one of the researchers accidentally or intentionally mis-programmed the stimulator software, where a different frequency was used? Is it relevant that there might be a software bug, or a loose wire somewhere that alters the stimulator's frequency? Is it relevant that the manufacturer accidentally or intentionally did not build the stimulator machine or its software to design specifications?

It doesn't matter whether any of these are *actually* the case, or whether anyone *knows* them to hold or not. It doesn't matter that our epistemic judgments about these matters are always fallible; skepticism is not at issue here. All that matters is whether they count as part of the background conditions held stable. For example, suppose the design includes eight frequencies, 5 Hz apart, starting at 5 Hz. It is possible that there is a difficult-to-detect software bug that includes those eight frequencies as well as 12 Hz into the protocol. If this is relevant, then the probability of getting 5 Hz on any given occasion is *not* 1/8, but something other than that. If this is *not* a relevant possibility, then it is

included as part of the background conditions held stable, and $p(s)$ for each frequency is 1/8. However, changing $p(s)$ changes the amount of mutual information that rate carries about frequency, as does changing $p(r)$. Thus, the amount of mutual information that firing rate carries about stimulus frequency *changes*, depending on which possibilities are relevant. But *relevance* is relative to a cognitive agent's judgment.

One might reply that, to get around this problem, we could simply stipulate that, *after* all the trial runs have been completed, the following frequencies did in fact occur …, with the following frequency of occurrence …. From that stipulation, it is a perfectly objective fact that, say, 10 Hz, occurred 12.5% of the time, and thus it had or has a probability of 1/8. But this is like saying, I flipped my coin 100 times and heads came up 56 times, therefore, the probability of getting heads in a coin-flip is .56. This misunderstands the frequency interpretation of probability: As the number of trials approaches the limit, the average function describing that experiment will converge to .5, not .56. Thus, if the coin-flip *really has* a probability of .5, it doesn't matter that in any particular run of trials the outcome did not exactly reflect that. Similarly, if the probability of observing 10 Hz *really is*, say, .09, rather than the .125 assumed in the information calculation, it doesn't matter that some particular run of trials did not reflect that.

The question of whether observing 10 Hz *really has* a probability of 1/8 depends on which possibilities are relevant, and relevance is dependent on a cognitive agent. This renders mutual information and the other concepts from engineering theory non-objective, and does so without conflating epistemic conditions with metaphysical conditions constitutive of information-carrying.

Objection 2: If what you say is correct, then you've proven too much. Shall we decline to be naturalists about radioactive decay, because the conditions that determine the probability that a uranium nucleus will disintegrate are dependent on minds?

Reply: No, we should still be naturalists about radioactive decay. Further, the use of probability and statistics to make inductive inferences is still a legitimate enterprise. The problem only arises when we *identify* representation or intentionality with information, because in so doing we have reduced mind to, essentially, itself, and thus have not *explained* mind (or representation, or intentionality). Or at least, we have not explained our target in naturalistic terms.

The background conditions that define probability are essentially relevance judgments, and those relevance judgments are, at the end of the day, probably ineffable. However, this does not imply that we are not responding to some real, objective regularity in the world in making those judgments. Given the success of science in providing theories that allow for prediction and manipulation of manifest phenomena, we probably are responding to real regularities in the world. Thus, we are justified in concluding that uranium nuclei decay at a regular rate described by our best physics.

Clarification: Are there no objective nomic regularities, since they are all relative to "relevancy judgments"?

Reply: There are real, objective, underlying regularities in the world, governed by natural laws. These regularities do what they do irrespective of humans' epistemic or doxastic states with respect to them.

While I have argued that neither representation nor intentionality can be reduced to information, this does not imply that the concept is unusable. Rather, since it is likely that our implicit

relevancy judgments are responding to objective regularities in the world, and since measures of information depend on them, it is likely that information calculations respond to some underlying regularities in the world as well. In chapters 5-8, where I develop and illustrate my theory, I make substantial use of mutual information. The various concepts of information from engineering are useful as *fallible epistemic guides* to, or *evidence* for, the existence of an underlying, lawful regularity in the universe. Just because information is not objective, this does not imply that it cannot be used as evidence for that which is. I have more to say on this later, particularly in 5.5 and 8.5.1.

Dretske's work on information theory and its application to semantics is seminal. Claims of "information-carrying" are ubiquitous in this literature. Philosophers typically either accept that the carrying of information is, in broad strokes, the correct foundation for mental representation and then try to show that it is compatible with misrepresentation, or, they reject information-carrying as the foundation for representation because they claim that it can't be made to account for misrepresentation. But for the most part, neither those that accept nor those that reject information as the foundation for representation, question its objectivity.

Dretske never abandons the thesis that information is objective and can serve as the foundation for semantics. In his later work, he refines the teleological component of his theory while continuing to rely on the informational component. Since the refined teleological component is orthogonal to the objections made with respect to information, the later versions of Dretske's theory of intentionality do not work either. I will discuss teleology in chapter 4, when I address Millikan's work.

Despite its shortcomings, Dretske's work on this topic is influential and important. Fodor for example has taken the foundations of informational semantics laid down by Dretske, and modified it to produce his asymmetric dependence theory, which I address in the following chapter.

Chapter 3: Asymmetric Dependence Theory

3.0 Introduction

The problem for informational semantics that has the received the most attention is misrepresentation. Any satisfactory theory of representation must have the resources to account for error, so a solution to this problem is crucial. Jerry Fodor provides his proposed solution from the background provided by Dretske's work. Fodor argues that informational semantics can be made to work by appealing to an asymmetric dependence of the causal/informational relationship between inaccurate tokenings of a symbol and what it represents, and the causal/informational relationship between accurate tokenings of a symbol and what it represents. In this chapter I critically review Fodor's asymmetric dependence theory.

3.1 Asymmetric Dependence: Version 1

Fodor offered three versions of his theory of content[40]. I'll examine each in turn, starting with the first version in his 1987 *Psychosemantics*.

Fodor's starting point is that there are only a handful of options for a naturalistic theory of content. He rejects anything having to do with inferential or functional role as well as teleology. He also

[40] Terminological note: I follow Fodor's conventions of using capitals to denote concepts, italics to denote properties, and single quotes to mention a symbol token. It should be noted that the symbol tokens that I refer to, unless otherwise stated, are *mental* representations, not linguistic tokens. So I am not actually mentioning 'horse' (qua orthographic sequence of marks) but referring to a mental representation that has the propositional content HORSE, which expresses the property *horse*. Fodor's expository convention already has metaphysical assumptions built in: The mental representation 'horse' has the concept HORSE as its propositional content, and that propositional content expresses the property *horse*.

rejects resemblance or image-based theories, and that leaves only causal theories. There are both actual historical and counterfactual causal theories. For the most part Fodor prefers the counterfactual variety, but in version 3 he proposes a mixed historical/counterfactual theory[41].

We begin with what he calls the *Crude Causal Theory of Content* (CCTC), which claims that "symbol tokenings denote their causes, and the symbol types express the property whose instantiations reliably cause their tokenings" (Fodor 1987, 99). He goes on to remark that "'reliable causation' requires that the causal dependence of the tokening of the symbol upon the instancing of the corresponding property be counterfactual supporting" (1987, 99), i.e.: were the property to be instantiated, then (ceteris paribus, or, under the right conditions) the symbol token would also get instantiated. That is, counterfactually, given the right circumstances, the instantiation of the property causes the instantiation of the symbol token[42]. He continues: "The Crude Causal Theory says, in effect, that a symbol expresses a property if it's nomologically necessary that *all* and *only* instances of the property cause tokenings of the symbol" (1987, 100). There are problems with both the 'all' and 'only' in this thesis, and he deals with each separately. Ultimately he retains the basic structure of the CCTC, but qualifies both clauses. It becomes "*all and *only Xs cause 'X'"[43], and the task is to articulate what the starred clauses amount to.

[41] Information theories are counterfactual theories. In principle, they do not need to be counterfactual *causal* theories, since all that mutual information, or Dretske's informational content, require, is statistical dependence between source and receiver. Typically, causation is the mechanism that mediates that dependence, so in practice, almost all informational theories are counterfactual *causal* theories.

[42] He explicates reliable causation a bit more: "I suppose that it is necessary and sufficient for such reliable causation that there is a nomological – lawful – relation between certain (higher-order) properties of events; in the present case, between the property of being an instance of the property *horse* and the property of being a tokening of the symbol 'horse'" (Fodor 1987, 99). His idea of nomological necessity as exceptionless is essentially the same as Dretske's conditional probabilities of one.

[43] Actually, I should have written "*all and *only Xs cause 'X's" (notice that the symbol token is plural). I'm going to always write "'X'" rather than "'X's" because it gets ugly and difficult to read otherwise.

In effect, Fodor's requirement, that '*A* reliably causes *B*' amounts to the nomological necessity of *B* given *A*, is identical to Dretske's requirement that the conditional probability of *B* given *A* be one. Fodor does not explicitly defend this requirement, but perhaps he is leaning on Dretske's arguments about the transitivity of information-carrying and the Xerox principle, which require the conditional probability to be no less than one. Additionally, there is a common worry that anything less than a conditional probability of one is an arbitrary cutoff point. However, this reading of 'reliable causation' is what leads to the requirement that 'all and only Xs cause 'X'', which is a difficult claim to substantiate, even with counterfactuals.

3.1.1 *Only Xs Cause 'X'

The problem with the 'only' clause is error or misrepresentation. It's not the case that, for example, only horses cause 'horse', since sometimes cows are mistaken for horses, and hence, sometimes cows cause 'horse'. This cannot be fixed by appealing to counterfactuals, by specifying the circumstances under which only horses cause 'horse'. Misrepresentation must be possible, or we don't have a satisfactory theory of representation. Second, the *disjunction problem* would arise with the counterfactuals anyway. This is the problem of error as it arises for causal theories. Since 'X' expresses whatever property it reliably covaries with, and since both Xs and Ys ($Y \neq X$) do, or would, cause 'X', it follows that the property that reliably covaries with tokens of 'X' is not the property of *being an X* but rather the disjunctive property *being an X or Y*. But then 'X' means *X or Y*, and so error is not possible. This is the problem that Dretske's learning period and other teleological considerations were designed to solve. It is generally considered the major problem for informational/causal theories.

Fodor's solution is as follows. He notes, first, that "It's an old observation – as old as Plato, I suppose – that falsehoods are *ontologically dependent* on truths in a way that truths are not

ontologically dependent on falsehoods" (1987, 107). With this as a motivation, the basic idea is that the causal connection between non-Xs and 'X' is asymmetrically dependent on the causal connection between Xs and 'X'. In other words, while both horses and cows can cause 'horse', if horses didn't cause 'horse', then neither would cows. However, horses would still cause 'horse' even if cows did not. So the causal connection between cows and 'horse' is asymmetrically dependent on the causal connection between horses and 'horse'. Notice that the "simple" formulation of version 1 is articulated in terms of causal connections between individuals (that is, horses and 'horse'). But this is short for a more precise version (the "parade" version, as he calls it; this is still version 1), which is articulated in terms of nomic dependencies among higher-order properties. The simple version is:

> In a world where B-caused 'A' tokens are wild [that is, false] (and express the property *A*), the nomic relations among properties have to be such that
>
> 1. A's cause 'A's.
> 2. 'A' tokens are *not* caused by B's in nearby worlds in which A's *don't* cause 'A's.
> 3. A's cause 'A's in nearby worlds in which B's don't cause 'A's (Fodor 1987, 108-109).

A caveat is that these conditions apply synchronically. The more precise version, articulated in terms of higher-order properties, is this:

> B-caused 'A' tokens are wild only if the nomic dependence of instantiations of the property of being an 'A' tokening upon instantiations of the property of being a B tokening is itself

dependent upon the nomic dependence of the property of being an 'A' tokening upon instantiations of some property other than B (1987, 164, n. 6).

This purportedly solves the disjunction problem because error is explicitly built into the story: a token is "wild" (i.e. false) if its causal relationship to the property that caused it is asymmetrically dependent on the causal relationship between the property the token represents, and the token.

The precise ("parade") version does not appeal to causation, only nomic dependencies and higher-order nomic dependencies. Presumably the first-order nomic dependence – that which obtains between the token and the property – is causally mediated. It is not clear what mediates, or what is responsible for, the nomic dependence of the one first-order nomic dependence (between falsely-expressed property and token) on the other first-order nomic dependence (between truthfully expressed property and token). It is not immediately obvious whether Fodor owes us a story about that (higher-order) metaphysical dependence, and in virtue of what that it obtains.

Fodor's basic claim is that 'X' means X if and only if *all and *only Xs cause 'X'. The 'only' clause gets qualified with the asymmetric dependence claim. Now we'll move to the 'all' clause, whose qualification breaks down into two parts.

3.1.2 *All Xs Cause 'X'

The problem with the unqualified 'all' clause is that it isn't true. Not all horses cause mental representations with the content HORSE. A retreat to counterfactuals is again in order, but the problem is to specify them. Under what circumstances would a horse (*must* is more like it) cause a 'horse'? These are what might be called "optimal" circumstances: They are circumstances in which the

instantiation of *horse* could not fail to cause the tokening of a 'horse'. Fodor's solution is to go the empiricist route and build everything from, essentially, sensory concepts.

He begins by making a sharp distinction between sensory/observable properties and concepts and theoretical ones. Concepts like PROTON are in the theoretical category, as are concepts like HORSE. The two parts of his qualification of the 'all' clause involve sensory concepts and non-sensory, theoretical concepts.

For the first part, Fodor appeals to psychophysics, the science that specifies those "optimal" conditions under which certain properties (the observables, or sensory properties, or the properties that are directly transducible by our sensory organs) always result in a particular mental state[44]. The specification of those conditions supports counterfactuals and is stated in nonintentional and nonsemantic terms, such as 'wavelength', 'rod', 'cone', etc. For those special properties at least, Fodor claims that we already have a science that nonintentionally specifies the optimal circumstances under which the mental state 'red' (for example) *must* occur.

There is a distinction between seeing and seeing *as.* For example, seeing a horse does not imply that one sees it *as a horse*. The latter requires something like the application of the concept HORSE to the thing in the world which the representation is about. For the alleged "psychophysical concepts" and their properties, the distinction between seeing and seeing as breaks down. This is what makes them psychophysical concepts: given the optimal circumstances described by psychophysics, were that property to be instantiated, the psychophysical concept could not fail to be tokened[45]. So the first part

[44] "If (enough of the) wall is (bright enough) red", he writes, "and if you're close (enough) to the wall, and if your eyes are pointed toward the wall and your visual system is functioning, then the Mentalese equivalent of 'red there' will get stuffed into your belief box *willy-nilly*" (Fodor 1987, 112).

[45] "But what makes RED special – what makes it a 'psychophysical concept' … – is that the difference between merely seeing something red and succeeding in seeing it *as* red vanishes when the observer's point of view is

is easy: psychophysics describes those very optimal conditions needed to describe the relevant counterfactuals, and hence, "psychophysics naturalizes the semantics of a certain – relatively quite small – set of mental representations; viz., those for which the distinction between seeing and seeing as vanishes in psychophysically optimal circumstances" (1987, 117).

This approach won't work for all properties. Partially this is because not all properties are transducible as such, but further because for most properties, there aren't any circumstances such that, given the instantiation of the property (say, *horse*), the symbol for that property ('horse') *must* be tokened. Certainly you might see the horse in psychophysically optimal circumstances, but this does not imply that you see it *as* a horse, and so does not imply that you'll have the right *intentional content*. The second part of Fodor's story builds on the first.

While psychophysics can't describe the optimal circumstances in which 'horse' will be tokened, it can guarantee the optimal circumstances in which the corresponding psychophysical concept HORSY LOOK (or something like that) will be tokened. As Fodor writes, "*horse* isn't a psychophysical property …; but instantiations of *horse* are, very often, causally responsible for instantiations of what are psychophysical properties" (1987, 118). The causing of those corresponding psychophysical properties is reliable, and since psychophysics can (at least ideally) describe the optimal circumstances for the tokening of a mental representation of any psychophysical property, it follows that we have a reliable causal chain from *horse* to the mental representation 'horsy look'. Fodor intends for this to apply even for concepts that are "more" theoretical than HORSE, such as PROTON. There are circumstances in which *proton* reliably causes certain corresponding psychophysical properties; namely, during experimental circumstances, when great care is taken to ensure that the changes in the photographic

---

psychophysically optimal. You can't – or so I claim – see something red under psychophysically optimal circumstances and *not* see it as red" (1987, 117).

plate, resulting in psychophysical properties, indeed are caused by the presence of *proton*, and not something else. Here physics plus psychophysics guarantees the relevant counterfactuals between properties in the world and something being in the belief box. But it isn't the right thing, since Fodor doesn't need 'horsy look' or 'protonish look' to be tokened, he needs 'horse' and 'proton'.

The final part of his story involves mediation by theory, knowledge-structures, or inference. That is, given the right background theory, a person who has 'horsy look' tokened will be caused to token 'horse'. Similarly, a person who has 'protonish look' tokened, and who has the right background theory (say, a physicist), will be caused to token 'proton' in her belief box. It is in virtue of background knowledge, and inferences, that observable or psychophysical concepts in the belief box get converted into theoretical concepts in the belief box. This is, of course, a question-begging version of the story, since it appeals to inferences and knowledge structures and hence intentional/semantic notions that make the naturalization project fail. Fodor is unconcerned. His non-question-begging version of this final element of the theory is merely to assert that *there are* mechanisms that reliably mediate between properties in the world and representations in the mind, and that he need not explain or characterize them. He writes,

> But though protons typically exert causal control over 'protons' via the activation of intentional mechanisms [those that cause the inferences based on knowledge structures; theory, for short], a naturalistic semantics doesn't need to specify all that. All it needs is that the causal control should actually obtain, *however* it is mediated. The claim, to put it roughly but relatively intuitively, is that it's sufficient for 'proton' to express *proton* if there's a reliable correlation between protons and 'protons', effected by a mechanism whose response is specific to

> psychophysical traces for which protons are *in fact* causally responsible. And *that* claim can be made in nonintentional, nonsemantic vocabulary. It just was … For purposes of semantic naturalization, *it's the existence of a reliable mind/world correlation that counts, not the mechanisms by which that correlation is effected* (1987, 121-122).

To recapitulate: 'X' means X if and only if *all and *only Xs cause 'X'. '*Only' gets cashed out in terms of asymmetric dependence. '*All' gets cashed out in two parts, based on the observation/theory distinction. For observable properties ("psychophysical properties"), the science of psychophysics describes the optimal circumstances that underwrite the necessary counterfactuals. For theoretical, non-observable properties (the ones where there is a seeing/seeing as distinction), there is a proprietary set of psychophysical properties that are reliably caused by each property, and psychophysics can describe the optimal circumstances for that. Then, knowledge of theory or the making of inferences is used to guarantee that the psychophysical concepts in the belief box get converted into theoretical concepts, thus completing the causal chain from property in the world to mental representation in the mind – from *horse* to 'horse'. However, all we can say is *that there is* a mechanism that mediates between psychophysical concept in the belief box and theoretical concept in the belief box, but we can't say anything about it, on pain of failing the naturalism constraint. So we'll just sort of keep mum about the fact that the mechanism involves inferences and knowledge of a theory.

3.2 Version 1 Doesn't Work

There are several fairly obvious reasons why this doesn't work, and it doesn't take long before Fodor begins to alter the theory. However, the revised versions 2 and 3 are only superficially but not substantially different from version 1 (as I argue), so they are still vulnerable to some of the same objections. I begin with his claim that a naturalistic semantics need not specify the mechanisms by which the property causes the symbol to be tokened.

Fodor claims that, so long as these mechanisms exist and can be denoted without using intentional/semantic terms, it is not a requirement that those mechanisms themselves be explained or described. The denoting part is easy: just use 'those mechanisms'. Clearly, simply adverting to "those mechanisms", and then remarking, *sotto voce* as it were, that those mechanisms are inferences generated by knowledge structures and the application of existing theory is unhelpful, not explanatory, and above all tacitly circular.

While Fodor need not explain every *mechanism* that mediates *horse* and 'horse', he must, according to his own requirement, specify the counterfactuals themselves. He must specify the circumstances under which *horse* causes 'horse'. However, part of what is involved in that specification is the indispensable appeal to intentional mechanisms. Even if psychophysics specifies the circumstances under which *horse* causes 'horsy look', it does not specify the circumstances under which 'horsy look' causes 'horse', and thus Fodor must extend the counterfactual specification to account for that additional step. In doing so, if he appeals to knowledge structures and inferences, he has not explained the semantics for 'horse' according to his own naturalistic requirement[46].

[46] According to Fodor, "what we want at a minimum is something of the form *'R represents S' is true iff C* where the vocabulary in which condition *C* is couched contains neither intentional nor semantic expressions" (Fodor 1984, reprinted in his 1990, 32). In the case under consideration, to be blunt, Fodor is trying to cheat, because he clearly does and must appeal to knowledge structures and inferences to specify condition *C*. If he need not specify those circumstances, what's the point of appealing to psychophysics? Why not simply say, by definition, that 'horse' represents *horse* iff all and only *horses* cause 'horse'?

On a related note, consider Cummins' (1989) objection against all covariance theories. The problem with any covariance theory, Fodor's included, is that the only way that it can deal with error is through idealization. Fodor's appeal to the "optimal circumstances" for perception, which he claims can be naturalistically specified by psychophysics, is an appeal to ideal conditions. But he must also appeal to ideal conditions for the *inferences* that he claims mediate between psychophysical concepts and theoretical concepts. For example, the cognizer should not be intoxicated, recently hit over the head, or a poor inference-maker in the absence of external circumstances like trauma or drugs. This leads to two separate objections.

First, Cummins' objection: An appeal to ideal circumstances that guarantee the tokening of the mental representation in question is inconsistent with the assumptions and/or findings of cognitive science, which tell us that, even when the perceptual machinery is working perfectly and the environment is "normal", no mental representation, including the so-called "psychophysical"/perceptual ones, is guaranteed to be tokened.

The second objection, related to one of Cummins', is that Fodor cannot specify the ideal conditions for inference without relying on intentional terms such as knowledge structures and correct inference. So his (silent, oblique) appeal to inferential mechanisms as causal mediators between psychophysical concepts and theoretical concepts is not only unexplanatory, it cannot be cashed. He still needs to appeal to ideal conditions for inference which cannot be done without presupposing intentional notions. The next group of objections involves his appeal to the theory/observation distinction.

Fodor needs to make a strong distinction between theoretical concepts and properties, and sensory/observable/psychophysical ones. He is aware of this, and argues that

> It doesn't … matter … that the observation/theory distinction isn't epistemologically or ontologically principled; for, in fact, I'm not wanting to do any epistemological or ontological work with it. All that matters is that there are concepts (Mentalese terms) whose tokenings are determined by psychophysical law (1987, 113).

But that doesn't follow. If there is no principled distinction between theoretical and observational concepts or properties, then there is no difference between the two, and the metaphysical work that Fodor surely *does* want to do with this distinction can't be done. You can't base laws on a non-existent distinction. Further, there is textual evidence that Fodor surely does rely on a robust and principled distinction between theory and observation. Consider the following:

> But what makes RED special – what makes it a 'psychophysical concept' … – is that the difference between merely seeing something red and succeeding in seeing it *as* red vanishes when the observer's point of view is psychophysically optimal … That is, perhaps, the hard core of truth that underlies the traditional doctrine of the 'theory neutrality' of observation: qua intact observers, we do have some concepts we token willy-nilly under circumstances about which psychophysicists can tell us the whole story. Perceptual applications of such concepts are, in that sense, independent of – not mediated by – the perceiver's background of cognitive commitments (1987, 117).

Fodor surely does need, and rely on, a principled distinction between theory and observation. Without it, his qualifications to the 'all' clause are ill-defined.

Finally, Fodor's implicit appeal to a one-to-one matchup between instantiations of non-psychophysical properties such as *horse* and psychophysical properties such as *horsy look* is problematic. The disjunctive set of retinal stimulations caused by horses, of different species, colors, and sizes, when viewed from different angles, in different lighting, and at different distances, is very large, perhaps potentially unlimited. But further, there is no reason to focus exclusively on retinal stimulations, as horses also have a peculiar smell, sound, texture, warmth, etc. It certainly seems a stretch to claim that there is a one-to-one matchup between *horse* and some set of psychophysical properties including retinal, tactile, auditory, olfactory, and even gustatory stimulations. It is a further stretch to claim a one-to-one matchup for *every* non-psychophysical property with a primitive predicate in the language of thought. Without this matchup, the reliable causal chain from *horse* to 'horse', *cow* to 'cow', or *proton* to 'proton', does not exist.

Fodor soon becomes aware of this problem, as he abandons the claim in his (1990) (see p. 109 of that work and below). I further address this issue below. Here, Fodor is diligently working to save the 'all' clause in the Crude Causal Theory. Later, he abandons the 'all' clause, so prima facie it is reasonable to abandon the theoretical commitments once used to justify it.

For the reasons set forth above, version 1 does not work. Most of the problems were centered on qualifications of the 'all' clause in the CCTC. In the next section we'll examine his refined versions 2 and 3.

3.3 Asymmetric Dependence: Versions 2 and 3

Versions 2 and 3 of the asymmetric dependence theory are found in Fodor's (1990) "A Theory of Content II: The Theory", to which I now turn.

Deep down, Fodor says (1990, 90), the disjunction problem isn't really a problem about error; it is a problem about the difference between meaning and information. Information, he claims, is about causal etiology, where the same symbol type, if caused by different things, can carry different kinds of information. The meaning of a symbol is what all of its tokens have in common, regardless of how they were caused. "So, information follows etiology and meaning doesn't" (1990, 90)[47].

Meaning has a property that he calls *robustness*. He explains robustness with an example early in the paper, and provides a more explicit characterization towards the end: "But surely this underestimates what one might call the *robustness* of meaning: In actual fact, 'cow' tokens get caused in *all sorts* of ways, and they all mean *cow* for all of that[48]." (1990, 91). Later he says, "The dependence of C's on B's is robust only if there are non-B-caused C's" (1990, 118), or to put it in the same terminology as the example, the dependence of 'cow' on cows is robust only if there are non-cow-caused 'cow's. The robustness of meaning amounts to the fact that symbol tokens mean what they do in spite of having various causal etiologies. So Fodor's task is to square the robustness of meaning with a reduction of meaning to information, which does respect causal etiology.

[47] Parenthetically, Fodor misunderstands *information*, which is not about causal etiology, but about statistical dependence or the reduction of possibilities at a source.

[48] Fodor uses double quotes rather than single quotes to mention symbol tokens in (Fodor 1990). In the text I stick with the earlier convention of using single quotes to mention, double quotes to quote, and double quotes as scare quotes to signal a non-standard or ironic use. When quoting Fodor I convert double quotes used to mention into single quotes.

Fodor's explicit characterization of robustness implies something stronger than the mere possibility of error; it implies that in order for a symbol token to have meaning at all, there must actually be false tokens. If 'cow' means *cow*, then there have to be instances in which non-cows (falsely) elicit 'cow'. I leave it open whether that is an acceptable result or not. Second, Fodor's characterization of robustness is curious, when considered in conjunction with his thesis that "meaning is information (more or less)" (1998, 12). Information respects causal etiology (on Fodor's view of information), and robustness is the property of not respecting causal etiology. On the face of it, the claim that meaning is robust is the claim that meaning is *not* information. For now, I'll just note the (prima facie) contradiction.

Fodor's task is to square the intuition that meaning is information with the intuition that meaning is robust; or, it is to provide for the possibility of error within a causal-informational theory. Version 2 of the asymmetric dependence theory is as follows: Assume that both cows and cats cause 'cow' tokens. 'Cow' means *cow* and not *cow-or-cat* because the existence of cat-caused 'cow' tokens depends on the existence of cow-caused 'cow' tokens, and not the other way around. Or, noncow-caused 'cow' tokens are asymmetrically dependent on cow-caused 'cow' tokens. "'Cow' means *cow* because *but that 'cow' tokens carry information about cows, they wouldn't carry information about anything*" (Fodor 1990, 91). In this way, false tokens are metaphysically dependent on true ones, thus respecting the guiding intuition with which he started.

The remainder of Fodor's paper consists of a multitude of objections and his replies, and through discussion of these, he refines and clarifies the theory. Before I turn to evaluation, we should note the difference between his earlier and later theories of content.

Version 2 does not assume that there are naturalistically specifiable conditions in which only cows cause 'cow', and neither did version 1: Asymmetric dependence allows for non-cows to cause 'cow', yet 'cow' mean *cow*, on both versions. But in version 2, he does not attempt to specify naturalistic conditions in which *all* cows cause 'cow'. In the first version this was attempted through the dubious strategy of appealing to a theory/observation distinction, psychophysics, unnamable inferential mechanisms, and one-to-one matchups between properties such as *horse* and sets of psychophysical properties. He says in the later work, "Nor does [version 2] assume that there are nonquestion-beggingly specifiable circumstances in which it's semantically necessary that *all* cows would cause 'cows'" (1990, 91). In the footnote to that sentence on he writes:

> Compare *Psychosemantics* …, in which I took it for granted – wrongly, as I now think – that an information-based semantics would have to specify such circumstances. As far as I can tell, I assumed this because I thought that any informational theory of content would have to amount to a more or less hedged version of "all and only cows cause 'cow's". This, too, was a failure to take semantic robustness sufficiently seriously. It's no more plausible that there are nonquestion-beggingly specifiable situations in which it's semantically necessary that all cows cause 'cow's than that there are such situations in which, necessarily, only cows do. How *could* there be such circumstances in which the content of a thought guarantees that someone will think it (1990, 131-132)?

Fodor does not provide an argument for dropping the 'all' clause, other than his rhetorical question and an appeal to the robustness of meaning. For robustness, if 'cow' means *cow* then there

are actual cases in which non-cows cause 'cow', and so it is impossible to specify the circumstances in which no non-cows cause 'cow'. But this doesn't imply that there are no circumstances in which, although no non-cows cause 'cow', all cows do.

Fodor's diagnosis of his own requirement for specifying the circumstances in which all cows cause 'cow' is that he didn't take robustness seriously. My diagnosis is that it is based on his views of information and reliable causation. As he states in his (1987) (I paraphrase), '*A* reliably causes *B*' amounts to the nomological necessity of *B* given *A*, and this entails that the conditional probability of *B* given *A* must be 1 for *B* to carry information about *A*. If information is reliable causation and reliable causation amounts to nomological necessity, then so does information. Nomological necessity entails the counterfactual "All cows, in certain circumstances, would cause 'cow'". Hence there seems to be a need to specify those circumstances for a naturalistic reduction.

Finally, taking robustness seriously leads him to deny this: Robustness is the claim that meaning is not information, so we don't need to specify those circumstances. But if meaning is not information, then why are we trying so diligently to reduce it to information? The contradiction reappears.

I will briefly describe version 3, which is essentially version 2 with the addition of an actual causal history requirement. Fodor claims that version 2 implies a sort of verificationism, which can be avoided by appending an actual causal history requirement:

> 'X' means X if: (i) 'Xs cause 'X's' is a law; (ii) Some 'X's are actually caused by Xs, and (iii) for all Y not=X, if Ys actually cause 'X's then Ys causing 'X' is asymmetrically dependent on Xs causing 'X' (1990, 121).

3.4 Versions 2 and 3 Don't Work

One widely discussed case in this literature involves frogs, flies, and BBs. In an experimental environment, frogs presented with erratic, fast-moving little black dots, such as small BBs, will snap at them. It is claimed that a theory of content should decide whether the intentional objects of a frog's snaps are flies or little black dots (or something like, *flies-or-BBs*). Asymmetric dependence theory decides the case in favor of little black dots, because the frog's snappings at flies are asymmetrically dependent on its snapping at little black dots. If it didn't snap at little black dots it wouldn't snap at flies (because all flies are little black dots) but it does snap at little black dots even though it does not snap at flies. This engenders the following objection.

Objection: How do you avoid saying that frogs are really snapping at their retinas (Fodor 1990, 108-109)? The frog would not snap at flies (or little black dots) except for some proximal retinal stimulation. So, if there were no retinal stimulation, the frog would not snap at the black dots, yet the frog might still snap if the retinal stimulation were present but not the little black dots/flies. The snappings at flies (de re) is thus asymmetrically dependent on the retinal stimulation. Thus, the objection goes, on Fodor's theory the intentional object of the frog's snapping is really its own retina.

Fodor's reply: There are *no* specific arrays of proximal stimulation upon which snapping at flies depends. "Since, – due to the laws of optics, inter alia – cows [for example] are mapped one-many onto their proximal projections, the mechanisms of perception – constancy, bias, sharpening, and the like – must

map the proximal projections many-one onto tokenings of COW" (1990, 109). Further, he doesn't think that a disjunction of specific retinal projections would work either:

> It might still be said, however, that the dependence of cow thoughts on distal cows is asymmetrically dependent on their dependence on *disjunctions* of proximal cow projections; distal cows wouldn't evoke COW tokens but that they project proximal whiffs or glimpses or snaps or crackles or … well, or what? Since, after all, cow spotting can be mediated by theory to any extent that you like, the barest whiff or glimpse of cow can do the job for an observer who is suitably attuned. Less, indeed, than a whiff or glimpse; a mere ripple in cow-infested waters may suffice to turn the trick … cow thoughts do *not*, of course, owe their intentional content to the belief systems in which they are embedded (1990, 109).

So, there is no (non-open) disjunction of retinal (or other proximal sensory) states that will map cows onto 'cow' tokens, and as it is for us and cows, so it is for frogs and flies. Hence the snappings at flies (de re) are not asymmetrically dependent on any specific retinal patterns of stimulation and the claim that the intentional object of a frog's snapping is its retina, is avoided.

Discussion: It is interesting to note that exactly the opposite claim was made to support Fodor's defense of the qualified 'all' clause in version 1. In that version he assumed that there *is* a limited set of proximal stimulations that can be defined psychophysically, and for which psychophysics can define the optimal circumstances for their instantiation, associated with each property. This included both *horse*

and *proton*. Whereas in the second version he uses the negation of that claim to defend his theory from a version of the causal chain problem. The causal chain problem is another common problem for causal theories of content: if a symbol represents what it was caused by, where in the causal chain should we put its content? If 'cow' means *cow* because it was caused by cows, why not say that 'cow' means pattern of photons/electromagnetic wavelengths, or retinal stimulation, or thalamic activation, etc.?

The question at this point becomes, in dropping the 'all' clause, what has he dropped? The route from the Crude Causal Theory, which says that symbol tokens represent what reliably causes them, to, "'X' means *X* if and only if all and only Xs cause 'X'" was fairly direct: Reliable causation amounts to the lawful relation between the cause and its effect, which amounts to "X causes 'X'" is counterfactual supporting, and finally to "in the right circumstances, all Xs would cause 'X'" (ignoring the 'only' clause). Then the circumstances need to be specified in naturalistically acceptable terms. Does he have a different conception of law, cause, or counterfactuals? I argue that, actually Fodor has not dropped the 'all' clause. It is now hidden and we'll have to dig it out.

Besides the obvious change of no longer explicitly defending a hedged version of "all and only Xs cause 'X'", the major change that took place from the 1st to the 2nd version of the theory concerns what asymmetrically depends on what. The official version of 1 talks of higher-order instantiations of properties[49]. Specifically, there is a first-order nomic dependence between *being a horse* and *being a 'horse' token*. The asymmetric dependence, or, the higher-order dependence, is of the first-order nomic dependence between *being a 'horse' token* on *being a non-horse*[50] and the first-order nomic dependence of *being a 'horse' token* on *being a horse*[51]. Causation is not explicitly mentioned here.

[49] A crucial reminder: I use italics to refer to properties. Thus, *horse* is the first-order property that horses have, whereas *being a horse* is a second-order property that *horse*s have.
[50] Actually, it's not the property of *being a non-horse*, it's the property of *being an F*, where *F* is not *being a horse*.
[51] For my reader's convenience, here is the quote again for comparison, from his (1987, 164, n. 6): "B-caused 'A' tokens are wild only if the nomic dependence of instantiations of the property of being an 'A' tokening upon

However we can presume that what mediates the first-order nomic dependencies is causality, because of the unofficial articulation (the "pocket version", as he calls it), which goes like this. 'X' means *X* if Xs cause 'X', and if non-Xs cause 'X', their causing of 'X' is asymmetrically dependent on Xs' causing 'X'[52].

Version 2 still speaks of a dependence between instantiations of higher-order properties, but different properties come into play: The dependence is between *being a horse* and *being a cause of 'horse' tokens*, whereas in the first version the second property mentioned was *being a 'horse' token*, not *being a <u>cause of</u> 'horse' tokens*. He writes,

> 'cow' means *cow* if (i) there is a nomic relation between the property of being a cow and the property of being a cause of 'cow' tokens; and (ii) if there are nomic relations between other properties and the property of being a cause of 'cow' tokens, then the latter nomic relations depend asymmetrically upon the former (1990, 93).

I argue that he still has to deal with the problem from version 1 of providing the naturalistically specified ideal conditions in which all cows would cause 'cow'. The reason, basically, is that in appealing to *the cause of 'cow' tokens* he is in effect appealing to the same nomic relationship in version 1, where all cows, under certain circumstances, would cause 'cow'.

Let:

---

instantiations of the property of being a B tokening is itself dependent upon the nomic dependence of the property of being an 'A' tokening upon instantiations of some property other than B".

[52] Here's the quote again, from (1987, 108-109): "In a world where B-caused 'A' tokens are wild [that is, false] (and express the property *A*), the nomic relations among properties have to be such that

1. A's cause 'A's.
2. 'A' tokens are *not* caused by B's in nearby worlds in which A's *don't* cause 'A's.
3. A's cause 'A's in nearby worlds in which B's don't cause 'A's".

C = is a cow

G = is a cause of *C*s

T = is a 'cow' token

I want to show that Fodor's appeal to something's being a *G* has the consequence that, if something is a *C*, then that thing in certain circumstances would cause a *T*, and hence, that all *C*s, in those circumstances, would cause *T*s.

1. "*C*s cause *T*s" supports counterfactuals.

This is just the claim that there is an informational relationship between cows and 'cow' tokens. The informational relationship between non-cows and 'cow' tokens asymmetrically depends on the informational relationship cited in 1. Fodor wants us to read this as follows.

2. "If x is *C* then x is *G*" is a law.

He remarks, "if the generalization that Xs cause Ys is counterfactual supporting, then there is a 'covering' law that relates the property of being *X* to the property of being a cause of *Y*s: counterfactual supporting causal generalizations are … backed by causal laws, and laws are relations among properties" (1990, 93). This is where the property *G*, that of being a cause of 'cow' tokens, enters the story. If a

statement is a law, Fodor remarks that "the satisfaction of the antecedent of a law is nomologically sufficient for the satisfaction of its consequent (I'll sometimes say that the truth of the antecedent of a law *nomologically necessitates* the truth of its consequent)" (1987, reprinted in his 1990, 144). So from here we get

3. If x is *C* then necessarily x is *G*,

with the modal operator being read as nomological necessity. Next, note that the predicate 'is *G*' stands for *is the cause of Ts*, so by definition, we get the following.

4. Things that are *G* cause things that are *T*.

Fodor holds a commonly accepted notion of causation, which he makes explicit in his (1987). He writes, "If an event *e*1 causes an event *e*2, then there are properties *J*, *K* such that: (i) *e*1 instantiates *J*, (ii) *e*2 instantiates *K*, and (iii) '*J* instantiations are sufficient for *K* instantiations' is a causal law"[53]. Applying that here we get

5. "*G* instantiations are sufficient for *T* instantiations" is a causal law.

[53] This is from p. 142 in the (1990) reprint. For clarity I've changed his predicate symbols from '*F*' and '*G*' to '*J*' and '*K*'.

Once again applying his definition of a law, we get the following.

6. If x is *G* then, necessarily, there exists a y that is *T*.

Finally by transitivity from 3 and 6 we get this:

7. If x is *C* then necessarily there exists a y that is *T*.

Note that the necessity in 7 is nomological necessity. If the antecedent of a conditional nomologically necessitates its consequent then the conditional (sans modal operator modifying the consequent) is a law (see p. 144 of Fodor's 1990), hence:

8. "If x is *C* then there exists a y that is *T*" is a law.

But since the kind of law that connects *G*s to *T*s (and hence *C*s to *T*s) is a *causal* law, we are justified in moving to this claim:

9. "If x is *C* then x causes a *T*" is a law.

Fodor explicitly connects lawhood with subjunctives as follows: "[W]hat laws subsume a thing is a matter of its *subjunctive* career; of *what it would do* (or would have done) *if* the circumstances were (or had been) thus and so" (1990, 58). So from 9 we get

10. If the circumstances were thus and so, then if x is *C* then x would cause a *T*.

Finally, since this is supposed to be a general law connecting properties and not individuals, we get to our target:

11. Under certain circumstances, all *C*s would cause a *T*.

Thus, version 2 has the same consequence as version 1. Version 2 says that "'cow' means *cow* if (i) there is a nomic relation between the property of being a cow and the property of being a cause of 'cow' tokens; and (ii) [the rest of the asymmetric dependence story]" (1990, 93). But the nomic dependence of the property of being a cow and the property of being a cause of 'cow' tokens implies that, under certain circumstances, *all cows would cause 'cow'*, and this in turn implies that (here I quote from *Psychosemantics*):

> The viability of the Causal Theory [as well as version 2 here under consideration] depends on its being able to specify (in naturalistic vocabulary …) circumstances such that … in those circumstances, 'horse's covary with horses; i.e., instantiations of *horse* would cause 'horse' to be tokened in my belief box… were the circumstances to obtain… Just which circumstances are those, pray (1987, 111-112)?

And he later remarks,

> It's no more plausible that there are nonquestion-beggingly specifiable situations in which it's semantically necessary that all cows cause 'cow's than that there are such situations in which, necessarily, only cows do. How *could* there be such circumstances in which the content of a thought guarantees that someone will think it (1990, 131-132)?

Fodor's apparent dropping of the 'all' clause of the counterfactual causal theory is only apparent. However, as Fodor correctly alludes with his rhetorical questions above, there aren't any circumstances in which it's nomologically necessary that the content of a thought guarantees that someone will think it. This becomes especially apparent when we consider that, regardless of the possibility of multiple realizability, human mental states are realized by the brain, and brain states are composed of neural states.

Many of the biochemical mechanisms responsible for neural activity are stochastic processes. Ion diffusion across the membrane, vesicle release and synaptic reuptake, and many of the mechanisms that open and close ion channels are stochastic processes. Thus, there will always be a certain amount of "noise" in the brain, and there will always be bits of energy at the periphery that do not get transduced into electrochemical changes in neurons. Even at the most rudimentary levels involving directly transducible properties, such as mechanical, thermal, or electromagnetic energy, there are simply *no* circumstances under which the instantiation of some property guarantees the instantiation of any particular brain state, and thus, no property guarantees the instantiation of *any* particular mental state. *Nothing* "gets stuffed into the belief box willy-nilly".

To clarify, Fodor has claimed that the viability of the causal theory depends on the ability to specify certain circumstances in naturalistically suitable terms. Cummins (1989) has argued that this cannot be done, because it will involve idealization which will ultimately violate the naturalistic constraint, by appealing to optimally working intentional mechanisms such as knowledge structures and inference-processes. Here I make a different claim. The problem is not with specifying circumstances in naturalistically acceptable language. The problem is that there are no such circumstances to be specified.

This is a fatal objection: If meaning, or intentional content, or representational content is identical to counterfactual causal covariation, since there aren't any circumstances in which that necessary causal connection holds, then meaning/intentional content is never instantiated. For example, suppose that 'cow' means *cow* iff all and only *cow*s cause 'cow', and suppose further that the asymmetric dependence component satisfactorily qualifies the 'only' clause. Nonetheless, 'cow' means *cow* just in case, in certain circumstances, *cow* must cause 'cow'. If there are no circumstances under which *cow* must cause 'cow', then it's not the case that 'cow' means *cow*. However, I claim that for

every property *p*, there are no circumstances in which *p* must cause 'p'. Hence, it's not the case that 'p' means *p*, for all *p*. Thus, the counterfactual causal theory implies that there is no meaning, intentional content, or representational content in the actual world. Needless to say, that is false; or at least, that is not an implication that a counterfactual causal theorist would welcome.

This is a bold statement, so I'll say it again to clarify. The counterfactual causal theory says that 'p' means *p* iff *p* counterfactually causally covaries with 'p', which translates to, 'p' means *p* iff all and only *p* causes 'p', which implies that 'p' means *p* only if, under certain circumstances, *p* must cause 'p'. I claim that there are no circumstances under which *p* must cause 'p', for any property. More carefully, there are no *nomologically possible* circumstances under which *p* must cause 'p'. In terms of possible worlds, in each nearby possible world, there are *p*s that don't cause 'p', where my metric of *nearness* is determined by the laws of physics and biology: If those laws are different, that world is not nearby. Since there are no nearby possible worlds in which every *p* causes a 'p', it follows that 'p' does not mean *p*, for every *p*, and thus the counterfactual causal theory implies eliminativism or semantic nihilism.

I defend my claim that there are no circumstances under which *p* must cause 'p' on the basis of neurophysiology. The brain simply doesn't work the way Fodor apparently thinks it does; psychophysics cannot specify the circumstances under which any property guarantees that a brain will represent it[54]. Thus, the counterfactual variety of the causal theory of content cannot be made to work, and this includes all versions of the asymmetric dependence theory, since version 3 is essentially version 2 + the requirement that *p* historically has caused 'p'.

[54] I suppose that a counterfactual causal theory might be saved by multiple realizability. If there are aliens whose biology is different in such a way that there are some properties for which their instantiation guarantees a mental representation of them, this might be a counterexample to my claim. No reasonable person should want their theory of mental representation, which is supposed to explain *us*, to depend on the possibility of a specific type of alien.

I will hedge my bets. Even if my sweeping rejection of all counterfactual causal theories doesn't work, asymmetric dependence theory still fails. All of the versions are relevantly the same, and rely on cashing out the naturalistic conditions under which *cow* always causes 'cow'. Doing that requires several unwarranted assumptions, including the theory/observation distinction, the appeal to psychophysics as articulating the naturalistic sufficient conditions for the tokening of observational concepts, and the one-to-one matchup between non-psychophysical concepts and a set of proprietary proximal sensory stimulations, as well as obliquely appealing to inferences and knowledge structures, thus running afoul of the naturalism constraint.

I now turn to Ruth Millikan's theory. Several key distinctions arise through discussion of her work, and these provide a breakthrough and a foundation upon which I'll build a naturalized theory of representation.

Chapter 4: Biological Categories, Teleofunction, and Teleosemantics

4.0 Introduction: Normativity at the Foundation of Representation

The guiding motivation behind Ruth Millikan's teleosemantics is that representation crucially involves truth and falsity and thus is a normative notion. Normativity is fleshed out in terms of biological function, itself explained in terms of natural selection: Roughly, whatever tokens of a type of device *did*, which is causally responsible for the existence or preservation of that type of thing, is the teleofunction of those things. Pumping blood, for example, is what hearts do, and it is because they pumped blood in the way that they did that organisms that had hearts were differentially adapted to their environment, and so the forces of natural selection caused hearts to proliferate.

Cognitive mechanisms, Millikan argues, have biological functions in the same way that hearts do, and it is in terms of these biological functions that we should understand cognitive states and processes, and ultimately representation. Her project is the very ambitious one of providing a unified theoretical framework within which we can understand physiological systems and processes as well as cognitive and linguistic states and processes. What unifies that framework is the concept of a thing's *teleofunction*, which is to be understood in terms of natural selection.

In the present chapter, I begin with a review of Millikan's theories of teleofunction and intentionality, after which I review some objections and say why they don't work. My chief objection to Millikan's theory is simply that it is incomplete. In brief outline, she explains what it is to be a representation in terms of proper function, and then explains representational content in terms of mapping rules from representations to the world. Ultimately, it is an explication of these mapping rules that is needed for a complete understanding of representation. However Millikan's theory of these rules does not work as it stands, so we will need an alternate theory.

4.1 Biological Categories

4.1.1 A Preliminary Distinction

There is an important distinction between (i) what makes something a representation, and (ii) what determines representational content, given that a thing is a representation. By analogy, consider the distinction between (i) what makes something money, and (ii) the *value* of any particular coin or bill, given that it is money or a unit of currency[55].

One immediate insight that we can take from Millikan's work is her recognition not only that there is a distinction between (i) and (ii), but also the possibility that they might take different answers. She says in her (1984, 100), that "Conditions (1) through (4) [which define what she calls *intentional icons*, what we may understand for our purpose as representations] tell us when something is an intentional icon. They do not, however, tell us what a given intentional icon is an intentional icon *of*". Her proposed conditions answer the question related to (i): What is it for some thing to be a representation? Her answer to that question involves the notion of a teleofunction. But her answer to the question associated with (ii), which concerns what determines the content of any particular representation, is based on the notion of mapping rules. The unifying concept of a teleofunction, which explains via the same basic conceptual machinery what it is to be a heart or a kidney as well as what it is to be a representation, is orthogonal to the further question regarding what determines what a representation represents.

Each of the authors previously discussed clearly attempt to construct a theory that answers question (ii), by attempting to define the conditions under which some thing *R* represents some

[55] Michael Levin posed this analogy during discussion at my prospectus defense.

property or state of affairs. While it is not clear whether they would acknowledge Millikan's distinction, the failure to recognize it is at least an exegetical problem, as it leads to a misreading of her theory.

To understand the nature of representation we need answers to both questions. My criticism of Millikan concerns the second question, but not the first. In broad strokes at least, something like Millikan's answer to question (i) has got to be correct, for the following reason.

The human organism is a biological organism, and the fundamental and unifying theory of biology is natural selection. Since we have been created by the forces of natural selection, we should expect a unified framework for explaining both cognitive and physiological states and processes, grounded in natural selection. I take this to be almost an immediate consequence of the naturalist supposition.

4.1.2 The Theory of Proper Functions

My goal in this subsection is to outline Millikan's theory of proper functions so that I can explain her theory of intentionality. What is most distinctive about her theory of teleofunction is that it is not based on a thing's causal powers or dispositions, but rather in its history. She writes, "My claim will be that it is the 'proper function' of a thing that puts it in a biological category, and this has to do not with its powers but with its history" (1984, 17). "The simplest idea, then, would be to define a thing's function as what something like it once did that helped cause it to be, to be where it is, or to be as it is" (1993, 33).

We begin with the concept of a *reproduction*, which is much like the concept of a copy. *B* is a reproduction of *A*, roughly, if and only if they have certain properties in common, the fact that they have those properties in common can be explained by a natural law, and the natural law that explains why *A*

and *B* have these properties in common does so in terms of explaining that, whatever determinate property characterizes *A* must also characterize *B*, and this is a causal law. "Roughly, the law … implies that *had A been different* with respect to its determinate character *p* within a specifiable range of variation, as a result, *B would have differed accordingly*" (1984, 20, emphases in the original). The properties by reference to which reproduction is explained are called *reproductively established properties* (or *characters*). From this is derived the concepts of first-order and higher-order reproductively established families.

Any set of entities that have the same or similar reproductively established properties, derived by repetitive reproductions from the same original model form a *first-order reproductively established family* (1984, 23). Higher-order reproductively established families are defined in terms of proper functions, but for the moment we can rely on an intuitive grasp of a biological function to define them, and later return with a more precise definition. There are three disjunctive conditions under which a group of things constitutes a *higher-order reproductively established family*. First, a set of similar items produced by members of the same reproductively established family, when it is a proper function of the producers of these things to produce them, constitutes a higher-order reproductively established family. Second, a set of items produced by the same device, when it is that device's function to produce items such that the later ones match the earlier ones, also constitutes a higher-order reproductively established family. Third, to make room for malformed members, if an item is produced by a device whose function is to produce members of a higher-order reproductively established family, and that item is in some respects similar to the other ("Normal") members of the group, to some unspecified degree, then that item is also a member of the group which constitutes a higher-order reproductively established family (1984, 23-25).

Here are some examples. Viruses and genes are examples of reproductions, whereas mass-produced items coming off an assembly line are not reproductions of one another (one does not cause the next to have the same properties), nor are hearts or dogs reproductions of each other. Examples of members of a first-order reproductively established family would be tokens of specific genes, handshakes in a specific culture, and various tokens of the same word. Hearts, people, and dogs are members of higher-order reproductively established families. Consider hearts: They are built according to the "instructions" of the gene tokens that are responsible for building them during development. Those gene tokens form first-order reproductively established families, and it is their function to produce hearts. Similarly, mass-produced products coming off an assembly line are members of a higher-order reproductively established family, since they are each produced by the same device which has the function to produce items where later members are to match earlier members.

The concept of a *direct proper function* is this: "a function *F* is a direct proper function of *x* if *x* exists having a character *C* because by having *C* it *can* perform *F*" (1984, 26). Through natural selection, items proliferate or continue to exist because of their adaptive value, or because of what they do. If what a state or structure *x* does is differentially adaptive for the organism that has *x*, compared to organisms that don't have *x* (and hence don't do whatever *x* does), then doing that is the direct proper function of *x*. "A direct proper function is a function that an item has *as* a member of a reproductively established family" (1984, 27). Thus if any member of a reproductively established family has *F* as its direct proper function, then every member does. However, it is crucial to note that it is not necessary that any device actually realize or perform its proper function in order to *have* that function. Sperm is an example frequently cited as a biological device which has a function that tokens of that type only very rarely actually perform, which is to fertilize an ovum. "Notice that it is not necessary that a device actually serve any direct proper functions of it ... [Sperm for example rarely perform their proper

function] … Having a proper function depends upon the *history* of the device that has it, not upon its dispositions" (1984, 29).

Frequently, a device's ability to perform its function depends on its environment, and often the environment includes other biological devices performing their functions. This leads to the concepts of *stabilizing and standardizing functions*, which are important for Millikan's account of language. If hearers did not respond in lawlike ways to uttered sounds their reactions would be unpredictable and speakers would stop speaking. Similarly for hearers: they would stop listening if speakers' sounds did not correlate in lawlike fashions to things of interest to hearers. It seems likely that there is a sort of crossover point between standard uses of language devices and hearers' responses to them that contribute to both the speaker's and the hearer's ends.

> The stabilizing and standardizing proper function … of a language device is that hypothesized function … that tends at the same time to keep speakers using the device in standard ways and to keep hearers responding to it in standard ways, thus stabilizing its function … (1984, 31-32).

The environmental conditions which were historically responsible for the proper performance of a device's function underlie Millikan's concept of *Normal explanations* and *Normal conditions* for the performance of proper functions. A Normal explanation is simply "an explanation of how a particular reproductively established family has historically performed a particular proper function" (1984, 33), and "Normal conditions to which a Normal explanation makes reference are preponderant explanatory conditions under which that function has historically been performed … these are the conditions to which the device that performs the proper function is biologically adapted" (1984, 34).

Millikan uses a capital 'N' in 'Normal' to signify that she intends for the term to be read normatively, historically, and relative to a specific function, but *not* statistically (1989, reprinted in Stich and Warfield 1994, 246). Thus, Normal conditions are *not* the conditions that have historically been present most often in the environment of a biological device. Rather, they are the conditions that have been present when a biological device performed its function properly, and are necessary environmental components for the proper performance of that function[56].

A *relational proper function* of *x* is the proper function of doing or producing something that bears a specific relation to something else. Given a device with a relational proper function, and a specific context in which that device is embedded, the device's relational proper function relative to that context, is the device's *adapted proper function*. The thing that has an adapted proper function is an *adapted device*, and the thing or context to which the adapted device is adapted is its *adaptor*. For example, a chameleon's skin has the relational proper function of producing pigment that matches whatever it is sitting on. But given some specific context, such as brown and green leaves, the chameleon's skin has the adapted proper function of producing brown and green pigment. The skin is the adapted device and the context of the leaves is the adaptor (1984, 39-40).

---

[56] In a later writing she alters the terminology to "normal mechanisms", although the basic idea remains the same:

> A normal mechanism for performance of a trait's function will pretty invariably involve the presence of other things that act in cooperation with it, acting on it or being acted on by it, and it will involve the presence of various supporting conditions. In the absence of these supporting things or conditions, probably it will not be able to perform these functions (2004, 69).

To fill out the analogy between the earlier and later terminology, a normal mechanism is analogous to a Normal explanation and the various supporting conditions necessary for the performance of a trait's function are the Normal conditions.

It may seem odd that the concept of an *explanation* gets traded for the concept of a *mechanism*. This is in fact why Millikan has decided to switch the terminology. The concept of an explanation with which she was working in her earlier writings should not be thought of in terms of a set of propositions, but what those propositions are about. "In earlier writings I referred to [normal mechanisms] as 'normal explanations' or 'Normal explanations' for performance of a trait's functions. This caused some confusion, since many think of an explanation as being a set of propositions rather than what these propositions are about." (2004, 69, fn. 5).

Derived proper functions are proper functions of adapted devices that are derived from the functions of the devices that produce them. "The proper functions of adapted devices are derived from proper functions of the devices that produce them that lie *beyond* the production of these adapted devices themselves" (1984, 41). Here's an example: "Our chameleon's brown and green pattern has as an invariant derived proper function to make the chameleon invisible to predators, hence to prevent it from being eaten" (1984, 42). So the derived proper function of the skin pattern is derived from the proper function of the skin-changing mechanism; hence, to camouflage. Notice that the skin *also* has the adapted (relational) proper function of being brown and green (since it is adapted to this particular context).

Technicalities aside, Millikan's theory of biological categories is relatively simple. Biological devices are what they are in virtue of their teleofunction, and the teleofunction of a device is whatever tokens of that type of thing did in the past that had adaptive value, and hence caused tokens of that type of thing to proliferate and continue to exist.

4.2 Millikan's Theory of Intentionality

4.2.1 Intentional Icons: Mapping Rules + Function

Every author has a slightly different "pre-theoretic" take on the target explanandum. In my case, I assume (following Cummins 1989) that intentionality involves the representational properties of folk psychological beliefs and desires, and further that representation is something distinct[57]. Millikan does not conceptualize the explanandum (that is, intentionality) in exactly the same way since she takes

[57] I don't follow Cummins, as noted earlier, in the goal of understanding what the nature of representation must be for it to play the role it does in classical cognitive science. Rather, I seek to understand what I take to be the core concept of representation, common to several explanatory and theoretical endeavors.

representation to be a special case of intentional states, rather than the other way around[58]. The basic elements that we will concern ourselves with, which are closest to what I would call *representations*, are Millikan's *intentional icons*.

Peculiar to states that have intentionality, Millikan notes, is that what they signify need not exist. "Intentional signs are clearly distinguishable, though peculiar and puzzling, in that what they mean need not exist or be actual" (1984, 86). Further, intentionality is not a clear-cut phenomenon:

> There is no clean distinction between intentional and nonintentional signs or between intentional and nonintentional senses of "means". Intentionality does have to do, very generally, with what is Normal or proper rather than what is merely actual. It also has to do with mapping relations – ones that are Normal or proper rather than merely actual or average. But the notion "intentionality", like the notion "sign", is unified not by a definition but by a paradigm. Indeed, there are *two* paradigms of intentionality, an indicative paradigm and an imperative paradigm (1984, 86).

The traditional way of thinking about the problem of intentionality, inspired by Brentano, is in terms of a relation one of whose relata need not exist. The general form of Millikan's solution to this problem is to appeal to teleofunction, since a thing can have a teleofunction yet fail to perform or satisfy it. In this sense, anything that has a teleofunction has intentionality. On a more constrained usage of

[58] Although it should be noted that she does think that folk psychology is relevant to and can form the basis of cognitive science. However, her conception of the nature of folk psychology is not as a theory (qua set of laws). Rather, she argues that folk psychology posits states that have functions and whose content is determined in accordance with these functions (see Millikan 1986). I won't worry about this issue here.

'intentionality', in keeping with its traditional use, not everything that has a teleofunction has intentionality. On Millikan's view, the basic elements that have intentionality are items that she calls *intentional icons*, examples of which are sentences.

> Intentional icons exhibit a number of the most striking features of sentences. Intentional icons are devices that are 'supposed to' map *thusly* onto the world in order to serve their direct proper functions … And they are devices that are supposed to be used or 'interpreted' by cooperating devices (1984, 95).

While sentences are intentional icons, not all intentional icons are sentences[59].

There are four conditions for being an intentional icon, but condition 4 has two parts corresponding to indicative and imperative icons. In what follows note that Millikan uses 'Normally' as an abbreviation for "when performing its proper functions in accordance with a Normal explanation". Also note that she speaks in terms of sentences since they are paradigmatic; however, the definition generalizes to non-sentences. Here are the four conditions (from 1984, 96-97):

[59] Further, all *representations* are intentional icons, but neither all sentences nor all intentional icons are representations; representations are a special kind of intentional icon. We have no need to get into Millikan's technical notion of a representation here, since it is really her theory of intentional icons that is aligned with what I've been calling representations. For the record though, on Millikan's theory a representation is an icon that has the function of allowing its real value to be identified, and then the act of identifying gets explicated in terms of the way that different intentional icons, with the same real value, get used by an interpreter device. Since none of this matters for my purposes, I use 'icon' and 'representation' in the same way in order to (hopefully) avoid confusion. My reader would be best advised to ignore everything I've just said in this footnote about Millikan's use of 'representation'.

(1) A sentence is a member of a reproductively established family having direct proper functions.

(2) Normally a sentence stands midway between two cooperating devices, a producer device and an interpreter device, which are designed or standardized to fit one another, the presence and cooperation of each being a Normal condition for the proper performance of the other.

(3) Normally the sentence serves to adapt the cooperating interpreter device to conditions such that proper functions of that device can be performed under those conditions.

(4a) In the case of imperative sentences, it is a proper function of the interpreter device, as adapted by the sentence, to produce conditions onto which the sentence will map in accordance with a specific mapping function of a kind to be described below.

(4b) In the case of indicative sentences, the Normal explanation of how the sentence adapts the interpreter device such that it can perform its proper functions makes reference to the fact that the sentence maps conditions in the world in accordance with a specific mapping function of a kind to be described below.

These four conditions define what it is for some thing, some biological device, to be an intentional icon. They do not, as Millikan appropriately notes (1984, 100), define that which an intentional icon is an icon *of*. That part of the theory involves the mapping rules, to which we will turn shortly. Another distinctive element of this proposal is that, rather than focusing solely on the mechanisms of icon (or representation) production, as is the case with Fodor and Dretske, Millikan's theory focuses equally on three distinct items: the mechanism of icon production, the mechanisms or devices that consume or use these icons, and the icons themselves. That the icon itself has a function is determined by the first condition, which mandates that it be a member of a (first-order or higher-order) reproductively established family and have a direct proper function. That the producer and interpreter/consumer[60] device have functions is mandated by the remaining conditions. Keep in mind

[60] I'll use, as does Millikan, the terms 'interpreter' and 'consumer' interchangeably. In later writings she uses 'consumer' more frequently, presumably because 'interpreter' may invoke inappropriate images of an intelligent homunculus "interpreting" some signal, thus initiating a regress. This does not occur in Millikan's theory since an

that 'adapt' is a technical term, described earlier[61]. It is the job of both the producer and consumer to work together, and proper functioning of each is a Normal condition for the proper functioning of the other. The icon adapts the consumer device to some condition in the world in such a way that the consumer's being adapted to that world-condition is a Normal condition for proper performance of the consumer's function. The fourth condition serves to differentiate the manner in which the interpreter device is adapted to the world. For imperative icons, the consumer device's function is to produce the condition to which it is adapted; for indicative icons, the consumer's being adapted to that world-condition is a Normal condition for proper performance of the consumer's function, whatever that function may be.

An example should be helpful. Honeybees engage in behavioral repertoires known as *bee dances*, and these dances bear determinate relations to the location of nearby nectar. When a honeybee engages in this behavior in the presence of its conspecifics, the other bees usually fly towards the nectar. Bee dances are intentional icons. First, they are members of higher-order reproductively established families because they are produced by mechanisms which themselves have the function of producing these dances. They have the relational proper function of bearing a specific relation, defined by the mapping rules to be discussed shortly, to the direction of the nectar.

---

interpreter is simply any device that is selectively caused by icons to do something that has adaptive value. For example, whatever motor mechanism in bees responsible for causing the bee to move in the appropriate direction upon witnessing a conspecific's nectar-location dance is an interpreter.

[61] Recall that a device has a relational proper function if its function is to do or produce something that bears a specific relation to something else (such as the chameleon's skin being so-related to its surrounding environment). Given a specific context, if a device has a relational proper function, then its function to bear a relation to *that context* is its adapted proper function. The device is the adapted device and the context is the adaptor. Notice however that while Millikan is usually meticulous with her exposition and definitions, she has not actually defined the *verb*, "to adapt". Thus it is somewhat unclear how to modulate the concept of an adapted relational proper function to its verb form, where some device *adapts* some other device to some condition in the world. This may of course be my failure of exegesis not Millikan's failure of exposition. Thankfully this subtlety can be ignored since, as I've argued above, regardless of the details, something *like* Millikan's answer to question (i) must be right, although I neither endorse nor challenge the details.

Second, Normal functioning of the dance-producing and dance-interpreting mechanisms are Normal conditions for the proper performance of each. Bee dances are not learned, so both the producer and interpreter mechanisms are standardized to fit each other via evolutionary processes. In this sense, the dance itself (the icon) stands "midway" between the producing and interpreting mechanisms. Notice that the interpreting mechanism is physically located in a bee other than the one in which the dance production mechanism is located: it is nowhere implied that both the producer and interpreter must exist in the same token organism.

Third, the proper function of the interpreter mechanism is to produce the appropriate direction of flight in the watching bee. The adaptor for the bee dance is the location of the nectar. The dance adapts the interpreter devices to that location. This is what enables the interpreter device to perform its own proper function by producing the appropriate direction of flight. Bee dances thus fulfill the third condition by adapting the interpreter mechanism to conditions under which the proper performance of that device can be performed.

Both parts of the fourth condition are satisfied:

> Intuitively it is clear that in some sense of "mapping," the bee dance that causes watching bees to find nectar … is one that maps in accordance with certain rules onto a real configuration involving nectar, sun, and hive. As such it is an indicative intentional icon. The bee dance also maps onto a configuration that it is supposed to produce, namely, bees being (later) in a certain relation to hive and sun – that is, where the nectar is. So the bee dance is also an imperative intentional icon (Millikan 1984, 99).

Many of the most primitive intentional icons are simultaneously indicative and imperative. In a later writing (1995) she introduced the term 'pushmi-pullyu' to describe these kinds of icons, and subsequently (2004) elaborated further on their nature. A full discussion depends on discussion of the mapping rules, so we'll postpone pushmi-pullyus until then.

This aspect of her theory of intentionality should be understood as follows. Millikan's theory of proper function is a theory of biological categories. It can be used to describe and categorize all sorts of physiological phenomena, as well as explain diseased and malfunctioning organs and other physiological states and processes. It is a naturalistic account of function and malfunction. As members of the group of biological categories, intentional icons also have proper functions, and the explanation of *what it is to be* an intentional icon is provided from within that conceptual framework. Interestingly, Millikan's account of icons does not explicitly say that a state must be capable of *being false* in order to be a state of an intentional icon. In fact, it does not say anything at all about what the intentional icon is about. That aspect of the theory – certainly an essential one – is described in terms of the mapping rules.

Finally I have two more pieces of Millikan terminology to introduce. What an intentional icon is an icon of is its *real value*. The mapping rule in accordance with which the icon is supposed to map to its real value is its *Fregean sense*. "[I]mperative sentences that are not obeyed do not have real values. Similarly, indicative intentional icons that do not map onto anything … do not have real values" (1984, 101). A real value is not to be identified with what philosophers of language would traditionally call a name's referent. Further, Fregean sense is not to be identified with sense or intension as traditionally understood. Fregean senses – the mapping rules between icon and real value – are the most fundamental element of intentionality, but they are not intensions.

4.2.2 Articulateness of Intentional Icons and the Relation of Sense to Reference

The traditional way of thinking about sense and reference, and the relationship among sentences and terms, is as follows. Terms, especially names, refer to or denote their object and hence correspond in some way to their object, and the correspondence of sentences to the world is a function of the correspondence of its parts to the world and its syntactic structure[62].

> The assumption is that any description of the mapping functions that correlate sentences with world affairs would, of course, begin by coordinating at least some words with some objects, that the kind of coordination involved is reference (or denoting), and that coordinations of sentences with world affairs must be built up out of these basic reference relations plus, perhaps, some added paraphernalia (1984, 102).

This suggests that the way that sentences map onto states of affairs is similar to the way that words map onto their referents. The problem with this is in dealing with false sentences, since there is nothing that the sentence maps onto. Frege's solution to this problem was to map names and subjects onto referents, which are things in the world, then claim that predicates express mathematical functions from objects either to other objects or to the true or the false. Sentences are like names, except they are very complex signs that all map either onto *the true* or onto *the false*. The problem with this is that it somehow loses the correspondence between sentence and world.

[62] This is not my analysis of previous literature as compared to Millikan's; I am still providing an exposition of her work. This particular subsection relies mostly on her (1984, 100-107).

Wittgenstein's early work tried something different, saying that words map onto the world, and the relations that words bear to each other is mirrored in the relations that the referents of these words bear to each other. He proposed essentially an isomorphism/picture theory for language.

Wittgenstein has it upside down, Millikan claims. Referential terms only denote in the context of a sentence, so being part of a sentence is one of the Normal conditions for a referential term to perform its proper function. Second, its function qua part of a sentence is to map onto its referent, however, "It maps onto its referent in the context of a sentence if and only if the sentence is *true*" (1984, 104). Because of this,

> the most basic or most direct kind of correspondence, then, is the correspondence between a true sentence and a world affair. When this correspondence occurs, we say that the sentence has a "real value" – namely, the affair it maps onto. A less direct, more mediated, kind of correspondence is the correspondence between a referential term *in the context of a true sentence* and its referent (1984, 104).

Another way of getting at this same point is this:

> My claim is that if we analyze the notion "reference" correctly, we see that it depends upon more fundamental kinds of relations, such as the relation of a true sentence to the world affair it maps, which relations cannot be analyzed in terms of reference for the same sort of reason that "pumping blood" cannot be analyzed in terms of "being designed to pump blood"… Somehow,

we must begin by correlating *sentences* with world affairs, correspondence of words with things coming after (1984, 106-107).

Fregean sense is the mapping rule in accordance with which an icon is supposed to map onto some complete state of affairs in the world. That state of affairs is the real value of the sentence or intentional icon. Reference, by contrast, is a derivative notion, that only applies *relative* to an entire sentence. It is only within the context of a sentence that, say, 'Aristotle' refers to Aristotle. 'Aristotle' does *not* refer to Aristotle merely as a name all by itself in virtue of, say, a dubbing ceremony and the causal links that pass down through the generations[63]. So in the sentence, "Aristotle was a philosopher", the complete state of affairs of Aristotle's instantiating the property of being a philosopher is the real value of the sentence, and Aristotle is the referent of 'Aristotle' only derivatively and within the context of the sentence.

This is, so far as I can tell, a key, fundamental difference between Millikan and most others. In a later work (1990) she calls this the *articulateness* of intentional icons. The idea is that an icon is about a complete state of affairs, or, the instantiation of a property by an object, rather than just an object sans property or property sans instantiating object[64]. By contrast, while Dretske makes a similar distinction between *topic* (that which the representation is about) and *comment* (what the representation "says"

[63] See (Kripke 1980).

[64] "A third contrast [between Millikan, Dretske, and Fodor]… is the special emphasis that Millikan alone places upon the *articulateness* of all complete representations. Complete representations represent complete states of affairs…A representation that represented something simpler than a state of affairs, one that represented, say, only an object or a property or a *type* of state of affairs (compare a propositional function) would make no *claim*, hence would fail to be true or false, to represent anything either correctly or incorrectly…" (Millikan 1990, reprinted in her 1993, 131). Notice that here she is using the word 'representation' in the standard way, and not in the special way that she has reserved for it in her 1984. In the context of our discussion in the text we should be thinking of the 'representation' in the above quote as 'intentional icon'.

about the topic), he also claims that the topic is *not* part of the content of the representation[65]. Millikan has made a fundamental and immensely important advance here, and this is one of the basic ideas upon which I will build my theory of representation. I'll return to this line of thought in chapter 5.

4.2.3 The Mapping Rules

If conditions 1-4 determine when some biological device is an intentional icon, then what determines what the icon is an icon *of*?

Millikan provides two different answers to this question in her (1984), but in subsequent works (I will focus on her 2004) clarifies the relationship among the different theories of content-determining mapping rules. I begin with the original proposal.

4.2.3.1 *Language, Thought, and Other Biological Categories*

One of the oldest theories of mental representation says that representation is a picturing relation, where the vehicle of representation bears structural similarities to, or shares properties with, that which it represents. The guiding idea here is that there is a kind of "picturing" or "mirroring" between representation and represented in virtue of which the representation relation obtains. I'll call this the *picture theory* of representation. An updated version of this sort of theory posits not a structural similarity between a token vehicle of representation and a token represented state of affairs,

[65] In his (1988, 70), Dretske says "there are always two questions that one can ask about representational contents. One can ask, first, about its reference – the object, person, or condition the representation is a representation *of*. Second, one can ask about the way what is represented is represented". In a later writing (1995, 26) Dretske notes that a representation has the function of indicating "the F [or, property] of those objects which stand in C to it [where C is a reference-determining context], but it does not have the job of indicating – does not therefore represent – which objects – or even whether there is an object – that stands in C to it". Hence, the topic or referent, that on which the representation comments, is not part of the representational content.

but rather, a structural similarity among a *system* of representations and a *system* of states of affairs. I'll call this the *system-isomorphism* approach.

On this latter theory, the guiding motivation is the same: the preservation of internal structural relations between representation and represented is of the essence of representation. However, the structural similarity obtains between a set of items and relations on that set, and another set of items and relations on it[66]. The first theory of the mapping relations that Millikan proposes, and clearly the one that she intends as her major solution, is a version of the system-isomorphism approach.

When an indicative intentional icon has a real value, the icon is related to its real value in the following way. First, the real value is a Normal condition for proper performance of the icon's direct proper functions. Second, there is a system of icons in which each icon can be altered according to a rule of transformation. The represented states of affairs also admit of transformations, and the two systems, icons and real values, bear structural similarities to each other: the transformations are structure preserving.

> The governing idea here is that, in the first instance at least, it is *transformations* of the icon that correspond to *transformations* of the real value – *operations* upon the icon that correspond to *operations* upon the real value – not elements of the icon that correspond to elements of the real value (Millikan 1984, 107).

[66] I argue that a system-isomorphism approach is a necessary element of the theory of representation. We will delve much further into this concept in chapter 5.

The theory of proper functions plays a minimal role in determining the content of intentional icons. The direct proper function of the icon is simply *to represent to* the producer that such and such is the case. But what that comes down to is that the icon has the function of bearing some particular correspondence relation to its real value, where that correspondence relation is determined by the function of the consumer, as follows. The function of the producer is to produce icons that stand in some correspondence relation to the world, and to act in concert with the consumer. The function of the consumer could be anything at all, but whatever its function is, the explanation of that function is going to rely on certain conditions, one of which is a particular kind of correspondence relation between the icon the consumer uses and the world. Whatever relation that is, the producer has the function of producing icons that stand in that correspondence relation to the world. Finally, the icon itself has the function of simply bearing the appropriate correspondence relation. An accurate icon is one which successfully performs its function of representing to the consumer that so and so is the case, and it does this when it accords with some condition in the environment according to the correspondence rule determined by the proper function of the consumer. Thus, all it is for an icon to perform its proper function is for that icon to bear some particular correspondence relation to the world; it is the correspondence relation itself that does most of the theoretical work here. What determines the content of the intentional icon is the systematic, one-to-one correspondence of transformations of the icon to transformations of its real value. This is system-isomorphism theory.

Millikan however does not leave it at that, but provides a second, and very different, explication of the mapping rules that determine content. At the end of chapter 8 she provides an appendix titled "Why Beliefs are Intentional Icons", where she attempts to show (based on some further assumptions) that beliefs satisfy conditions 1-4. While the earlier description of the mapping rules involved structure-preserving transformations, now she speaks in terms of *information*:

> These other [inner consistency-testing] programs are good programs and should pass muster only if they are helping to produce [inner] sentences [i.e. beliefs] that map onto the world in accordance with some definite rules *for a reason* – a reason mentioning conditions under which the programs often operate and mentioning laws of nature which, under these conditions, *connect* these sentences with what they map. (That is, these sentences must bear *information* concerning what they map onto roughly in the sense that Dretske defines in *Knowledge and the Flow of Information*.) These programs are then associated with definite mapping rules (1984, 146).

Thus, to satisfy condition 4, Millikan appeals to Dretske's information, not transformation rules.

In her (1984), from which I am mostly drawing my exposition now, there is an apparent wavering between information and isomorphism. In later writings Millikan provides a more in-depth analysis of Dretske's information. In her (2001), "What has natural information to do with intentional representation?", she argues that Dretske vacillates, both within his (1981) and between his (1981) and his (1988), between different senses of 'information', and that the canonical formulation from his (1981) does not do the work needed by a theory of representation. In her (2004), to which we now turn, Millikan provides her own concept of information, different from Dretske's, and clarifies the relationship of information (in her sense) to isomorphism in her theory of intentional icons.

4.2.3.2 *Varieties of Meaning*

Briefly, Millikan rejects Dretske's information because of his requirement of a conditional probability of one, thus appealing to natural necessity rather than mere statistical correlation between sign and signified. First, the information (non-technical use of 'information' here) that we and every other organism need in order to survive, cannot possibly be gained by relying solely on such strict necessities. Organisms must make use of the statistical correlations in their environment. Second, creatures must be able to *learn* from the information-bearing signals they encounter, and Dretske's concept does not seem to account for this[67]. Further, individuals do not enter into laws, and so it is impossible in principle for an information theory based on natural necessity to provide an account of how organisms represent individuals (2004, 35).

Having rejected Dretske's information, Millikan proposes a "softer" information, which she calls *local natural information*. The corresponding signs that carry local information are *local natural signs*. The basic natural signs needed for a theory of intentionality can't be isolated or one-time-only signs. They must recur, and with the same signification or meaning. Hence, she dubs them *recurrent natural signs*. Since Millikan rejected Dretskean information on the grounds that it ignores statistical correlations less than one, and is not obviously connected to an organism's ability to learn, she builds these attributes into her theory of local information. At root, she says, the notion of a natural sign is an epistemic notion:

> The central thing common to all of these examples [described in Millikan's text] of natural signs, I suggest, is that in each case it is possible for a true belief to be reached about one thing from

[67] "Nearly all of the kinds of information needed by us, and by all other organisms as well, for securing what we need in an inclement world, is information that cannot possibly be acquired without leaning on certain merely statistical frequencies" (Millikan 2004, 32-33) … "The mere fact that a signal carries certain natural information seems not to bear on whether a creature could learn anything from encountering that signal" (Millikan 2004, 33).

> knowledge of the other… A natural sign of a thing is something else from which you can learn of that thing by tracking in thought a connection that exists in nature. The notion of a natural sign is at root an epistemic notion (Millikan 2004, 37).

That which makes it possible to learn about *B* from *A* is that there exist correlations in nature between the two, and it is not required that these are necessary correlations. But correlations, as Dretske noted as well, must be defined relative to a reference class, and these reference classes cannot be arbitrary. Millikan writes that, relativized to any arbitrary reference class, the concept of a natural sign "does no work" (2004, 38). However the problem is deeper. Without a principled understanding of the reference classes to which the statistical correlations are relative, the very probabilities are themselves non-objective in a way that threatens the entire reductive naturalist project, as argued in chapter two (section 2.4.2). This is exactly the same problem that we encountered with respect to Dretske's work. So what to do?

The work that we want our concept of recurrent natural signs to do, according to Millikan, is to explain why organisms can *use* a natural sign as an indicator of something else; "we want it to be possible for an animal to come to *learn* of *B*s from encounters with those *A*s, where what counts as learning is acquiring true beliefs nonaccidentally" (2004, 39). What we need then "is some way to delineate *relevant natural* classes" (2004, 39, emphases in the original). Millikan's canonical formulation of her solution to this is as follows.

> A natural reference class for a sign – the natural domain within which certain *A*s are "locally recurrent signs" of certain *B*s – is a domain within which the correlation of *A*s with *B*s extends

from one part of the domain to other parts for a reason, and it must be a domain that it is possible for an organism to track (2004, 40).

In order to be able to use recurrent natural signs in the environment to learn about other parts of the environment, the organism must "stay" within the bounds of the relevant domain, since the statistical correlations exist only relative to that domain. That does not imply that the organism must be able to identify the boundaries of the domain as such. It is entirely likely, for example, that the rabbit will never leave the natural domain within which various locally recurrent signs of the fox bear the necessary correlations (2004, 42).

An important question that in part led Dretske to insist on conditional probabilities of one, is that of how much correlation is enough for a sign to count as a sign, or for local natural information to exist. With her teleological framework, Millikan has a ready answer: "A strong enough correlation to count in determining a local sign to be such is that one is strong enough to have actually influenced sign use, either through genetic selection or through learning" (2004, 44).

Finally, Millikan notes that it is not necessary that there be any causal connection between a sign and that which it signifies in order for that thing to be a sign (2004, 44). Many items, for example, constitute natural signs merely in virtue of conservation laws: Things tend to stay as they are. Thus, the relation between sign and signified may be reiterated not because of a causal connection but simply because the signs and the signifieds tend to persist.

Locally recurrent natural signs have another, very important property, which they share with sentences. Sentences exhibit productivity: They can generate novel sentences through the combination of parts. Natural signs also exhibit productivity, because they are structured world affairs which have

structured world affairs as their contents. This is the idea of the articulateness of signs introduced earlier. Natural signs are not analogous to predicates or names in the sense that they signify some property abstracted from the individual that instantiates it, or vice versa. Rather, natural signs exhibit their meaning architecturally, in virtue of their abstract shape or structure:

> Natural signs are structured world affairs and the things of which they are signs are also structured world affairs, analogous to the correlates of complete sentences rather than open sentences or sentence parts … the meaning of the signs is determined as a function of values of significant values or determinables exhibited by the sign (2004, 47-48).

The compositionality of sentences, Millikan writes, is really just a special case of architecturally determined meaning more generally. Natural signs exhibit their meaning architecturally, and this is what allows them to have novel contents and thus productivity. As described above, Millikan argues that the signification of parts of signs to parts of what they represent is derived from the signification of the whole to the whole, not the other way around.

As should be apparent, this is another description of the transformations and operations on icons as bearing structure-preserving relations to significant transformations and operations on real values discussed earlier. It is a description of a system-isomorphism theory. The *semantic mapping functions* (here 'function' is used in its mathematical not teleological sense) that define natural signs are relations from signs to signifieds. "Semantic mapping functions define isomorphisms between the set of possible signs in a certain domain and the set of their possible signfieds. Natural signs are abstract 'pictures' of what they represent" (2004, 49-50).

Given this new, "softer" version of natural information, we're now in a position to articulate Millikan's answer to the second element of our representation question: given that something is an intentional icon, what determines what it represents?

There are three basic kinds of intentional signs: descriptive (analogous to indicative icons), directive (analogous to imperative icons), and pushmi-pullyus, which are a combination of both, but evolutionarily more basic than either. In all three cases, the content of the intentional sign is determined by a definite mapping function, which itself is determined by a locally recurrent natural sign. Being a locally recurrent natural sign is determined, as described above, by a combination of both system-isomorphism and statistical correlations, where the correlations are defined relative to a natural reference class. This clarifies the relationship between information and isomorphism that was not apparent in her (1984). Namely, system-isomorphism and something analogous to covariation (causal or otherwise) are both built in to the theory of local information, and this is what determines the mapping rules which determine content.

We should be careful to note here the difference between the Normal conditions for performance of a teleofunction, and the teleofunction itself. When intentional signs are produced by normal mechanisms, those signs *are* natural signs, even though it is not the function of the producer to *produce* natural signs:

> [I]f we focus clearly on the function of the sign-producers, carefully distinguishing their function from the normal mechanisms by which they fulfill this function, we see that their function is only to produce for their consumers what the consumers need. Their function is only to produce representations that correspond to world affairs by a certain mapping function. Their purpose

> or function is not to achieve this in any particular way … In every case [of different kinds of representations], when the production and use of these representations proceeds by normal mechanisms, they are local natural signs (2004, 77-79).

Further, on her theory of local natural signs something can be a sign of future events[68], hence, there are natural signs for both descriptive and directive signs (2004, 79). For descriptive signs, "the content of the descriptive sign is not determined by the tasks its consumer performs. It is determined by what the sign needs to correspond to if the consumer is to perform its tasks in a normal way" (2004, 79-80). For directive signs, "that the sign maps in the right way will be a result of the consumer's activity" (2004, 80), and for pushmi-pullyus, there is a combination of both. There are thus two ways that the producer and consumer mechanisms can work together.

> First it might be that the producer is the one primarily responsible for making the sign correspond to the world [this is a descriptive sign] … Second, the consumer may be the one primarily responsible for making the world correspond to the sign [this is a directive sign] (2004, 80).

[68] For example, the Canada geese traveling overhead are a natural sign of the approach of winter. There is a statistical correlation between the geese traveling and the future arrival of winter, within a certain reference class. Notice that the geese do not cause the arrival of winter. See (Millikan 2004), chapter 3.

4.3 Some Extant Critiques, and Why They Don't Work

There are several criticisms of Millikan's work in the literature which are clearly off the mark, and this can be appreciated once we have an adequate grasp of her theory. I will discuss some of the more prominent.

The first criticism, articulated by Cummins (1989, chapter 7) among others, is that Millikan's theory is not even a candidate, since the concept of representation with which she is working is diachronic, not synchronic. That is, the nature of representational content must be based entirely in the present moment, not in the past. Since Millikan's theory of proper functions is based on historical factors, and the theory of proper functions is used in part to explain content, it follows that Millikan's teleosemantics does not work.

There are three replies to this objection. First, the claim that representation must be synchronic is unsupported. Perhaps it is based in intuition, but I must admit that I have no (*pre-theoretic*) intuitions on whether the content-determining factors for mental representation must occur simultaneous with the representation, or not. Further, I suspect that anyone who has this "intuition" is espousing the implications of one's favored theory.

Second, Millikan (1989) asks, regarding purposes, why not look at history? This is analogous to asking the same question regarding representational content, since conscious purposes have representational content (for example, "I intend that *p*"). The answer is instructive: Meaning rationalism, the doctrine that we know what our purposes are, and what our thoughts are about, says that our purposes are immediately available in consciousness. They are "given". "Hence [if meaning

rationalism is true] what one's explicit conscious intentions are could not possibly depend on facts about one's history" (Millikan 1989, reprinted in her 1993, 28).

But, Millikan argues, meaning rationalism is false. She cites Wittgenstein's (1953) *Philosophical investigations*, Sellars' entire corpus, Putnam (1975), and Burge (1979), as each contributing to the argument against meaning rationalism. For example, what one intends, for example when intending to follow a rule (Wittgenstein) is not given to consciousness, that *nothing* is epistemically given to consciousness (Sellars), and that what one means is not determined entirely by what is "in one's head", but rather on external relations from the head to the world (Putnam/Burge).

> If these philosophers are right and meaning something or intending something or purposing something depends on relations *not* packed inside an epistemic consciousness, then why are historical relations not as good candidates for this position as any other relations? (Millikan 1989, 29 in the reprint).

If meaning rationalism is false, we need not accept the objection that thought contents must be determined synchronically.

Third, and decisively, the objection misunderstands Millikan's theory. Teleofunction determines that a thing is a representation, whereas the mapping rules determine the content of any particular representation. Her theory of the content-determining mapping rules is grounded in system-isomorphism and statistical correlations, not teleofunction.

Fodor (especially 1990, chapter 3) makes several misguided objections. First, he argues that Millikan has made a sort of a "distributive fallacy" (1990, 65-66). Just because the mechanisms of belief production and use have functions, this does not imply that any individual beliefs have functions.

> But the assumption that the mechanisms that make/use cognitive states have functions does not entail that cognitive states themselves do… It's a sort of distributive fallacy to argue that, if having beliefs is functional, then there must be something that is the distinguishing function of each belief (1990, 66).

Fodor also apparently thinks that Millikan seeks to determine the content of representational states in terms of the function of the representations themselves. "[That mechanisms of belief production have functions does not imply that beliefs do, and] it a fortiori does not imply that beliefs … can be *individuated by reference to their functions*" (1990, 65).

This would be a distributive fallacy, and trying to determine content based on it would be a problem, but this is not Millikan's theory. Rather, she focuses equally on the functions of all three items (producers, consumers, and representations/icons). The function of the *representation* is only to bear a particular correspondence relation to the world, where which particular relation it is supposed to bear is determined by the function of the consumer, which could be anything at all. The function of the producer is to produce states which bear the appropriate correspondence relation. What Fodor means by 'the individuation' of icons is the determining of content. As discussed above, on both the earlier and later versions of Millikan's work, content is not determined by the function of the individual icon but by the mapping rules or semantic mapping function. Millikan does not claim that there must be a

distinguishing function for each belief, nor does she claim that beliefs/icons/representations should be individuated by their functions.

Fodor continues, "[Millikan's] sort of account leaves it mysterious why the identification of content with function works *only* for intentional states; why beliefs have intentional content in virtue of their functions but hearts, eyes, and kidneys don't" (1990, 66). This is a misreading of Millikan on two levels. First, as mentioned above, content is not identified with teleofunction and beliefs do not have content in virtue of their teleofunction on Millikan's theory. Recall from Millikan (1984, 100): "[The conditions articulated in terms of proper functions] tell us when something is an intentional icon. They do not, however, tell us what a given intentional icon is an intentional icon *of*".

Second, Millikan explicitly answers Fodor's question (in her 1984, chapter 6). She notes that, in the tradition of Brentano, intentionality is typically understood to be a relation one of whose relata need not exist. On this minimal account then, anything that has a function has intentionality, since it is of the essence of a teleofunction that it need not be performed. However, "in a narrower and more usual sense of 'intentionality' not everything that is meant to perform a certain function displays intentionality. For example, the heart does not display intentionality…" (1984, 95). Millikan then proceeds to develop a theory according to which biological devices are intentional icons on this narrower and more usual sense of 'intentionality', and provides the four conditions discussed in section 4.2.1. Hearts fail to satisfy conditions 2-4.

Fodor continues. Desires do not, even under Normal circumstances, cause their own satisfaction.

> It is simply intrinsic to the logic of wants that they can be causally isolated from the states of affairs whose occurrence would satisfy them … That it is possible to have wants that are arbitrarily causally inert with respect to their own satisfaction is, indeed, one of the respects in which wants are intentional (Fodor 1990, 68-69).

This is correct, but it is also exactly what Millikan says, and thus it does not constitute an objection to her theory. Recall condition 4a on being an icon: "In the case of imperative [icons], it is a proper function of the interpreter device, as adapted by the [icon], to produce conditions onto which the [icon] will map in accordance with a specified mapping function" (Millikan 1984, 97). Recall also that it is of the essence of something that has a proper function that it need not ever perform the function: "Notice that it is not necessary that a device actually serve any direct proper functions of it" (Millikan 1984, 29). Thus, the proper function of the *consumer* of imperative icons (not the icon itself) is to produce conditions onto which the icon will map in accordance with the specified mapping function. But devices with proper functions need not ever actually perform those functions. Hence on Millikan's theory it is intrinsic to the logic of wants that neither they nor their consumer mechanisms need ever *actually* cause their own fulfillment. As Fodor notes, this is precisely the connection between teleofunction and intentionality, and it is precisely what Millikan writes: "The general solution [to the problem of there being a thing which can apparently stand in a relation to something that need not exist] … is to see that intentionality is at root properness or Normalness" (1984, 95). In fact the most basic and distinctive idea behind Millikan's theory of proper functions is that current causal dispositions are *divorced* from teleofunction: "My claim will be that it is the 'proper function' of a thing that puts it in a biological category, and this has to do *not with its powers but with its history*" (Millikan 1984, 17, my emphasis).

Finally, Fodor has a general criticism of teleological theories which says that ascriptions of function are indeterminate in principle. Because of this, teleofunction cannot be used to determine content, either individually or in combination with an informational theory as a solution to the disjunction problem. This objection does not apply because Millikan's theory does not use teleofunction to determine content. Rather, on Millikan's theory it is the mapping rules that determine content. In her earlier work the theory of the mapping rules was, for the most part, a system-isomorphism theory. In her later work it is based on her concept of local natural information, which combines system-isomorphism with a kind of nomically grounded covariation. In both cases, the teleofunction of a particular representation does not determine its content.

Millikan's theory is complicated and her writing is dense, both of which unfortunately lead to the tendency to oversimplify or outright mischaracterize her work. As a result, many of the published objections are badly off the mark. I have focused on Fodor because his work exemplifies some of the common misreadings of Millikan.

4.4 Critique of Millikan

Millikan's theory of teleofunction is important, not least because it provides a conceptual framework for unifying physiology and psychology, which is exactly what we want if we take the naturalist viewpoint seriously. With respect to the first representation question – what makes a thing a representation – something like Millikan's theory has got to be correct. I do not propose to challenge that, nor to endorse the details. What I do challenge is her answer to the second question, which is what most people are focused on as well: Given that something is a representation, in virtue of what

does it have the content that it does? This is the question of content or aboutness, and it is to this question that Millikan provides her theory of mapping rules as an answer.

In 4.4 I proceed as follows. First I address the early system-isomorphism theory and discuss how it might fare against some traditional objections to resemblance theories. Then I argue that her theory of local information falls prey to the same problems as Dretske's information. Finally I address the role of normativity in a theory of mental representation. At the heart of any version of teleosemantics is the claim that truth is a normative notion, and so a naturalization of content depends on a naturalization of normativity. However, there's normativity and then there's *normativity*. In the final subsection I provide a more careful analysis of the relationship of normativity to mental representation, and argue that it is not quite as simple as teleosemanticists like Millikan and Dretske assume. Once we get clear on the relationship between normativity and representation, we'll see that it actually doesn't do the work that Millikan and Dretske want it to do.

4.4.1 Isomorphism

The basic idea of Millikan's system-isomorphism approach to the mapping rules is that there is a system of icons whose members can be transformed according to rules. Further, there is a system of real values that are also transformable. Importantly, what makes any particular icon about any particular real value is that both of the *systems* bear structure-preserving relations to each other. They are isomorphic[69]. The icons map one-to-one to real values, in a way that preserves transformations[70].

[69] In the next chapter I will propose my solution to the representation problem, which will crucially involve isomorphism. I will provide a more careful discussion of the concept of isomorphism in that chapter. For our purposes here a general overview will suffice.

[70] "[R]epresented conditions are conditions that vary, depending on the *form* of the representation, in accordance with specifiable correspondence rules that give the semantics for the relevant *system* of representation. More precisely, representations always admit of significant transformations (in the mathematical sense) which accord

Note that this is *not* a version of the antiquated picture theory of Aristotle, which says that the icon itself and that which it represents bear resemblance or structure-preserving relations to each other, the way for example a photograph of President Obama resembles President Obama.

Bee dances, for example, are icons whose intentional content is determined by their membership in a system that bears structural similarities to the system of real values (locations) that they represent. Variations in the tempo of the dance, as well as of the angle of its long axis, are operations upon the icon. Variations in the distance and direction of the location of nectar relative to the hive and sun constitute significant operations on the real value. Thus, operations on any icon results in a different icon, literally, because of their structure. Further, operations on any world affair results in a different world affair, or, a different location. Finally, a (mathematical) function can be described which maps, one-to-one, icons onto real values and which preserves the internal relations of the two systems. For example, the *faster tempo than* relation preserves the *further distance from* relation. Let's call the mathematical mapping function a *semantic mapping function* (as Millikan does in her 2004 and elsewhere). With respect to a particular semantic mapping function, each icon in the icon-set maps to and thus represents some particular location in the location-set.

There are some traditional objections to structural preservation theories of all kinds. Two can be quickly dismissed as not applicable to Millikan's. The first is that resemblance cannot be used as a reductive ground for a naturalist theory of mental representation because resemblance itself is an intentional notion. Seemingly everything shares some property with everything else; what matters is choosing the salient properties for the purpose of ascribing resemblance. But salience depends on an observer's judgments of salience, and hence, depends on an intentional agent. System-isomorphism is

---

with transformations of their representeds, thus displaying significant articulation into variant and invariant aspects" (Millikan 1989, reprinted in Stich and Warfield 1994, 248).

not susceptible to this objection: it is not the case that "everything is isomorphic to everything else". Proving that isomorphism exists is a difficult and non-trivial task, and whether or not isomorphism connects two relational systems is an entirely objective matter that does not depend in any way on an observer. There is a complication here: we must not confuse, on the one hand, whether or not there is an isomorphism, with *defining* relational systems in such a way so that they are isomorphic. The fact that arbitrary sets with arbitrary relations on them can be defined in such a way that they are in fact isomorphic to some relational system of interest, does not imply that isomorphism is a non-objective or intentional relation among independently specified systems[71].

Another traditional objection is that isomorphism is symmetrical while representation is not. While my thoughts might represent stars, stars don't represent my thoughts. This objection can be applied to both the picture theory version as well as the system-isomorphism version of structural preservation theories. If you can define a mapping function that preserves internal relations from set *A* to set *B* then its inverse will also preserve internal relations. Thus, icon-1 represents location-1, but further, location-1 represents icon-1. But that can't be right.

Millikan however has a ready reply to this. Her theory is not a *pure* isomorphism theory. The first requirement, recall, is that the real value is a Normal condition for proper performance of the icon's direct proper functions (see 4.2.3.1). I argued that this requirement doesn't really do much work, since the teleofunction of the icon is only to stand in some correspondence relation, and that correspondence relation is defined by the teleofunctions of the consumer device and by the transformation rules that define the system of icons. However, in this context it does do some work: the location represented by the icon is not a member of a biological category and so does not have any proper functions. Because of this, there aren't any Normal conditions for proper performance of its functions and the icon thus

[71] I have much more to say on this crucial point in chapters 5 and 6.

cannot be a Normal condition for the location, whereas the location is a Normal condition for the proper performance of the function of the icon[72]. So while the symmetry objection applies to pure isomorphism theory it does not apply to Millikan's version that appends the Normal condition requirement.

Another related objection is the non-uniqueness objection: there are infinitely many relational systems that are isomorphic to any given relational system (to the system of bee dances, for example). Because of this, there are infinitely many mapping functions from the system of bee dances to other systems. Why then, do we not say that icon-1 represents an infinite number of other things as well as location-1, to all of which icon-1 maps? Another way of getting at this objection is that sometimes, even though there exists an isomorphism between two relational systems with respect to function *f*, there is also a different function, *g*, that preserves internal relations between the same two relational systems. For example, icon-1 maps to location-1 with respect to *f* and icon-1 maps to location-2 with respect to *g*, and both *f* and *g* are isomorphism-defining functions between the same two relational systems.

This is a difficult problem for all structural preservation theories. I argue that, while we can fashion a reply using Millikan's machinery alone, it turns out that we actually don't have any way of explaining representing falsely. This argument is going to take some set-up; I'll begin that work now, but we'll have to wait for 4.3.3.3 for the resolution of this.

The basic trouble is that there are too many (mathematical) functions that define isomorphisms. For ease of discussion, let's name some functions. Consider the system of bee dance icons mapping onto the real value system of locations. The function from icons to locations that determines intentional

[72] Just to clarify: the proper function of the icon is to stand in an appropriate correspondence relation to its real value. But it can't stand in that relation to something that does not exist; thus, a Normal condition for performance of the icon's function of bearing some particular relation to the real value is that the real value exist. This is why it is a Normal condition for the icon.

content is the semantic mapping function: This is identical to the mathematical function that defines the isomorphism between the two relational systems. Let's call the function that we would intuitively want to say is the content-determining one *f*. With respect to *f*, icon-1 maps to location-1. Let's call some arbitrary isomorphism-determining function, one that we would intuitively say does not determine intentional content, *g*. With respect to *g* icon-1 has location-2 as its image (and thus, potentially, its content). So the question is: Why does icon-1 have location-1 and not location-2 as its content if structural preservation is all that matters for determining content?

To answer this, we can use the rest of the machinery Millikan has provided to specify which mapping function is the *semantic* mapping function. As a preliminary, it is important to note that Millikan has not (to my knowledge) discussed this particular problem and so has not provided the solution that I am about to discuss. Rather, in her early work she clearly intends system-isomorphism to be the major aspect of her theory of the content-determining mapping rules, but she also begins to flirt with information. In her later works she clarifies the role of information in her theory of the mapping rules. Thus, it should be understood that my discussion here isn't really of Millikan's theory, and is probably not something that she would endorse. Rather I am exploring what could be done, using the conceptual tools that Millikan has provided. This is an important task because the notion of a *semantic mapping function* (which I also call a *representation function*) is a crucial element of my own theory, which I develop based on the following discussion.

First, there is an important similarity between the above question and another question. Why does 'dog' mean *dog* and not *cat*? The meanings of our linguistic terms are, to an important extent, arbitrary with respect to the actual *form* of the symbols. There's nothing intrinsic to the form of 'dog' that maps it to *dog* and not *cat*. The problem with isomorphism-determining mapping functions is the same: there's nothing intrinsic to one mapping function (say, *f*) that makes *f* determine content and not

some other function, *g*. From a certain perspective, it is just as arbitrary that *f* determines content as it is that the mapping function known as the English language determines content for English speakers (for example, so that 'dog' means *dog*).

Millikan's answer to the linguistic question is to appeal to stabilizing and standardizing teleofunctions. The idea is this: if hearers did not respond to uttered sounds in lawlike ways, then speakers would stop speaking; similarly, if speakers' sounds did not correlate in lawlike ways to things of interest to hearers, then hearers would stop listening. There is a crossover point at which standard uses of uttered sounds (such as using 'dog' to mean *dog* and not *cat*) and hearers' responses to them contribute to both the speakers' and the hearers' ends.

> The stabilizing and standardizing proper function … of a language device is that hypothesized function … that tends at the same time to keep speakers using the device in standard ways and to keep hearers responding to it in standard ways, thus stabilizing its function … (Millikan 1984, 31-32).

How did it originally come about that 'dog' means *dog* and not *cat*? While the answer to that may be forever lost to history, what matters is that language users *did* in fact utter that symbol token, correlated in a lawlike way with dogs, often enough that this use became standardized, and this

standardization of use stabilized the teleofunctions of the producers and interpreters of language devices in such a way that now, 'dog' means *dog*[73].

The same sort of answer can be given to our initial question. It is to a certain extent arbitrary that it is *f* and not *g* that is the *semantic* mapping function, rather than just an isomorphism-determining mapping function. However, whatever the specifics, enough bees did *in fact* behave in such ways that their dance behavior correlated in lawlike ways with locations, and those locations were of interest and adaptive value since they contained nectar. The correlation that did in fact occur between the dances and the locations is described by *f*, not *g*. Since that is the mapping function which defines the correlation between icon and nectar-containing location, and this correlation is part of the Normal explanation for the proper performance of the various relevant teleofunctions (for example, the teleofunction of the interpreter-mechanism in the bee that results in muscle contraction which ultimately results in the bee's arrival at the nectar-containing location), then *f* is the *semantic* mapping function. That *f* maps icon-1 to location-1, and that bees behaved in such a way that the producer and consumer mechanisms associated with the bee dance system of icons had their teleofunctions stabilized and standardized to *f*, is why *f*, and not *g*, is the mapping function that determines the content of bee dance icons. Thus, icon-1 means location-1 and not location-2, even though there does in fact exist an isomorphism-determining function (namely, *g*) from icons to locations, taking icon-1 to location-2.

[73] This is a vast oversimplification of Millikan's theory of language. On Millikan's theory, there are three aspects of semantic content for language tokens. *Dictionary sense*, somewhat analogous (but not entirely) to the sorts of entries you would find in a dictionary, is determined by stabilizing and standardizing functions. *Fregean sense*, which we have discussed above, is determined by the content-determining mapping rules presently under discussion. This is the most basic aspect of meaning; it is distinct from sense as traditionally understood. *Intension* is something like the rules that determine which symbol tokens can be replaced with others ('Morning Star' can replace 'Hesperus'). I'll not be discussing any of this in the dissertation. My only point in the text above is that Millikan does have an answer, the details of which certainly bear working out, to the related question about the apparent arbitrariness of uttered symbol tokens. We can co-opt that answer to the present question about the apparent arbitrariness of semantic mapping functions.

The role that teleology plays here is not to determine content, or at least, not directly. The claim is not that each individual icon has its own unique content-determining teleofunction. Rather, it is the mapping function *ƒ* that determines content. What teleology does is select among various isomorphism-defining mapping functions to determine which is the content-determining mapping function. We have a plausible response to the non-uniqueness objection to structural preservation theories, and this response is a key element in my own theory. A further problem arises, but we'll return to it in 4.3.3.3.

Millikan begins to mention the possibility of a causal or an informational connection between world and mind in her (1984). She explicitly mentions Dretske's information in (1984, 146), but later repudiates it (2001, 2004). This is a good thing, since Dretske's information doesn't work. Let's leave the isomorphism section then, and have a look at her theory of local information.

4.4.2 Local Information

Millikan's theory of local information incorporates elements from both system-isomorphism theory and something like Dretskean information. While she does reject Dretske's information with its insistence on natural necessity based in exceptionless laws, she also accepts that local information must have something to do with correlations between sign and signified. How much correlation must exist for the sign-signified relationship to hold is explained in terms of teleology: however much correlation was necessary to make an evolutionary difference for the biological structures that made use of those correlations, is how much is necessary. Further, intentional signs which carry local information are structured world affairs and the real values that they signify are also structured world affairs. The system of icons admit of transformations that result in new icons, the system of real values do so as

well, and the two systems bear the isomorphism relation to each other. Thus, both statistical correlations as well as system-isomorphism are built into Millikan's theory of local information.

We've previously discussed the isomorphism aspect of her theory and will return to it. Presently however we'll take a look at the statistical correlation in local information. The motivation here is that local information has to be something that an animal can learn from. Reliable statistical correlations between things in the world and recurring natural signs are optimal candidates for this. However, Millikan correctly notes that statistical correlations only exist relative to a reference class or domain, and further, that domain cannot be arbitrary or the concept of statistical correlations won't do the work for which correlations are needed by a theory of intentionality (2004, 37-40). Thus, we need some principled way of delineating natural reference classes. Millikan's solution is this:

> A natural reference class for a sign – the natural domain within which certain *A*s are "locally recurrent signs" of certain *B*s – is a domain within which the correlation of *A*s with *B*s extends from one part of the domain to other parts for a reason, and it must be a domain that it is possible for an organism to track (2004, 40).

This is the canonical formulation of Millikan's view of natural reference classes, and it has two parts: the correlation must extend from one part of the domain to other parts for a reason, and it must be possible for the organism to track the domain. Let's take a look at the first part.

That the correlation in question extends from one part of the domain to another "for a reason" is not enough to delineate non-arbitrary natural reference classes, since the reason why a correlation

extends from one part of the domain to the other could simply be a consequence of the delineation of the reference class itself. For example, Millikan writes about tracks in the woods that may have been made by either quail or pheasant (2004, 38-39). She asks: Why can't we combine a discontinuous set of locations from woods in Massachusetts and Minnesota where there happen to not be any pheasants but there are quail, call that set Q-woods, and then say that there is a reliable statistical correlation between tracks in Q-woods and quail? If we did, then we could say that those tracks carry local information about quail (with respect to the reference class of Q-woods). Her reply is that the (likely correct) inference from "this is a track in Q-woods" to "there is a quail nearby" is not explained by citing the statistics of Q-woods (2004, 39). Rather, "what is needed is some way to delineate *relevant natural* classes" (2004, 39).

But this isn't right: the inference from "All or most *F*s are *G*" and "a is an *F*" to "a is a *G*" does in fact explain why a is a *G*[74], and hence, from above, would explain why we are justified in believing that there is a quail nearby. The problem to which Millikan is pointing is not that statistical correlations are not explanatory. Rather, the problem is that *what counts as an F* is relative to an arbitrary reference class. However, ruling out arbitrary reference classes is exactly the problem that we started out with. It simply reappears here in different form.

Millikan is right that there is something not explanatory about saying that tracks in Q-woods are natural signs of quail. However, this is consistent with the correlation between those tracks and quail extending from one part of Q-woods to another part for a reason. That correlation does extend throughout Q-woods, and for a reason; the reason is that those tracks are located in Q-woods and most tracks in Q-woods are made by quail.

[74] "Why did that bar expand when heated?", one might ask. A good explanation would be this: "That bar is made of metal, and all metal expands when heated."

The only other way to go here would be to appeal to exceptionless natural necessity. That way the reference class is the entire universe and is not arbitrary. However, Millikan rejects this on the grounds that there is no way that organisms could survive and proliferate without making use of the statistical correlations of less than one that exist in nature. So the first part, by itself, is not enough to rule out arbitrary reference classes.

The second aspect, that the animal must be able to track the domain, does not work either. Millikan writes, "Conventional signs have domains that must be tracked, just as the domains of local natural signs" (2004, 129; see also the quote including what I've termed the 'canonical formulation' above). But if the notion of local natural information is supposed to underwrite a reductive, naturalistic explanation of intentionality, then the claim that an animal or human must "track" the domain leads to a regress because it presupposes some notion of intentionality or representation. If one must track the boundaries of the domain then one must represent those boundaries. But if one must represent boundaries in order to represent anything within those boundaries, then one must represent the "outer" or 2nd-order boundaries, which themselves define the "inner" or 1st-order boundaries which define the original represented item. This leads to an obvious regress.

Millikan recognizes this, and states that one need not represent the boundaries of the relevant domain in order to "stay within the boundaries": "To interpret a locally recurrent natural sign successfully you must keep within its natural domain … [But] it may not be necessary to discriminate the boundaries of the sign domain in order to stay within them" (2004, 42). But we must ask: if the ability of the animal to track the boundaries of the domain does not define the boundaries, then what does? What does the real work here is the claim that the correlations extend from one part of the domain to another for a reason. As I've argued above, that isn't sufficient for ruling out arbitrary reference classes.

Millikan's local information is distinct from Dretske's informational content and from mutual information. Nonetheless, it shares an important aspect with them both: If local information is to be an objective commodity, we must use the frequency interpretation of probability to generate the statistical correlations. But, as I argued in 2.4.2, the frequency-based interpretation of probability depends on implicit relevancy judgments. In Millikan's case, those implicit judgments appear as judgments about natural reference classes. This is analogous to the stable background conditions that define frequency-based probability, and analogous to the stable, enduring conditions that define the difference between a signal and a channel. In all three cases, a cognitive agent must decide what is and what isn't relevant. In the discussion above, we need a way to rule out things like Q-woods as arbitrary and irrelevant. But Millikan's two suggestions do not do so.

In chapters 5-8 I make significant use of correlations in explaining representation. But it's important to recognize the difference between identifying representational content with statistical correlations, as Millikan and Dretske have done, and using statistical correlations as fallible evidence of something else, as I do. Second, while an animal certainly can make use of reliable covariations in the environment, and can learn from them, we have to also distinguish between the identification of representational content with statistical correlations, and an animal's use of, or ability to learn from, a correlation. The latter element does not presuppose anything to which a naturalist is not entitled, whereas the former does, as argued in 2.4.2.

In addition to the problems with the non-objectivity of local information, Millikan's theory has the following further problem. She explicitly argues that information is at root an epistemic notion, and it is something from which an animal can learn about something else: "A natural sign of a thing is something else from which you can learn of that thing by tracking in thought a connection that exists in nature. The notion of a natural sign is at root an epistemic notion" (2004, 37). But this is unsuitable as a

reductive base for a theory of representation. Learning involves (at minimum) the acquisition of new representations, so a theory of representation based on information cannot explain what information is in terms of learning or "tracking".

4.4.3 Normativity

At the root of any teleology-based theory of semantics is the claim that truth and falsity are *normative*: there is something wrong with false beliefs or representations. Because of this, in order to naturalize representation we must naturalize normativity. This demand is vague and must be examined more closely. In this section I distinguish different kinds of normativity and different kinds of error or misrepresentation, and argue that the component of Millikan's theory that relies on normativity only accounts for one kind of misrepresentation but not both.

4.4.3.1 The Basic Claim

One way to object to Millikan's theory is to ask whether false beliefs can be adaptive, or rather, if there might be biological systems whose function is to produce false representations. If so then the fundamental claim of teleosemantics, that false representations are defective, is blocked. Millikan attempts to deal with this question in her (2004, 86), but only asserts that "Falseness itself *could* not be the point" [my emphasis]. She writes in her (1984, 9): "Put roughly, the meaning of a sentence is its own special mapping functions – those in accordance with which it 'should' or 'is supposed to' map onto the world. (Sentences are supposed to be true, aren't they?)". Later she says, "[beliefs] display the characteristic mark that all things defined by proper-function categories display. It makes sense to speak of their being *defective* … Beliefs … are *essentially* things that can be true or false, correct or defective" (1984, 94).

The basic idea is that representations are *supposed to be true*, and there is something wrong – normatively wrong – with false representations. This is where teleology comes into play: normativity is then "naturalized" or reduced to teleology, which is in turn explained in terms of natural selection and adaptive value. However, let us have a closer look at the claim that there is something wrong with false beliefs.

Consider this simple argument. The alleviation of suffering is a good thing. False beliefs, especially religious beliefs about an afterlife around the time of bereavement, tend to alleviate emotional suffering. Therefore, sometimes false beliefs and hence false representations are good things. Notice the form of Millikan's thesis and how it relates to this argument. First, normativity is reduced to adaptiveness, and then Millikan's claim is that false representations are not adaptive, and hence, not good. This argument bypasses the reduction of normativity to adaptive value and argues directly for the claim that sometimes false representations are *good*.

I am well aware that something seems acutely out of place with my argument from religious beliefs. However, I emphasize that Millikan claims that (i) false beliefs are defective, that something has gone wrong, and that (ii) this is in a *normative* sense of wrong or defective. The normativity of truth and falseness is what motivates the teleology. But surely something *is* out of place: the argument from the alleviation of suffering through religious beliefs is a non sequitur. Where is the error? The error is in assuming that all normativity ought to be lumped together into the same category. It should not.

4.4.3.2 Distinguishing Kinds of Normativity

There are two separate issues here. First, there is the claim that normativity reduces to adaptive value. Second, there is the claim that normativity is unitary. Let's have a look at them both.

Assume both that normativity is unitary and that it reduces to adaptive value. Then we get the naturalistic fallacy in ethics: the way that things ought to be reduces to the way that things are (or rather, were, in virtue of a thing's adaptive value). But that's not right. At the very least, we should not adopt the naturalistic fallacy as a result of our theory of representation. So either normativity is not unitary or it does not reduce to adaptive value.

However, Millikan's reduction of purposing and physiological functioning to adaptive value is a viable, useful, and explanatorily powerful theory in its own right, so we ought to accept it. Thus, to avoid the naturalistic fallacy and to accept Millikan's reduction of physiological proper functioning to adaptive value we should reject the claim that normativity constitutes a unitary category.

Given that normativity is not a unitary category, how should we split it? One place to start is at the distinction between *moral* normativity and non-moral normativity. Claims involving the value of the alleviation of suffering are clearly in the former category. Claims involving the purpose of the heart, or conscious intelligent purposing, are not. Let's name the moral type of normativity *strong normativity*, and the non-moral kind *weak normativity*.

There is a further element which we'll need to either assign to strong or weak normativity, or to its own category: truth and falsity. Clearly the statement '2+2=5' does not fall into the strongly normative category. But on what grounds should we assign it to the weakly normative category? That is, why should truth and falsity fall into the same category as purposing, rule-following, or teleofunctions such as the physiological function of the heart to pump blood? We've already seen that normativity is not all the same kind of thing so we should not lump them together on that basis. But if not that, then what?

Clearly, it makes sense to say that biological organs can malfunction. There is something defective with a heart that cannot pump. But the same consideration does not apply to '2+2=5'. There is nothing obviously *defective* about the statement. It is false, certainly. But to claim that falseness = defectiveness is to beg the question. So, given that we have reason to believe that normativity is not a unitary category, and in the absence of any reason to lump truth and falseness together with physiological functions and rule-following, and in the presence of the claim that the category of defectiveness clearly applies to hearts but not so clearly to sentences or representations, I propose to split normativity once again. In addition to strong and weak normativity, let's call the kind of normativity associated with truth and falseness *super-weak normativity*. This tripartite distinction is going to do some work for us. So an additional argument for making this distinction is going to come from working out a theory that relies on it, and from the clarifications that it will allow.

There are some further elements of normativity, one of which is rationality. Being rational is something like, if you believe that *p* and you believe that *p* implies *q*, then you ought to believe that *q*. Another kind of normativity is involved in explanation. The difference between a good explanation and a bad explanation is to be found in normativity, where the good explanation approaches something like a normative ideal. I'm not sure where these should fit in the above scheme, however it doesn't matter. I don't intend this to be a serious analysis of normativity. All that I need for present purposes is to make it plausible that (i) normativity is not a unitary category and (ii) the kind of normativity that reduces to adaptive value is different than the kind of normativity involving truth.

Millikan's claim that normativity reduces to adaptive value applies solely to weak normativity, but not to strong or super-weak normativity. I'm not going to say anything about strong normativity in this dissertation. But we'll still need an explanation of truth and falseness, and hence, super-weak normativity. This leads us to the final section.

4.4.3.3 Distinguishing Failure to Represent from Representing Falsely

Millikan's theory of teleofunction unifies physiology with psychology. Representations, qua mental states, have representational content. But further, qua *biological* states, representations are members of a biological category defined in terms of their teleofunction. They are what they are in virtue of whatever it is that they did, historically, that provided a differential advantage to the organism that had representations (or that had "better" representations), than those that did not. Millikan's theory is a complicated one involving the producers and consumers of representations as well as the representations themselves, plus a theory of mapping rules. Recognizing that this is a misstatement of her theory, but for ease of exposition, let's just say that representations are supposed to represent; that is their teleofunction and that is what puts them into their biological category.

Hearts have a teleofunction as well, which is to pump blood. A biological device that is supposed to pump blood but fails is still a heart, although it is a defective heart. There is something weakly normatively wrong with a heart that does not pump blood. Similarly, a biological device that is supposed to represent, but fails to do so, is still a representation, although it is a defective representation. There is something weakly normatively wrong with a representation that fails to represent. But *failing to represent* is one thing; representing *falsely* is something else entirely.

A representation that fails to represent is supposed to have a content (not necessarily any particular content, notice); but it has none. A false representation, by contrast, does indeed have a content. What is (super-weakly normatively) wrong with a false representation is that what it says just isn't true. But it at least says *something*.

This is an important distinction that bears emphasis. The difference between the failure to represent and representing falsely is this: a false representation has some content (that is, it "says"

something), whereas the representation that fails to represent does not have content. It fails to say anything at all. By distinguishing strong from weak from super-weak normativity, as well as representing falsely from failing to represent, some of the principal claims of the teleosemanticists now lack support.

First, the claim that normativity underlies representation is unsupported. Normativity underlies representation to the same extent that normativity underlies cardiac output: all physiological (and hence psychological) states are biological states, and biological states are defined in terms of their teleofunction, which involves weak normativity. There is nothing special about representational states with respect to normativity. The only thing that might be considered normatively special about representational states is that they can be true or false, which involves super-weak normativity. But as we've seen, this is distinct from both strong and weak normativity. I'd just as soon not even call truth and falsity a kind of normativity at all, but since this seems like a fairly entrenched idea, and since it's really just a matter of terminology, I'm happy to concede that truth and falseness involves normativity – it involves *super-weak normativity*, but nothing else. Second, the claim that false representations involve something having gone wrong is also unsupported, so long as we understand that when an author writes 'wrong' she generally does not mean 'super-weakly normatively wrong', but rather something else. Once we've got a clear grasp of the distinctions here, we see that false representations don't involve weak normativity, and no one claims that they involve the strong, moral kind, so they are really only super-weakly normatively wrong. And that's just another way of saying 'false'.

Weak normativity reduces to adaptive value in something like the way that Millikan claims in her theory of teleofunction. The defectiveness associated with weak normativity is the failure to perform a thing's teleofunction. With respect to representations, this defectiveness (or "wrongness") is the failure to represent. This is one kind of error or misrepresentation. Super-weak normativity does not reduce to

adaptive value. Representing falsely is the sort of defectiveness that is associated with super-weak normativity, and it is not explained by teleology. This is another kind of error or misrepresentation. So, we are in an important sense back to square one. We still don't have a theory of content, either for true or false representations. But that is to come in the next chapter.

These distinctions show us what was wrong with the argument from religion above. It is a non sequitur because I argued that false representations can be good on the basis of their strong normative value. But what is at issue is not their strong normative value but their super-weak normative value. This also show us something else: questions about whether false representations have adaptive value, when considered within the context of whether there is something normatively wrong with false representations, are also non sequiturs. One of Millikan's fundamental claims is that there is something wrong with false representations. Other authors have challenged that claim by trying to provide cases where having false beliefs does in fact have adaptive value. But this is not relevant. The only kind of normativity that reduces to adaptive value is weak normativity; this is the kind of normativity associated with the conditions on what it is to be a representation, not the conditions that determine representational content. The normativity involved with falsity is in an altogether different category.

Finally, let us return to our discussion from 4.4.1. We had been discussing the problem of the non-uniqueness of isomorphisms, where we needed some way to discover which of the numerous (and potentially infinite) isomorphism-determining mapping functions between a system of icons and other relational systems was the content-determining function, and hence, which was the *semantic* mapping function. The proposal on offer was that teleology can be used to select among the various mapping functions, and specifically that this could be done by co-opting Millikan's notion of standardizing and

stabilizing functions[75]. I argued that this would work, but only up to a point. Now I can state where that terminal point lies. Using teleology to select among the various mapping functions in order to determine which is the semantic mapping function does solve the non-uniqueness problem. But it only allows for an explanation of the failure to represent, not representing falsely.

Let's take another look at the bee dance. The teleofunction of the system of icons is to be related in a certain way (according to isomorphism-determining mapping function *f*) to the system of real values constituted by locations. That gets determined by the standardizing and stabilizing function of the producers and consumers of the icons. But once you have that, there is no representing falsely. Say that the honeybee makes a left instead of a right in its dance, such that this icon, according to *f*, maps to (and hence represents) location-3, not location-1. There is no way to say that this is false: the dance maps, according to *f*, onto its real value, and the existence of the real value is a Normal condition for the proper performance of the icon's teleofunction of mapping onto the real value, according to *f*. What we can say, however, is that this icon has *failed to represent*, because it has failed to satisfy the conditions on being an icon that it is supposed to satisfy.

Condition 3 on being an icon is that Normally, the icon adapts the cooperating interpreter device to the world such that the interpreter device can perform its proper functions. Further, the explanation of how the interpreter device performs its function makes reference to the fact that the icon maps conditions in the world in accordance with a specific mapping function (this is condition 4b, on being an indicative icon). The specific mapping function here is *f*. We can say that the dance in which the bee mistakenly made a left rather than a right fails to satisfy condition 3, because it has failed to mediate between producer and consumer in the way that historically allowed the consumer to

[75] This is just a reminder that this is, as far as I can discern, not Millikan's theory anymore. Her solution to representational content involves local information, which I've argued above doesn't work. This is still a useful exercise since I'm going to make use of it for my theory in the following chapter.

perform its function. The teleofunction of the consumer is to guide the bee to the honey; if the dancing bee misses a step, then when the interpreter bee interprets the dance according to mapping function *f*, the interpreter bee will not be guided to honey but somewhere else. Normally, however, bee dances do allow for the consumer to perform its function, and this is what (historically, causally) explains the existence and proliferation of honeybee dances and the representational machinery associated with them.

Thus, the bee dance has not represented falsely. Rather, it has failed to do what it is supposed to do. The dance has failed to map onto the world the way that historically has allowed the consumer to perform its function of guiding the bee to honey, even though that is the function of the dance. So it is an icon or a representation, but it is a (weakly normatively) defective one because it has failed in its function. By failing in its teleofunction, it has simply failed to represent in the same way that a heart can fail to pump while remaining a heart. It has not however represented falsely, and the theoretical tools that we are currently working with are not going to supply an explanation of that.

To get representing falsely, we need something more. We need something like subject/predicate structure. Millikan has noted this (Millikan 1990), but has failed to take its lesson sufficiently seriously. Millikan has further argued that representations or icons must have *structure* or be *articulate*. This is the way to go. However, again, she has not taken it far enough. The structure of icons in Millikan's theory does not provide an explanation of how icons can be false. What the structure or articulateness of representations does in her theory is to provide an explanation of productivity, but not of representing falsely.

To conclude this chapter: Millikan's work is insightful, useful, and provides some important pieces of the foundation for building a theory of representation. Specifically, the distinction between

what makes something a representation, and what determines representational content, is crucial. I've argued that something like Millikan's theory of teleofunction has to be right regarding the first question, but not the second. With respect to the second question, we still need a theory of mapping rules. By distinguishing representing falsely from failing to represent, as well as distinguishing the different kinds of normativity, we've made some progress toward that theory. But we still need a theory of content or aboutness that allows not just for the failure to represent, but the ability to represent falsely.

Chapter 5: The Nature of Representation I –Structural Preservation Theory

5.0 Introduction

The present chapter is the major positive chapter of this dissertation. In it, I construct a proposal for a naturalistic theory of representation based on what we have learned in the previous chapters. In brief, representational content is a structured relation with content determined jointly by structural preservation and causal history. The name of my proposal is the *structural preservation theory of original representation*, or just *structural preservation theory*. I'll proceed as follows. I begin with a brief review of our explanatory goals in terms of the core concept of representation, followed by discussion of some explanatory adequacy conditions for our explanandum. Then I'll characterize important points drawn from previous chapters, including several distinctions necessary for understanding representation. Then we'll get to the core of the chapter where I begin theory construction, through a discussion of isomorphism and several related concepts, then causal covariation and causal history, and the role that each of these, as well as teleology, play in constituting representations.

5.1 Adequacy Conditions and Review of Explanandum

Chapter 1 is largely devoted to identifying the target explanandum and defending the way that I've posed my research question. I provide only a brief review here.

Intentionality as traditionally understood by philosophers involves a relation one of whose relata need not exist, aboutness or directedness, the possibility of error, fine-grainedness and the

generation of referential opacity, and perhaps the assumption of minimal rationality. Most importantly, intentionality is traditionally understood and investigated in terms of folk psychological states, with a focus on language and the relations between language and thought (usually conscious thought accessible through introspection). By contrast, representation is a theoretical construct shared by several theoretical endeavors as well as our commonsense folk psychology. While folk psychology, philosophical semantics, the computational theory of cognition, connectionism, and the various neurosciences all make use of different notions of representation, there is a shared nucleus to each of their ontological posits as well. Specifically, on all of these theories, representations are states that have aboutness or directedness, the capacity for error, and are causally efficacious in the production and guidance of behavior.

Representation is my target explanandum, and my questions are as follows: (i) What is representation? (ii) How is it physically implemented? Question (i) is the *constitution* question, which asks about the nature of representation. Question (ii) is the *implementation* question[76], which asks about how representation, whatever it is, is implemented in the physical world. In chapters 5 and 6 I address the constitution question, and in 7 and 8 I adopt the incremental, dual-approach strategy introduced in 1.3.4 to simultaneously address the constitution and implementation questions.

What we want is a naturalistic theory of representation, consistent with a wider body of established physical theory that is mind and interpretation-independent. It should explain the capacities for aboutness and error in a physical system, while making room for the causal efficacy of representational states. It need not be overly constrained by language considerations or intuitions from folk concepts, but it should be implementable in the nervous system. It is desirable for it to be extendable to, and integratable with, other theories of the nervous system, behavior, and cognition.

[76] This is Michael Devitt's terminology. See my chapter 1, section 1.3.4.

5.2 Foundations

5.2.1 Two Preliminary Distinctions

There is a distinction between what makes something a representation and, given that something is a representation, what determines what it represents. The first element involves the metaphysics of what it is to be a representation, whereas the second involves the semantics of representations, or representational content. One of Millikan's significant contributions to this field was simply to recognize not only that these are distinct questions, but more importantly, that the answers to the two questions might not be the same. Millikan answers the metaphysical question with her theory of biological categories defined through teleofunction. As I've noted in the last chapter, while I do not here endorse the details, I do endorse the general strategy.

A general theory of biological categories should incorporate a unified explanation of the nature of cardiac and renal states just as much as it explains the nature of states of the nervous system. This provides a welcome unification of psychology with physiology. Further, biological devices have something that they are supposed to do, but can fail to do. What unifies biological devices, and what places them in their biological category, is the teleofunction or group of teleofunctions that they have. States of the brain aren't any different, and therefore neither are those states of the brain that are representational. Something like Millikan's theory of proper function is a good place to start for the general theory of biological categories, and something like Millikan's proposed conditions on intentional icons is a good place to start for an explanation of what it is to be a representation.

While we've got at least a framework and a starting place for answering the first representational question, we don't have an answer to the second. Questions about representational

content are at the heart of almost all discussions of representation. One of the chief obstacles to most theories of representation is an explanation of representational error. Fortunately, the above distinction helps to elucidate a further distinction, recognition of which is necessary to explaining representational error.

On the one hand, a representation is mistaken if it represents falsely. That is, the representation "says" of whatever it is about, something that is not true. The representation has representational content, but it is false. On the other hand, a representation is faulty or in error if it simply fails to represent. In this second case, the representation lacks representational content, even though it *should* represent. This is to be understood by analogy to a heart that does not pump: it is still a heart even though it does not fulfill its teleofunction. There is something weakly normatively defective with a heart that does not pump, just as there is something weakly normatively defective with a representation that fails to satisfy its teleofunction. It is still a representation, even though it "says" nothing at all. This is a different species of representational error than the first[77].

Our explanation of the second kind of representational error can be drawn straightforwardly from Millikan's work, as it goes hand in hand with her explanation of what it is to be a representation in terms of teleofunction. Our explanation of the first kind of representational error, by contrast, cannot be provided from within the context of a teleological theory of biological categories. The first kind of error involves representational content, and the explanation for the capacity to represent falsely must be built in to, or consistent with, our explanation of representational content (that is, our answer to the second representational question). To explain representational content, and representing falsely, we need representations with structure.

[77] I introduced strong and weak normativity in 4.4.3.2, and introduced this distinction between kinds of representational error in 4.4.3.3. See those subsections for further elaboration of these points.

5.2.2 Truth and Structure

By itself, neither a predicate nor a name is either true or false. Truth-evaluability, that is, the having of a truth value[78], is dependent on the application of a predicate to a thing, or, on something like the structure of a declarative sentence. For example, consider the sentence, "Johnny has green hair". The sentence is true if Johnny does indeed have green hair, and false otherwise. The subject operator 'Johnny' serves to refer to some *specific* thing, while the predicate 'has green hair' is applied to whatever the subject operator refers to (in this case, Johnny). One of the major problems for all of the authors previously discussed is that, while they each recognize in their own way something very much like this point, they fail to see its import. No explanation of error or representational content is forthcoming without *both* the subject and the predicate being part of representational content. By attempting to provide the semantics for predicate-like symbols in a language of thought (for example), it is not possible to, at the same time, provide an explanation of representational error, because predicates by themselves are neither true nor false, nor are names.

Goodman (1976) recognized something like this. He was evaluating representation more generally, rather than in the specific context of a naturalistic reduction of original representation, however, his general point transfers into our context. He says, "nothing is ever represented either shorn of or in the fullness of its properties. A picture never merely represents *x*, but rather represents *x as* a man, or represents *x to be* a mountain, or represents *the fact that x is* a melon" (Goodman 1976, 9). That is, a representation does not merely point to or refer to something, but it also predicates some property of whatever it refers to. He later introduces the terminology *representation-as*, as opposed to

[78] The phrase 'truth-evaluability' is here used to mean simply the having of a truth value, which is a metaphysical issue, not epistemological. It should not imply that anyone does, can, or must *evaluate* the truth value of some truth-valued representation.

representation (Goodman 1976, 27-31), where the distinction is essentially that between predication and reference, respectively. "A picture that represents a man denotes him; … a picture that represents a man *as a man* is a man-picture denoting him" (1976, 27-28). For Goodman, this latter, predicative function of a picture is actually a classification of the picture itself, rather than of the thing denoted by the picture. As a matter of exegesis, it would not be correct to say that Goodman's representation-as is predicative; however, the basic idea underlying his concept of picture-classification is just that.

> Thus with a picture as with any other label, there are always two questions: what it represents (or describes) and the sort of representation (or description) it is. The first question asks what objects, if any, it applies to as a label [we should see this as reference]; and the second asks about which among certain labels apply to it [this is the classification of the picture as, say, a man-picture or a horse-picture. But we should see it as analogous to predication; the picture predicates horseness (say), of whatever it refers to] (1976, 31).

Building on these distinctions, Dretske introduced what he called the *topic/comment* distinction. He says "there are always two questions that one can ask about representational contents. One can ask, first, about its reference – the object, person, or condition the representation is a representation *of*. Second, one can ask about the way what is represented is represented" (Dretske 1988, 70). The referent is the topic and the comment is what the representation says about that topic. In a later writing (1995, 26) Dretske argues that a representation has the function of indicating "the F [or, property] of those objects which stand in C to it [where C is a reference-determining context], but it does not have the job of indicating – *does not therefore represent* – which objects – or even whether

there is an object – that stands in C to it" (my emphasis). So for Dretske, the topic or referent, that on which the representation comments, is not part of the representational content. But without both reference and predication as part of representational content, we have something like an open sentence or a lone predicate, and these have no truth value.

Fodor also is aware of this distinction, but makes the same mistake that Dretske does. In his (1998, 23-39), Fodor discusses what he takes to be five "nonnegotiable" conditions on a theory of concepts. One of these is that "concepts are categories and are routinely employed as such" (1998, 24), by which he means the following. "To say that concepts are categories is to say that they apply to things in the world; things in the world 'fall under them'… Much of the life of the mind consists in applying concepts to things" (1998, 24). Predicates, like concepts, apply to things in the world, or, things in the world "fall under them". Further, Fodor takes predicate-like symbols in the language of thought to have concepts as their contents, where those concepts express properties. Fodor argues, "If, looking at Greycat, I take him to be a cat, then … I apply the concept CAT to Greycat. (If looking at Greycat I take him to be a meatloaf, I thereby apply the concept MEATLOAF to Greycat; incorrectly, as it happens)" (1998, 24).

Leave aside the talk of concepts and instead think in terms of predicates (either in thought or language), and Fodor's error should be apparent. He is correct in saying that, if he applies the predicate 'is a meatloaf' to Greycat he has made a mistake. However, the predicate 'is a meatloaf', by itself, is not in error, and neither is the subject term, 'Greycat'. The error only arises as a result of their concatenation. To have a truth value at all, both the subject term 'Greycat' and the predicate term 'is a meatloaf' must be a part of the representational content. But his asymmetric dependence theory does not even attempt to account for that:

> Let's start with the most rudimentary sort of example: the case where a predicative expression ('horse', as it might be) is said of, or thought of, an object of predication (a horse, as it might be). Let the Crude Causal Theory of Content be the following: In such cases the symbol tokenings denote their causes, and the symbol types express the property whose instantiations reliably cause their tokenings (Fodor 1987, 99).

He then proceeds to construct an elaborate theory (a "less crude" causal theory) of how 'horse' can misrepresent. But 'horse', 'cow', 'proton', and the other terms that Fodor tries to provide a semantics for, are neither true nor false. They lack a truth value.

Millikan noticed this problem, but failed to take it sufficiently seriously. She writes,

> A third contrast [between Millikan, Dretske, and Fodor] … is the special emphasis that Millikan alone places upon the *articulateness* of all complete representations. Complete representations represent complete states of affairs … A representation that represented something simpler than a state of affairs, one that represented, say, only an object or a property or a *type* of state of affairs (compare a propositional function) *would make no claim, hence would fail to be true or false, to represent anything either correctly or incorrectly*…" (Millikan 1990, reprinted in her 1993, 131; I've emphasized the last phrase).

Millikan recognized this early on in her (1984) when she argued that representations are structured entities[79]. However, the role of structure in her theory of content is only to provide for the productivity of representations, not truth-evaluability. In other words, Millikan does indeed argue that representations are structured entities and that the basic units of representation are analogous to complete sentences rather than sentence parts, and further she re-interprets the relation of sense to reference in light of this. However, what she fails to do is use that insight to do any significant work in her theory of the mapping rules that determine representational content. Relatedly, she fails to recognize that the structure of representations is what underwrites truth-evaluability and hence, is what makes possible representing falsely.

In his (1996), Cummins apparently had something like this distinction in mind, but further, he made some use of it by providing separate theories of what he called *target content* and *representational content*. Target content is what the representation is supposed to represent, while representational content is what the representation does represent. The content of the propositional attitude is the application of the representation to the target. Brian Cantwell Smith, commenting on Cummins (1996), has this to say: "*Targets*, that is, not representational contents, are what systems are intentionally directed towards … What representational contents are, in contrast, are *what the systems represent those targets as being like*" (Cantwell Smith 2002, 177). On this reading of Cummins, targets are much like Dretske's topic, while representational content is like Dretske's comment.

However, Cantwell Smith's interpretation of Cummins is overly charitable, and is in fact a misinterpretation. Rather than seeing Cummins' distinction by analogy with Dretske's distinction, we should see it as the distinction between what a representation is supposed to represent (its target), and what it does represent (its representational content). Cummins writes,

[79] See section 4.2, and especially 4.2.2 for detailed discussion of these points.

> The *theory of representational content* must explain what it is for something to be a representation, and what it is for a given representation to have a particular content [parenthetically, notice the conflation of the distinct metaphysical question and content question] … The *theory of target fixation* must explain 'the function of tokening a representation *r* is to represent *t*.' A crucial constraint is that target fixation must be independent of representational content (Cummins 1996, 20).

Cummins' theory fails because, while he does provide separate theories for the two elements of his distinction, he's got the wrong distinction in the first place. What makes representational content possible is the joining of something like predicative content with something like subject content, but his representation/target distinction does not map onto this.

The idea that there is a distinction between what a representation is about, or points to, and what a representation says regarding whatever it is about, is not by any means a new one. While Dretske and Fodor both recognized something like this distinction, they failed to see that, for truth-evaluability and thus error, both elements must be part of representational content. Millikan did recognize that both elements must be part of representational content in order to get truth-evaluable representations, but failed to do anything with it. Within the context of her work, she needed to make this distinction within the mapping rules that determine content, but instead, she used the idea of articulateness or structure to explain productivity, not content. Cummins also provided us with a clue, by arguing that what we need are two separate theories – one for each element of the distinction. But Cummins went off course because he was working with the wrong distinction. Putting each of their

contributions together, the next natural step in the evolution of this literature is clear: First, we need the distinction between predicative content and subject content, but they must both be part of representational content. Second, they must be determined separately. This is what will allow for an explanation of truth and error, because they can thus come apart.

A theory of representation must have many elements. It must answer the metaphysical question: what is it to be a representation? Something very much like Millikan's theory of biological categories is the best theory to answer that question. It must also be able to explain representational error, one component of which is the failure to represent when a thing has the function of so doing. This is also answered with Millikan's teleofunctions. A theory of representation must also answer the content question: given that something is a representation, in virtue of what does it represent what it does? Related to this question is the need for an explanation of the other component of representational error, which is representing falsely. In order for representations to have a truth value at all (and hence, be able to represent falsely), they must have something like subject/predicate structure, and this has to be part of representational content.

Representations are not analogous to sentence *parts*, they are analogous to complete sentences, and they have something like subject/predicate structure. However, it is crucial to emphasize that representations must have *something like* subject/predicate structure, but not necessarily be sentences in a language or in a language of thought, and further, must not necessarily be translatable into sentences in English. What makes this situation so difficult, as Millikan has pointed out[80], is that the only medium that we have for talking about the content of representations is language.

[80] She writes: "By stressing in the previous chapters that all complete signs signify complete world affairs, I may seem to have implied that complete signs are always translatable by sentences. To see why this is wrong we need to understand how signs are used to represent other signs. The difficulty lies in the fact that the only direct way we have to speak of what nonsentential signs represent is by misleadingly comparing them with sentences" (Millikan 2004, 87).

When we communicate in English about representations, we make use of further representations (through language use), and hence, have representations of representations. This causes some problems.

Language systems are relatively recent arrivals in the evolution of kinds of representations. There are simpler and evolutionarily older kinds of representations, and these are the ones of chief theoretical interest for someone interested in the naturalistic foundations of mind. However, language use dominates our lives, and it is likely the most common form of representation at work in our mental lives of which we are consciously aware. This may be one of the reasons why there has been such an overwhelming focus on language in theories of representation. Further, since we can only use language to discuss our theories of representation with each other, it is very likely that the properties of language, including constraints on what *can* be represented, on what *can* be the content of linguistic representations, get inadvertently slipped into constraints on what can be represented with non-linguistic representations, or on the form that the content of non-linguistic representations must take.

Millikan argues that what this suggests is that the possibility of saying precisely what a non-linguistic representation "means" (note scare quotes) depends upon the availability in our home language of an expression that has a matching teleofunction and a matching semantic mapping function (Millikan 2004, 89). This follows from her theory of intentionality. However what we can take from her discussion without importing any assumptions about a theory of representational content is that we need not be able to precisely translate what a non-linguistic representation "means" into English. Rather, we should be content with *describing in English*, as best as possible, the representational content of the non-linguistic representations (which are presumably the more basic ones), even though the linguistic content does not in fact match the representational content of the non-linguistic representation.

We'll need some terminology. For a representation to be true or false, it must (i) refer to, point to, or be about something, *and* (ii) it must say something, that is, predicate some property of, whatever it refers to. The representation is true if the object instantiates the property predicated of it, and the representation is false is the object fails to instantiate that property. Given the above warning not to interpret this in terms of a language, and given that this is intended to be a theory of the most fundamental kinds of representation, I'll use the terms 'f-reference' and 'f-predication', with the 'f' intended to connote 'fundamental'. F-reference and f-predication are analogous to reference and predication, but they are not to be construed as implying that all representations are symbols in a language of thought, nor, importantly, that a straightforward translation of the content of basic representations into English sentences is possible.

One final caution: The requirement that representations have something analogous to subject/predicate structure does not imply that the vehicles of representation have one physical component that is the referential expression, and another physical component that is the predicative expression. Rather, the structure can (but perhaps need not) be abstract. One example that Devitt likes to use in discussing whether representations can be simple is the yellow flag hung on a ship's mast to signify to other passing ships that the ship has yellow fever (Devitt and Sterelny 1999, 139). This seems like a simple, non-articulate, non-structured vehicle of representation. But it isn't. The fact that the flag *is yellow* signifies that, whatever ship is flying it, has yellow fever. But it is not the yellowness of the flag that signifies which ship has yellow fever. The fact that the flag is attached to *this* ship's flagpole is what determines the referent of the predicate, 'has yellow fever', as *this particular ship.* Thus, just one apparently simple vehicle can have different aspects to it, and those different aspects can separately determine the different aspects of representation. Presuming that, if the representation must have something like subject/predicate structure then there must be something analogous to a referring

symbol expression and a distinct predicative *symbol* expression is likely a result of not taking seriously enough the caution to avoid importing language requirements.

We now have all of the foundations and preliminary work behind us. In the next section I'll begin to develop the theory itself.

5.3 Theory Schema

I propose the following: A causal relation determines f-reference, and something like isomorphism determines f-predication. In this brief section all I seek to do is outline the motivation for this theory schema. In the following sections I'll provide more detailed discussions of these concepts, which will further motivate their involvement in representation.

Consider causal-informational semantics, but ignore the non-objectivity problem as well as the problem of providing the semantics for predicates alone. The major difficulty that informational theorists focus on is misrepresentation and the disjunction problem. No solution is forthcoming because, as I noted in chapter 3, this is a self-contradictory project. Fodor argues that the problem is showing how meaning reduces to information, which respects causal etiology, while simultaneously not respecting causal etiology[81]. And that isn't going to work.

We can draw some insight from this: there is no "misinformation", and the problem for informational theories is showing how an information-based semantics can account for falseness. But suppose also that, just as there is no "misinformation", neither is there any "mis-f-reference". That is, a representation either f-refers or it does not. (Perhaps it does not because it has failed to satisfy its

[81] See chapter 3, section 3.4 for discussion.

teleofunction in accordance with something like Millikan's theory; hence the representation is still a representation, but it has not performed its function.) We cannot reduce meaning to causal etiology while simultaneously showing how meaning does not respect causal etiology. But what we can do is reduce *f-reference* to causal etiology, without needing it to fail to respect causal etiology. Referring expressions are neither true nor false; similarly, f-referring "expressions" are also neither true nor false. Rather, they either f-refer or they don't.

Consider the major difficulty for isomorphism, structural preservation, and resemblance theories. The basic underlying issue is that each of the relations are nearly unconstrained. If isomorphism is the sole determinant of content, then representations are about or represent far too many things. We further lose the ability to have false representations, since a representation may be true under one isomorphism mapping but false under another, and if there is no fact of the matter as to what the content is, then there seems no way to account for error.

We can draw insight from this as well. Predicates, unlike subjects or referring expressions, are not specific. They can apply to many things; this is because properties are multiply instantiated whereas individuals are not. The multiplicity of isomorphisms, and the multiplicity of things to which predicates apply (due to the multiple instantiability of properties), suggests that the element responsible for f-predication in basic representations is isomorphism, or something like it. This also provides further motivation for thinking that causation is responsible for f-reference. Representations are specific, and that specificity derives from their f-referential component. Similarly, while isomorphisms and f-predication are not specific, f-reference and causation is.

Finally, as mentioned earlier, I propose that something like Millikan's work on biological categories provides the answer to the first, metaphysical question of representation as well as the species of representational error that involves the failure to represent.

We've got some bare bones of a theory to work with. What remains is to put some meat on them. We'll start with the concept of isomorphism, and through that discussion will provide further motivation for thinking that the preservation of structure is crucial to the nature of representation.

5.4 Isomorphism and Structural Preservation

5.4.1 Distinguishing Picture Theory from System-Isomorphism

The basic motivation behind all resemblance and isomorphism theories is that a representation represents what it does in virtue of either similarity or resemblance among representation and represented, or by sharing properties or relational structure. Aristotle's view[82], for example, was that there were two kinds of substrates that could have properties: the mental and the physical. For a mental state to represent something was for it to instantiate the same properties in the mental medium as the thing it represented, whose properties were instantiated in the physical medium. This view is not amenable to our present-day naturalism because it posits a non-physical substrate, but it makes the basic idea of resemblance clear.

To begin, we need some distinctions. No one these days argues that representation is underwritten by resemblance in the form discussed above. We'll start by distinguishing, as we did in chapter 4, picture theories from system-isomorphism theories. Aristotle's view was a picture theory: representations represent in virtue of the token vehicle of representation literally sharing properties or

[82] I rely on Haugeland (1985, 16) for this interpretation of Aristotle.

relations with what it represents, in much the way that a photograph or realistic portrait is thought to represent whatever it is of.

There is a different way of conceiving of resemblance however, which is the standard way that it is thought of these days. For now I'll refer to this with 'system-isomorphism'. Rather than sharing properties, *system-isomorphism* is the sharing of structure. *Isomorphism* is a concept from mathematics: A relational system is a set with relations on it, and two relational systems are isomorphic if there exists a certain kind of structure-preserving function between them. What this means is that the elements of the two sets map to each other in a one-to-one fashion, and further, the internal relations on the elements in one set are preserved in the other. This is not to say that the relations are the same, but only that there is a function *f* such that, if relation *R* holds between elements *a* and *b* in the first set, then relation *S* holds between elements *f(a)* and *f(b)* in the second, *a* maps to *f(a)*, *b* maps to *f(b)*, and *R* maps onto *S*. In this way, there is the preservation of internal structure that abstracts away from the particular properties or relations actually instantiated[83].

The concepts that we're going to work with draw their inspiration from measurement theory. In what follows we will take a look at some of the basic elements of that theory, and see how we can marshal them for our purpose in constructing a theory of original representation.

[83] I've chosen the terms 'picture theory' and 'system-isomorphism theory' in order to make immediately obvious the difference between the two. To connect this to the literature, Shepard & Chipman (1970) use 'first order isomorphism' and 'second order isomorphism' to describe this difference, while O'Brien & Opie (2004) use 'first order resemblance' and 'second order resemblance'. Cummins (1996) misleadingly calls his theory the "Picture Theory of Representation", even though it is a second-order or system-isomorphism view. Everyone wants their own terminology, I suppose. I will soon scrap this terminology for more precise language.

5.4.2 Measurement Theory

The application of numbers to empirical phenomena, or measurement, is a pervasive element of the physical sciences. As psychology broke from its early introspectionism and moved towards its current emphasis on inter-subjective verifiability and operational definitions, psychologists and other behavioral scientists began to devise ways of measuring psychological phenomena, such as intelligence. At this point the need for a systematic treatment of several importantly related questions became more pressing. Most importantly: what, if anything, justifies the application of numbers to empirical phenomena? Are we justified in applying numbers to phenomena, then reasoning about relations in the numbers, and thus reaching conclusions about the empirical phenomena on that basis?

In the following sections I'll present some of the elementary ideas of measurement theory which, with suitable adaptation, can be applied to my research. By the end of 5.4.5 I will have characterized a very general kind of resemblance or structural similarity, closely related to isomorphism, which I call *structural preservation*. Structural preservation is what accounts for f-predication.

I will be guided by two canonical treatments of basic measurement theory: Suppes and Zinnes (1963) and Krantz et al. (1971). In addition I will draw from an important philosophy article, Swoyer (1991). Swoyer argues that a number of seemingly disparate forms of representation are all species of a single relation, which he calls *structural representation*. He provides formal and informal characterizations of structural representation, also drawing on measurement theory. Further, he provides a treatment of what he calls *surrogative reasoning*: If you have a system that shares structure with whatever you want to reason about, you can reason about the surrogate system (e.g. a numerical system), then trace your steps back into the initial system to reach conclusions about the elements of the initial system. A general treatment of surrogative reasoning and its justification provides an answer

to the question from above: are we justified in reasoning about numerical systems in order to reach conclusions about empirical phenomena?

Swoyer does not attempt a naturalistic reduction of representation. Rather, his focus is a more general one, that of providing a unifying conceptual framework for understanding various types of representations, including measurement, mental representation, and linguistic representation. He also argues that certain phenomena not typically considered to be representational can be fruitfully studied in structural representational terms, such as ontological reductions and possible worlds semantics for intensional logics. Thus, foci typical of naturalistically inclined philosophers are not evident in Swoyer's work, such as misrepresentation, causal efficacy, or reductive explanation in non-intentional terms. This is not a criticism, but it does imply that we cannot simply import Swoyer's work wholesale for our purposes.

The use of measurement in the sciences is in essence an application of surrogative reasoning. We systematically assign numbers to empirical phenomena, then use familiar and convenient operations and relations over the numbers in order to reach conclusions about the empirical system. A major thesis of Swoyer's work is that the preservation of structure is what justifies this. Determining whether this holds is aided by some formalizations.

5.4.3 Representation Theorem

Measurement theory is centered around two fundamental problems: justifying the assignment of numbers to objects or phenomena, and specifying the degree to which that assignment is unique (Suppes and Zinnes 1963, 4). The first problem is solved by demonstrating that an isomorphism or homomorphism obtains between the empirical system and a numerical system. This is called proving a

Representation Theorem[84]. The second problem is solved by proving what is called a Uniqueness Theorem, to which we will turn in 5.4.4.

A *relational system* $\mathfrak{A} = \langle A, R_1, \ldots R_n \rangle$ is a set with relations on it, where A is a nonempty set called the *domain* of the relational system, and $R_1 \ldots R_n$ are relations on A. The *type* of relational system is defined in terms of the arity of the relations in $\mathfrak{A}$. For example, the relational system composed of $\{1,3,5\}$ with $<$ is of type $\langle 2 \rangle$, because $<$ is a binary relation. More generally, "if $s = \langle m_1, \ldots, m_n \rangle$ is an *n*-termed sequence of positive integers, then a relational system $\mathfrak{A} = \langle A, R_1, \ldots R_n \rangle$ is of type *s* if for each $i = 1, \ldots n$ the relation R_i is an m_i-ary relation" (Suppes and Zinnes 1963, 5). A relational system $\mathfrak{B} = \langle B, R, S, T \rangle$ with R and S binary and T a ternary relation is of type $\langle 2,2,3 \rangle$. For relational systems to be isomorphic they must be the same type.

Let $\mathfrak{A} = \langle A, R \rangle$ and $\mathfrak{B} = \langle B, S \rangle$ be relational systems of type $\langle 2 \rangle$. $\mathfrak{A}$ and $\mathfrak{B}$ are isomorphic if there exists a bijective[85] function $f: A \to B$ such that for every $a, b \in A$,

$$aRb \text{ iff } f(a)Sf(b).$$

If f is surjective but not injective, then $\mathfrak{A}$ and $\mathfrak{B}$ are *homomorphic*. These definitions generalize naturally to relational systems of other types[86].

[84] I will always capitalize the first letters in 'Representation Theorem' in order to remind my reader that this is a concept from measurement theory, not a philosophical treatment of naturalized original representation.

[85] A function is bijective if it is *injective* and *surjective*. A function is injective (or one-one) if each member of the range is mapped to by only one element of the domain. A function is surjective (or onto) if every member of the range is mapped to by some element of the domain.

[86] "Let $\mathfrak{A} = \langle A, R_1, \ldots R_n \rangle$ and $\mathfrak{B} = \langle B, S_1, \ldots, S_n \rangle$ be similar relational systems [that is, of the same type]. Then $\mathfrak{B}$ is an *isomorphic* image of $\mathfrak{A}$ if there is a one-one function f from A onto B such that, for each $i = 1, \ldots, n$ and for

A *numerical relational system* is a relational system with numbers in its domain, while an *empirical relational system* is a relational system whose domain includes "empirical" objects or properties, such as weights, heights, judgments, velocities, and so forth[87]. This last is not a rigorous definition, but I trust that the intuitive distinction is clear: empirical relational systems involve things in the world and their properties and relations, while numerical systems involve numbers and their relations (leaving open whether numbers are "things in the world" or not).

From the perspective of measurement theory, the first basic problem is to show that an empirical relational system of interest is isomorphic to an appropriately chosen numerical relational system. This justifies the application of numbers to the phenomena as well as surrogative reasoning about the empirical things with the numerical system as surrogate. The qualification "appropriately chosen" is added because with finite sets there always exists some numerical relational system that is isomorphic, but it may not be of any use because it does not facilitate surrogative reasoning. From our perspective, we need something different. We want to show that *two* empirical relational systems are isomorphic to each other, rather than each isomorphic to some numerical system.

Why would measurement theorists be interested in homomorphism, which involves only a surjective or onto function, but not injective or one-one, as is the case with isomorphism? In measurement, two rods for example may be the same length, and hence we would want to assign the same number to them, but this does not imply that they are identical. Only a function that is not injective, and hence a homomorphism and not isomorphism between relational systems would allow this.

---

each sequence $\langle a_1, \dots, a_{m_i} \rangle$ of elements of *A*, $R_i(a_1, \dots, a_{m_i})$ if and only if $S_i\left(f(a_1), \dots, f\left(a_{m_i}\right)\right)$" (Suppes and Zinnes 1963, 6). If the function is onto but not one-one then $\mathfrak{B}$ is a homomorphic image of $\mathfrak{A}$.

[87] The general definition of relational system is attributed to (Tarski 1954). To the best of my knowledge, the terms 'numerical relational system' and 'empirical relational system' were introduced in (Suppes and Zinnes 1963).

The choice of the domain in the numerical relational system is important, as it will become relevant when we apply these concepts to a naturalistic theory of representation. While Krantz et al. (1971) do not explicitly acknowledge this, if the domain of the numerical system $\mathfrak{N}$ is the real numbers then the only way for the function to be onto is if there are as many items in the empirical set as there are real numbers. Rather than make this assumption, measurement theorists typically require that $\mathfrak{A}$ be isomorphic or homomorphic to $\mathfrak{N}'$, whose domain is a subset of the reals, with the same relations as $\mathfrak{N}$. Swoyer takes note of this issue by defining isomorphic *embedding*: "if a mapping has all of the features of an isomorphism except being onto, it is an *isomorphic embedding*" (Swoyer 1991, 456). If there exists an injective (but not surjective) structure-preserving mapping $f: A \to \mathbb{R}$ then we would say that $\mathfrak{N}$ isomorphically embeds $\mathfrak{A}$. Alternatively, we can define N, a subset of the reals, as the image of A under f, define $\mathfrak{N}'$ in terms of N, and say that $\mathfrak{A}$ is isomorphic to $\mathfrak{N}'$. If f is onto but not one-one, then $\mathfrak{A}$ is homomorphic to $\mathfrak{N}'$. Swoyer does not mention this, but we can extend the notion of embedding to allow for homomorphic embeddings, in the obvious way: If f is neither onto nor one-one, yet it is structure-preserving wherever it is defined on both sets, then $\mathfrak{N}$ homomorphically embeds $\mathfrak{A}$.

5.4.4 Uniqueness Theorem

Demonstrating that a numerical system is isomorphic or homomorphic to an empirical relational system proves a Representation Theorem, and demonstrates the preservation of relational structure. This is what justifies the systematic assignment of numbers to things, or, this justifies measurement. But how unique is the assignment? Could others do just as well? These are the questions answered by proving a Uniqueness Theorem.

Consider a set A of rigid, straight rods[88]. We'll define a binary and a ternary empirical relation. When placed beside one another, with one end coinciding from each rod, either one rod extends beyond the other or they appear to coincide. (Intuitively, either they are the same length or one is longer than the other; but the notion of length presupposes a notion of measurement, which presupposes that numbers can be systematically applied to these empirical objects in a structure preserving way, which is just what is at issue for the theory of measurement.) When placed in this fashion and a extends beyond b then $a >_E b$, when b extends beyond a then $b >_E a$, and when they coincide then $a \sim b$. The subscript 'E' is there to remind us that this is an empirical relation, not its familiar numerical analogue. '$\sim$' denotes equivalence, not identity: a and b may indeed be of the same length, but this does not imply that $a = b$. We define a concatenation operation $\oplus$ by placing two rods end to end, and define it as $(a \oplus b) \sim c$ iff c coincides with a and b placed end to end. Let $\mathfrak{A} = \langle A, >_E, \oplus \rangle$. Let $\mathfrak{N} = \langle \mathbb{R}, >, + \rangle$.

One measurement procedure we might use that would take into account not only their ordering under $>_E$, but also their concatenation, would be something like this. Assume that a', a'', a''' etc., are all perfect copies of a, in the sense that when placed beside a, each appears to coincide at the other endpoint with a. If $a \oplus a' >_E b$ and $b >_E a$, then we would want to assign numbers so that $f(a \oplus a') > f(b) > f(a) = f(a')$. Further, we want to be able to characterize that $a \oplus a'$ is twice as long as a; hence, $f(a \oplus a') = 2f(a)$. We do this by constructing a *standard sequence*: $a, 2a = a \oplus a', 3a = (2a) \oplus a'', \ldots$. Then we set $f(a) = 1$. When measuring rod b, if it falls between na and $(n+1)a$, then we assign it a length anywhere between $nf(a)$ and $(n+1)f(a)$. As the selection of the first rod (a) gets smaller, measurement gets more precise. Thus, the construction of a standard sequence requires

[88] This example is from (Krantz et al. 1971), pp. 1-5 and 8-12.

the selection of a unit (a was selected as the unit above). The meter stick, for example, consists of the first 1000 members of a standard sequence with a millimeter-long rod chosen as unit.

With the selection of any one rod as the unit, the measurement of every other then falls into place as a ratio with respect to the unit. But the selection of any rod as having unit length was arbitrary. Assuming that a Representation Theorem has been proven for the system described above, the question remains: was the measurement completely arbitrary? What if we had chosen a different rod to have unit length? The answer is that the measurement of length is unique *up to a similarity transformation*. Let me describe what that means.

Suppose f defines a homomorphic embedding from $\mathfrak{A}$ to $\mathfrak{N}$, as does g $(g \neq f)$. Note that a similarity transformation is a function ϕ, from the real numbers to the real numbers, if there exists a positive real number α such that for every real number x, $\phi(x) = \alpha x$. To say that the measurement of length is unique up to a similarity transformation is to say that f and g are related to each other by a similarity transformation ϕ. That is, $g = \phi \circ f$. While the choice of a unit is arbitrary, the ratios of lengths are not. Letting a be an arbitrary rod and u be the rod chosen as the unit, $\frac{f(a)}{f(u)} = \frac{g(a)}{g(u)}$ for all f and g. Recall the discussion of the objectivity of the height of the Empire State Building (2.4.2): while we have the freedom to choose a unit, it is completely objective that the Empire State Building stands in some ratio to that unit length. For this reason measurement that is determined by functions that can be related via a similarity transformation are called *ratio scales*.

There are several other well-known types of scales, such as ordinal, interval, and log-interval scales. I've presented a brief discussion of ratio scales because it is a simple one that can be used to illustrate the concept of a Uniqueness Theorem. Ordinal scales are scales where only order is preserved (as opposed to the ratio scale which preserves the $>_E$ ordering as well as the concatenation operation).

These will become important for our discussion shortly. Interval and log-interval scales will not concern us here.

Swoyer provides an interesting discussion of the philosophical significance of proving Representation and Uniqueness Theorems. Proving a Representation Theorem explains the applicability of mathematics to reality (Swoyer 1991, 462-463), while proving a Uniqueness Theorem helps us to differentiate which aspects of a representational system are conventional (or artifacts) from those that are not (1991, 463). For example, the choice of a unit in a ratio scale is a convention, while the ratios of lengths to units are not. He makes a further distinction between *systemic conventions* and *mapping conventions* (1991, 467).

The conventional choice of a unit is a mapping convention. The choice only arises after we've chosen a numerical relational system. But there is a further convention of choosing *the system itself*. For example, instead of defining the numerical system with the addition operation, why not use the multiplication operation? Both of these systems preserve the relational structure of the empirical systems that we would like to measure, so the choice of one over the other is also a convention. It is a systemic convention.

With respect to my use of these concepts, proving Uniqueness Theorems does not have the same import as it does for measurement theory. Unlike measurement theory, I am interested in whether these relations obtain among two empirical relational systems, not an empirical and a numerical relational system, and this is a crucial difference. Further, we are investigating this in order to develop a theory of representation, not measurement. Proving a Uniqueness Theorem within the context of measurement theory shows (i) which homomorphism-determining functions are equally good (i.e. also preserve structure) and (ii) which aspects of the measurement are conventional and which are

not. Within the context of a naturalistic theory of representation, investigating transformations of the homomorphism or isomorphism-determining function should be done with an eye to ruling them out, not ruling them as equally acceptable. There should be *no* convention involved in which states are representations and what they represent, otherwise we will have run afoul of our naturalistic constraint. We'll need to investigate both systemic choices and mapping choices. To preview, we will be able to combine a causal requirement with teleological conditions in order to constrain which of the various systems and mappings determine representational content.

5.4.5 Extensions and Relaxations

The definitions of isomorphism, homomorphism, and their embeddings can be straightforwardly generalized to relational systems of other types, with the type of relational system defined in terms of the arity of the relations in the system. These definitions can be further generalized to allow for multiple sets in the domain and vector-valued functions[89]. This is important because many brain states are usefully defined along several dimensions and thus in terms of vectors.

Swoyer proposes several relaxations to homomorphism that may be appropriate for understanding representation. His goal, recall, is to explain what he has called *structural representation*. For structural representations, it is in virtue of the preservation of relational structure that a representation represents what it does. However, he is not focused on a naturalistic reduction of representation. As a provisional characterization, he defines the structural representation of one relational system by a second as the isomorphic embedding of the first in the second (Swoyer 1991,

[89] "Still more generally, we may have n sets $A_1, \ldots, A_n$, m relations $R_1, \ldots, R_m$ on $A_1 \times \cdots \times A_n$, and a vector-valued homomorphism ϕ, whose components consist of n real-valued functions $\phi_1, \ldots, \phi_n$ with ϕ_j defined on A_j , such that ϕ takes each R_i into relation S_i on Re^n" (Krantz et al. 1971, 9). For a function to "take" a relation into another is the familiar requirement that the elements $a, b, \ldots \in A$, stand in relation R iff the corresponding members $f(a), f(b), \ldots \in B$ stand in relation S.

457). He argues that this is not general enough, because there are several other cases that, intuitively, should count as representations but do not fit the isomorphic embedding characterization.

I will not evaluate his argument that these other cases ought to count as representations. He and I have different explanatory goals, and we would just be talking past each other. What I will do is list some of his concerns, and the subsequent relaxations on isomorphic embeddings that he proposes. My goal in this and previous sections is to provide characterizations of several related but distinct concepts, partly to draw them apart, but also to characterize the very general notion of *structural preservation*, which subsumes these various other concepts. What follows is an exposition and discussion of section 6, pp. 470-476, of (Swoyer 1991). I begin by listing some (but not all) of the concerns that motivate Swoyer's various relaxations.

(i) The requirement that relational systems be of the same type is too restrictive.

(ii) Sometimes f does not respect all of the relations in the original system but only some.

(iii) Sometimes relations are only preserved in one direction or the other, but not both. For example, aRb only if $f(a)Sf(b)$ preserves R in one direction.

(iv) Under the current definition, the representations are in the range of the function. There may be cases where we want them to be in the domain because the relevant mapping from representation to represented is many-one and hence not a function. For example, one person may have two names.

(v) There may be cases where we want to include elements in A that don't map to anything in B. So the requirement that the function be total (i.e. defined everywhere on A) is too strict.

Let θ be a subset of the relations in $\mathfrak{A}$. Then we will say that f is a θ-morphism just in case f respects all of the relations in θ[90]. We say that f respects relation R in $\mathfrak{A}$ when, for $a, b, \ldots \in A$, R holds of them iff S holds of their counterparts $f(a), f(b), \ldots \in B$. Allowing θ-morphisms accommodates (ii), so that f need respect only some, but not all, of the relations in $\mathfrak{A}$. Weakening the requirement so that f may be a θ-morphism also makes room for dropping the requirement that two relational systems be of the same type, which accommodates (i). For example, if $\mathfrak{A}$ is of type $\langle 2,2,3 \rangle$ (with the relations in $\mathfrak{A}$ R, S, and T, respectively) and $\mathfrak{B}$ is of type $\langle 2,2 \rangle$ then $\mathfrak{A}$ and $\mathfrak{B}$ are not isomorphic because they are not the same type. However, we can define $\theta = \{R, S\}$ and then investigate the existence of a θ-morphism from $\mathfrak{A}$ to $\mathfrak{B}$.

To say that a function respects a relation is to assert a biconditional. I will here consider relations R and S as binary for ease of exposition, but this can be extended to an n-place relation straightforwardly. For a mapping to *respect* R, it must be the case that aRb if and only if $f(a)Sf(b)$. But we can split the biconditional and define its parts as follows. A function *preserves* R only if $aRb \rightarrow f(a)Sf(b)$. A function *counter-preserves* R only if $f(a)Sf(b) \rightarrow aRb$, and thus a function respects R only if it preserves and counter-preserves R.

Let Δ and Ψ be subsets of the relations in $\mathfrak{A}$, and f a function from A to B. f is a Δ/Ψ-morphism just in case it preserves all of the relations in Δ and counter-preserves all of the relations in

[90] Swoyer defines isomorphism as a relation between what he calls *Intensional Relational Systems*, which are distinct from the relational systems that we have been discussing here. Because of this, his definition of isomorphism is slightly different than the one I have provided above, although the basic idea is the same. To make the changes that he proposes applicable to (non-Intensional) relational systems I have altered some of his definitions. I retain his core idea in all cases, as well as his notation wherever possible.

Ψ. Splitting the biconditional and defining its constituents in this way allows us to account for (iii), and makes it possible, if there is good reason, to define representations in these terms[91].

Considerations (iv) and (v) are easily accommodated. Regarding (iv), we no longer require that the representations are in the range, but that they can be either in the domain or range of the function. We will turn to a discussion of this in 5.4.8. In deference to (v), we do not require that f be defined everywhere on A. Swoyer does not provide a specific name for this last relaxation, so I'll use the following terminology. Let ϕ be a proper subset of A. If f is defined everywhere on ϕ but not everywhere on A, and f defines an isomorphism from $\mathfrak{a} = \langle \phi, R \rangle$ to any relational system $\mathfrak{R}$ then f defines an isomorphism* from $\mathfrak{A} = \langle A, R \rangle$ to $\mathfrak{R}$. We define homomorphism*, isomorphic and homomorphic embedding*, θ-morphism* and Δ/Ψ-morphism* in an exactly parallel way.

Structural preservation is the most general relation that subsumes each of the above. A function $f: A \rightarrow B$ structurally preserves $\mathfrak{A}$ in $\mathfrak{B}$ only if f defines an isomorphism, homomorphism, an isomorphic or homomorphic embedding, a θ-morphism, a Δ/Ψ -morphism, or the starred version of any of the previous morphisms.

The question of which kind of structural preservation is necessary for original representation is partly an empirical question and partly a conceptual question. The conceptual work is (at least) two-fold. First, I'll have to argue that some kind of structural preservation (at all) is necessary for representation. Second, the way we type relational systems makes a difference with respect to which kind of morphism obtains. When we're dealing with an isomorphism from an empirical to a numerical relational system it doesn't make much difference whether we type N in such a way that it is the image of f under A, and hence, $\mathfrak{A}$ is isomorphic to $\mathfrak{N}'$, or if we let $N = \mathbb{R}$ and thus, $\mathfrak{A}$ is isomorphically

[91] Swoyer for example would have us define representations of A in B as the items in B to which a Δ/Ψ -morphism maps, so long as Ψ is non-empty; that is, so long as at least one relation in A is counter-preserved (Swoyer 1991, 474).

embedded in $\mathfrak{N}$, but $\mathfrak{A}$ is *not* isomorphic to $\mathfrak{N}$ because f is not onto. Similarly, the difference between whether f is defined everywhere on A is a matter of whether certain items are included in A or not. For measurement theory and numerical systems these may be pedantic considerations, but for constructing a theory of original representation this has philosophical significance. For the time being, I simply wish to point out that there is a close relation between the way relational systems are typed and the various kinds of structural preservation (or lack of structural preservation) that obtain. This warrants further discussion, which is to be found below.

5.4.6 Empirical Axioms

I'll summarize what we've learned thus far, in order to get our bearings and remain focused on the task at hand. I've argued that a fruitful research strategy is to investigate the possibility of a theory of representational content which involves different components, one for f-reference and one for f-predication. The considerations I mentioned in 5.2.2 and 5.3 suggest that something like isomorphism be responsible for f-predication and causation be responsible for f-reference, and in 5.3 I provided a theory schema. Now we're beginning to fill in the schema.

Following many before me, I've argued that we must distinguish picture theories from system-isomorphism theories. To get a feel for the concepts of measurement theory of relevance to this project, in 5.4.2-5.4.5 I've reviewed the fundamentals of measurement theory. It is important to understand that the results of measurement theory do not automatically apply to the case of two empirical relational systems, which is the scenario relevant for a naturalistic theory of representation. See Appendix A for more detail on which aspects of measurement theory do and which do not apply to this special case. Additionally, in Appendix A I provide schemas for proving that one empirical relational system structurally preserves another.

I've introduced the term 'structural preservation', which denotes a broad class of relations among relational systems. Isomorphism, homomorphism, embeddings, θ-morphism, Δ/Ψ morphism, and the starred version of each of the above are all types of structural preservation. I've briefly noted the important relationship among the way relational systems are typed and the kind of structural preservation (or lack thereof) that they may bear to each other.

We may view the existence of structural preservation among empirical relational systems through the lens of empirical "axioms" on relational systems[92]. That is, we may investigate the existence and nature of structural preservation among empirical relational systems simply by determining certain of their properties. Here are some examples: Is the relational system finite or countable? Does the empirical relation induce a total or weak ordering over the elements of the set? Shall we assume order density? And so on. These are difficult questions whose answers must be both solidly conceptually grounded as well as empirically justified. Further, even once we have come to plausible, justified answers to these questions about typing empirical relational systems, that will only help determine *that* structural preservation exists, and what kind. Answers to the typing questions will not complete a theory of representation. That requires further work.

As I mentioned above, it is important to realize that measurement theory is concerned with the relations between an empirical and a numerical relational system. Properly speaking, measurement theory *does not apply* to the case of two empirical relational systems. Nonetheless some of the concepts from measurement theory are routinely, and appropriately, appealed to by philosophers seeking to understand representation. While it is easy to see the broad outlines of how some of these concepts might be applied to a theory of representation, the technical details of how they apply are not so obvious. Numerical relational systems have properties that not all empirical systems are guaranteed

[92] I borrow this idea and terminology from Krantz et al. (1971, 6-7).

to have, so a different strategy is called for. In Appendix A I address this issue, and show how to extend some of the results from measurement theory to the distinct case of two empirical relational systems[93].

A list summarizing the results from Appendix A should prove helpful, and can be referred to as we continue our investigation into the nature of representation. Below I speak of *range equivalence classes*, whose basic idea is this. Should there be a plausible empirical reason for associating members of an uncountable, totally ordered set with members of a countable set that partitions the uncountable set as a group of ranges, we can treat each range as something akin to an equivalence class. Then we construct a structure-preserving mapping using the *range equivalence classes*, thus having the technical machinery associated with countable sets available to us.

SUMMARY OF RESULTS FROM APPENDIX A

Finite Case

F1. If $\mathfrak{A}$ and $\mathfrak{B}$ are weak orders with finite domains, then $\mathfrak{A}'$ and $\mathfrak{B}'$, the associated relational systems generated by constructing equivalence classes, are total orders.

F2. There exists $f: A/\sim \rightarrow B/\sim$, a bijection, such that $\mathfrak{A}'$ and $\mathfrak{B}'$ are isomorphic only if $[A/\sim] = [B/\sim]$[94].

F3. f defines a homomorphism from $\mathfrak{A}$ to $\mathfrak{B}'$.

[93] Some readers may wish to pause here and read Appendix A to get a better feel for how to connect two empirical relational systems while preserving structure.

[94] Notation: I use '[X]' to denote the cardinality of the set X. The set $A/\sim$ is the set of equivalence classes constructed of members of A.

F4. If $[A/\sim] > [B/\sim]$ then f defines an isomorphism* from $\mathfrak{A}'$ to $\mathfrak{B}'$. If $[A/\sim] < [B/\sim]$ then f isomorphically embeds $\mathfrak{A}'$ in $\mathfrak{B}'$.

F5. If we assume that $\mathfrak{B}$ is a total order, then $\mathfrak{A}'$ is isomorphic to $\mathfrak{B}$, $\mathfrak{A}$ is homomorphic to $\mathfrak{B}$.

F6. If we assume that both $\mathfrak{A}$ and $\mathfrak{B}$ are total orders then $\mathfrak{A}$ and $\mathfrak{B}$ are isomorphic. Similar remarks regarding the cardinality of A and B as those in F4 apply here.

Infinite Countable Case

C1. If $\mathfrak{A}$ and $\mathfrak{B}$ are countable total orders, and if B has an order dense subset then there exists a function $f: A \to B$ that defines an isomorphic embedding of $\mathfrak{A}$ in $\mathfrak{B}$ or an isomorphism of $\mathfrak{A}$ and $\mathfrak{B}$.

C2. If $B = Im(f|A)$[95] then $\mathfrak{A}$ is isomorphic to $\mathfrak{B}$. If $Im(f|A) \subset B$ then $\mathfrak{A}$ is isomorphically embedded in $\mathfrak{B}$.

C3. If we do *not* assume that $\mathfrak{A}$ and $\mathfrak{B}$ are total orders, but instead only assume weak ordering, then $\mathfrak{A}'$ and $\mathfrak{B}'$, the associated relational systems generated by constructing equivalence classes, are total orders. Then if $B/\sim$ has an order dense subset then there exists a function $f: A/\sim \to B/\sim$ that defines $\mathfrak{A}'$ as isomorphically embedded in $\mathfrak{B}'$. Under these assumptions, $\mathfrak{A}'$ and $\mathfrak{B}'$ are isomorphic only if $B/\sim = Im(f|A/\sim)$.

C4. Similar remarks apply here as to the finite case. If we assume total ordering on B but not A then $\mathfrak{A}'$ is isomorphic to $\mathfrak{B}$ and $\mathfrak{A}$ is homomorphic to $\mathfrak{B}$.

[95] Notation: '$Im(f|A)$' denotes the image of A under f.

C5. If there are upper and lower bounds to B (or $B/\sim$, as the case may be), that is, if we assume the existence of b_0 and b^{super} (or their associated equivalence classes), then if $b_0, b^{super} \in B$ then f defines an embedding. If $b_0, b^{super} \notin B$ then f defines an isomorphism, and mutatis mutandis for the function on equivalence classes.

C6. Let A^{RE} partition A, with *A* countable and A^{RE} finite ('RE' is intended to connote 'range equivalence'). Then the function described above for the finite case can be used to map range equivalence classes of *A* to members of *B* (with *B* finite), defining an isomorphism, embedding, etc. from $\mathfrak{A}^{RE} = \langle A^{RE}, \geq_{AR} \rangle$ to $\mathfrak{B}$. The relation $\geq_{AR}$ is defined in terms of $\geq_{AE}$, as $a \geq_{AE} b$ iff $a^{re} \geq_{AR} b^{re}$, where a^{re} and b^{re} are the range equivalence classes associated with *a* and *b*, respectively.

C7. Given range equivalence classes as above, the function that determines isomorphism from $\mathfrak{A}^{RE}$ to $\mathfrak{B}$ is associated with a function (that is, $f_1(a) = f_2(a^{re})$) that defines a homomorphism from $\mathfrak{A}$ to $\mathfrak{B}$.

Uncountable Case

U1. If *A* and *B* both have countable order dense subsets, and $\mathfrak{A}$ and $\mathfrak{B}$ are both total orders, then $\mathfrak{A}$ is isomorphic to $\mathfrak{B}$.

U2. Let A^{RE} partition A, with *A* uncountable and A^{RE} countable. Then the function described above for the countable case can be used to map range equivalence classes of *A* to members of *B* (with *B* countable), defining an isomorphism, embedding, etc. from $\mathfrak{A}^{RE}$ to $\mathfrak{B}$.

U3. Given range equivalence classes as above, the function that determines isomorphism from $\mathfrak{A}^{RE}$ to $\mathfrak{B}$ is associated with a function that defines a homomorphism from $\mathfrak{A}$ to $\mathfrak{B}$.

5.4.7 Typing and Idealization

5.4.7.1 What Has Been Proven?

Could there be more than one "equally good" structure-preserving mapping between empirical relational systems? What role does the empirical analogue of a Uniqueness Theorem play in a theory of representation?

Assume that we prove that structural preservation obtains between two empirical relational systems. What, specifically, would be proven? The answer is simply that some function, with particular properties, exists. This would show *that* structural preservation exists, and further, what kind (given assumptions on how the relational systems are typed). What it wouldn't do is determine which out of (possibly) infinitely many functions, is the *representation function* (or, Millikan's *semantic mapping function*).

This issue should be understood in relation to Swoyer's *mapping conventionalities* (1991, 467). Given that a relational system has been chosen, it is partially a matter of convention (in the context of measurement theory- *not* original representation) which mapping function is used to measure the empirical phenomenon of interest. Proving a Uniqueness Theorem then helps to distinguish which elements are truly conventional or artifacts, and which are not. In the context of a theory of representation, the situation is different. To satisfy the naturalistic/reductionist constraint, no aspect of representation can be a matter of convention. Thus, there is no empirical analogue of a Uniqueness Theorem. Changing the function, even if it is with an empirical similarity transformation, will change the

content of the representation. What we need then is a way to distinguish which of the various mappings is *the* representation function.

Another way to think of this issue is as one element of the non-uniqueness problem discussed in 4.4.1 and 4.4.3.3. In those sections we discussed Millikan's theory of the mapping rules that determine representational content, and in particular, the role played by system-isomorphism. Given the tools provided by Millikan's theory of proper functions, we were able to provide criteria by which we may choose which of the various mapping functions from $\mathfrak{A}$ to $\mathfrak{B}$ is the *semantic* mapping function. A second component of the non-uniqueness problem, which we have not yet discussed, is something analogous to Swoyer's *systemic conventionalities*. Here choosing among the various mapping functions from $\mathfrak{A}$ to $\mathfrak{B}$ is not what is at issue, but rather what is at issue is what the nature and properties of $\mathfrak{A}$ and $\mathfrak{B}$ themselves are. Or, which relational systems are involved in representation at all? This component of the non-uniqueness problem has an answer as well, and that answer involves at least two parts. First, causation helps to determine which of the relational systems are those involved in representation. Second, we need independent considerations on how to type empirical relational systems in general as well as for each particular system. It is to this second element that we now turn.

5.4.7.2 Typing Empirical Relational Systems

Near the beginning of 4.4.1, while briefly mentioning some of the traditional objections to resemblance theories, I warned of the following potential source of confusion. We must not conflate or confuse *stipulating* relational systems that are isomorphic to some relational system of interest, with *discovering* whether independently specified relational systems are in fact isomorphic. This distinction begs the important question of this subsection: How are we to independently specify empirical relational systems? This was not as acute in the context of measurement theory, as it didn't make much

difference whether we typed numerical systems in such a way as to define an embedding or an isomorphism. Further, the ability to prove a Uniqueness Theorem allowed us to see which of several mappings (which can involve different elements in the numerical sets and hence different systems) are equally good, at least from the perspective of their ability to preserve structure. Here it does make a difference because different mappings and different relational systems, while they may preserve structure equally well, will result in different representational content when structural preservation is marshaled for a theory of representation.

This issue introduces difficult and foundational questions in ontology. While a full treatment is beyond the scope of this dissertation, the questions are too important to be ignored. I begin the work here, and discussion of this issue is going to carry over into the remainder of this dissertation.

Given the results from measurement theory plus my extension of those results to empirical relational systems, the question of whether structural preservation connects two empirical relational systems can be investigated by way of an investigation into the properties of those systems. The investigation into the properties of those systems is mostly an empirical matter to be settled by experiment and evidence; or, it is going to involve establishing whether brain states and non-physiological states of the world satisfy certain empirical "axioms" or conditions. It also involves conceptual work to determine whether typing brain states in some particular way is a legitimate, sensible, justified way of so doing.

The first consideration I offer to aid in the conceptual task is an obvious one: The manner in which empirical relational systems are typed must not be motivated solely to provide support for, or falsification of, my theory. Here is a brief example. Most neurons, as we've learned from empirical work, reach a certain threshold in the voltage difference across the cell membrane. When voltage

reaches that threshold, certain biochemical and mechanical processes occur which result in the cell's voltage increasing exponentially, then decreasing to a point below its "resting" level. These exponential increases are known as spikes or action potentials. One way of typing the states of a neuron that fires action potentials is in terms of the rate at which it fires them. The members of the set are rates of action potential firing, and the relation over that set is the greater-firing-rate relation. A constraint here is that there should be an element of the set that corresponds to the neuron's not firing at all as a lower bound, and an element of the set that corresponds to tetanic firing, which is the physiological upper limit to firing rate. Whether or not we type the cardinality of the set as infinite (countable or otherwise) depends on whether we should idealize firing rate, in the way we idealize the permeability of containers and the elasticity of molecules in the ideal gas law. We'll return to idealization. By contrast, other neurons exhibit what are called graded potentials, where there is no voltage threshold beyond which the voltage spikes exponentially. It makes no sense to type the states of these neurons in terms of firing rate, but it does make sense to type them in terms of their voltage. This introduces further considerations in typing specific to voltage. What I wish to show with these two brief examples is simply that *there are* considerations, independent of structural preservation or my theory of representation, that motivate the typing of empirical relational systems. This is what underlies the difference I alluded to in 4.4.1, between stipulating isomorphic systems, and discovering whether independently specified systems are in fact isomorphic.

The requirement that typing considerations be independently motivated is too weak. Set membership is unconstrained[96], relations are subsets on Cartesian products of sets, and so they too are unconstrained, making relational system membership unconstrained. Requiring independent

[96] Set membership is unconstrained, except for some relatively technical constraints motivated by the need to avoid Russell's paradox. We can ignore this for our purposes.

motivation for how we specify relational systems is a good start, but it does not provide enough guidance. Further guidance is to be found in two places.

First, it seems prima facie plausible that some idealization is going to be involved in typing empirical relational systems. Second, the issue of how best to type empirical relational systems is most fruitfully understood only from within specific contexts, rather than as a set of general guidelines. The two examples provided above give a taste for some of the independent considerations that might come into play, but those considerations are not general principles. Thus, the question of typing is best understood from the perspective provided by specific experimental contexts.

Part of the problem, as mentioned above, is that set membership is unconstrained. There are an infinite number of types and hence relational systems, and while we have the freedom to name any arbitrary type we want, we don't have the freedom to effect type membership, and hence relational system identity, in the first place. That is an objective matter[97]. However, while there are infinitely many types and hence relational systems, there must be ways of discovering, not stipulating, which of those are involved in representation. Nature herself must have a way of specifying empirical relational systems, and the task for science is to discover them. From the investigative paradigm that I have set up, the task for philosophy is to investigate, given the relational systems discovered by science, whether they are connected by structural preservation or not, and further, whether we have any reason to believe that those empirical relational systems have something to do with representation. A prior, independent task for both philosophy and science, is that of justifying the legitimacy of the way that science has typed the relational systems in the first place.

[97] The 'freedom we have and don't have' is Devitt's terminology, from (Devitt 1997, 245). This point with respect to realism about kinds and kind-membership is his.

5.4.7.3 Idealization

Idealization may become involved in the typing of relational systems. Is this detrimental to my theory? Can it be avoided? Unfortunately, I don't have an answer to either of these questions. In lieu of answering them, I'll present several concerns that are relevant, and leave the answers for another day (and hope that those answers do not prove fatal to my theory).

First, consider what Swoyer has to say on this (and keep in mind, as discussed above, the difference between Swoyer's and my use of 'representation' for what follows):

> The use of [relational systems] in dealing with actual cases of representation … involves an element of idealization, for most of the things we represent, and most of our representations of them, are not literally [relational systems]. Moreover, there is a risk of confusion here, since [a relational system] may itself be regarded as a sort of model or representation of the real-life situation we use it to study… Strictly speaking, the relationship of structural preservation holds only between [a relational system that models] a real-life situation and [a relational system that models] a representation of that situation (Swoyer 1991, 457).

Swoyer's theory of structural preservation is couched in terms of what he calls *Intensional Relational Systems* (or *IRS*s), which involve non-extensional properties. These are distinct from the relational systems that I have been discussing. It is fine for Swoyer's purpose to use 'Intensional Relational System' in the way that he does, where they are diverse from the real-life situations that they model. However for our purposes this won't do. I use the notion of a relational system simply as a tool to understand the nature of original representation. I need actual, non-abstract states of nervous

systems to bear some naturalistically suitable relation to other states; hence, my empirical relational systems do not *model* a real-life situation, they *are* real-life situations (or more carefully, the elements of the domain of the relational system are states of affairs that are related via the empirical relations that constitute the remainder of the relational system).

In the footnote to the above quoted paragraph, Swoyer considers the possibility of treating *IRS*s as identical to, not modeling, real-life situations. He says (Swoyer 1991, 501, fn. 9):

> In some cases an *IRS* model of a situation or phenomenon … injects a hefty dose of idealization … We frequently treat actual things as point masses or ideal speaker-hearers or objects that have perfectly definable lengths, even though we know that there really aren't any such things. This often enables us to provide reasonably tractable structural representations of actual systems in well-understood mathematical systems. In such cases, we can still think of the *IRS* model as a faithful depiction of something actual, though now it is the scientist's idealized version of a real-life system, rather than the system itself, that the representation depicts.

Thus, Swoyer claims that even when we treat *IRS*s as real-life things, they are really only idealizations of the real-life situations, but not identical to the actual situations themselves. This is counter to what I have claimed, but again, Swoyer's *IRS*s are distinct from my empirical relational systems. However, Swoyer's claim that even when we consider *IRS*s as real-life situations we must always treat them as idealizations of and hence non-identical to the actual situation suggests a parallel claim for the empirical relational systems that I wish to use. We must ask, for example, if idealization is involved in treating greater-voltage-than, or greater-firing-rate as inducing a weak or a total order on

the system. We must ask if treating the domain of any empirical relational system as infinite involves idealization and if so, (i) whether it is justified and (ii) even if it is justified, whether this creates a problem for a theory of original representation that attempts to use these idealized relational systems.

Let's have an example. Assume that it is justified and unproblematic to type the domain of some empirical relational system (either physiological or non-physiological) as non-finite. If we are to do any work at all with this, we need to make the assumption that there is a countable order dense subset for at least one of the two relational systems in question. On what grounds should we assume order density for an empirical relation? I don't know how we would go about determining whether order density actually holds for an empirical relation, or what would justify making that assumption. Given this admission, what options are available to me?

The first option is to eschew the empirical assumption as implausible or unmotivated. In that case we may still type the (non-physiological) empirical relational system in such a way that we can partition it into ranges, and then make use of the technical machinery available for countable or finite classes. A second option is to explicitly idealize the relational system. Rather than claiming that, say, a set of voltages and the greater-voltage-than relation that orders that set are order dense, we say that the *ideal* voltage relational system is order dense. Then claim that this is no more problematic than the claim that, under ideal conditions, the pressure of a gas in a closed container varies inversely with its temperature. Everyone agrees that these ideal conditions are never satisfied, but nonetheless everyone also agrees that Boyle's ideal gas law is *true*. Perhaps idealization with respect to typing empirical relational systems is no more or less problematic than it is elsewhere.

5.4.8 Structural Preservation and Representation

Two important questions yet to be addressed are these: Are original representations to be found in the domain or range of the function that connects empirical relational systems? Are there any reasons, specific to a theory of representation, that lend support for some particular kind of structural preservation's involvement in representation, to the exclusion of the other kinds? Let's start with the second of these, although keep in mind that answers to one constrain answers to the other.

O'Brien and Opie (2004) present what they call a *structuralist* theory of mental representation. They distinguish first from second-order isomorphism, and then characterize weaker forms of second-order isomorphism, ending up with what they call *second-order resemblance* (O'Brien and Opie 2004, 11). Second-order resemblance is weaker than both isomorphism and homomorphism; it is essentially what I have called a *homomorphic embedding* (I will continue to use my terminology for continuity). These authors claim that this very weak form of structural preservation is what is important for *representation*:

> In the literature on second-order resemblance the focus is often placed on isomorphism …, but where representation is concerned, the kind of correspondence between systems that is likely to be relevant will generally be weaker than isomorphism. In what follows, therefore, we will tend to avoid this restrictive way of characterising [sic] resemblance (O'Brien and Opie 2004, 13).

This assertion lacks argument, and nowhere in their paper do they provide one. I see neither an intuitive nor other prima facie reason why any particular kind of structural preservation should be

involved in representation to the exclusion of others. Mandik, Collins, & Vereschagin (2007) have argued that for representation, isomorphism, to the exclusion of homomorphism, is relevant to understanding representation. We argued as follows.

First, for independent reasons, we argued that causal-informational relations must underwrite representation. Further, for these causal-informational relations to obtain in evolved organisms, we argued that the information-bearing states of an organism must enter into isomorphisms with the states that they carry information about. Information (contingently) requires isomorphism, at least given the way that Darwinian selectional processes occur:

> The representations attributed are states of the nervous systems of the creatures that represent environmental (and bodily) states in virtue of carrying information about those and a requirement on the acquisition by the organism of such states is that the states enter into isomorphism relations between neural and other structures (Mandik, Collins, and Vereschagin 2007, 90).

I no longer hold an informational view of representation, for the reasons discussed in chapter 2. What concerns me here is our argument for why isomorphism but not homomorphism is relevant to understanding representation.

The crux of our first objection to homomorphism is that we wanted determinate content; or, we wanted to rule out multiple contents: "We especially want to avoid attributing multiple contents to one and the same representation, as in, saying of a height of the mercury column that it represents

multiple temperatures" (Mandik, Collins, and Vereschagin 2007, 91). With homomorphism, it seemed to us that it was possible that one and the same representational vehicle could have multiple contents, since multiple items from the domain of the function could map to the same item in the range.

Our second objection involved discriminability:

> We do not want to attribute multiple contents if the organism is not capable of distinguishing them … The attribution of contents to an organism is an attempt to portray the world as it is carved up by the creature's point of view: Elements of the world that the creature cannot distinguish cannot make a difference discernible from the creature's point of view (Mandik, Collins, and Vereschagin 2007, 91).

Our arguments in that paper were flawed. First, according to our own stated theory, isomorphism is irrelevant to representation. We claimed that isomorphism is contingently necessary for information, as an organism could not evolve to have information-bearing states unless it entered into isomorphisms with whatever it carries information about. But it is also the case that an organism could not evolve to have information-bearing states unless (presumably, I suppose) $e = mc^2$, or the 2nd Law of Thermodynamics held, or any number of other fundamental physical laws and processes that are necessary for evolution to work the way it does, held. But that doesn't imply that any of these things determine representational content. What we proposed was an information theory accompanied by a proposed explanation of what makes the carrying of information possible. We did not propose an isomorphism theory of representation. Thus, each of the considerations based on worries about

multiple contents do not apply to our theory. However, let us assume that they did and discuss each of them in turn.

Our first problem with homomorphism is that we feared it would allow for the attribution of multiple contents. However, this worry depends on an unstated and unjustified premise: the representations are in the range of the empirical function. We stated that it was a desideratum that we can say, for example, "of each height of the mercury column, whether it represents a temperature and, if so, which one" (2007, 91). Suppose that the representation function runs from the heights of the mercury column to temperatures, and defines a homomorphism or homomorphic embedding from the two empirical relational systems. Then we can say, of each height of the mercury column, whether it represents a temperature (is f defined over that height?) and if so, which one (what is the value of f at that height?). Further, the worry about attributing multiple contents is ruled out: if f maps any element in its domain to more than one object in its range then f is not a function and hence structural preservation does not exist. We also claimed that, "we want to say of each temperature, if it is represented by a height of the mercury column and, if so, which one" (2007, 91). But wanting this (coupled with the previous desideratum just above) is just another way of saying that we want to define an isomorphism. It does not constitute an argument for why we *should* want to define an isomorphism. Nor does it constitute an argument for why isomorphism but not homomorphism is relevant to representation. Further, it does not seem out of the question that multiple vehicles of representation may represent the same thing. I go by different names in different company, for example. Thus, each of these worries only arises when we assume that the representations are in the range, not the domain, of f. If these are worries (it has not yet been *argued* that multiple contents are problematic), then this constitutes an argument that representations are in the domain. It does not tell for or against

isomorphism to the exclusion of other structural preservations. What about the multiple contents worry?

One reason that we need some level of determinacy is to allow for error. If representational content is wildly indeterminate then it may turn out that no sense can be made of representations ever being false. But this consideration does not speak for or against multiple contents. For example, content can be indeterminate but bounded: Perhaps firing rate x has energy range $[y, z]$ as its content. Does x determinately represent every element within the range, or just some, or an endpoint, or the class itself? Further work would be needed to answer this, but indeterminate yet bounded content does allow for x to have content $w \notin [y, z]$ in error. The worry about multiple contents, I think, arises from the failure to divorce folk psychological intuitions about the contents of beliefs and desires from our theory construction regarding basic-level representations. While I reject meaning rationalism in general, all I need here is to make it plausible that meaning rationalism does not hold for the basic, lowest-level, underived representations at issue here. There is certainly no reason to believe that we, or lower organisms, have unerring introspective access to the contents of our fundamental, underived representations. Whether meaning rationalism holds for beliefs is a separate question.

Our second concern was with discernibility; we claimed that we did not want to attribute multiple contents if the organism is not capable of distinguishing them. This is problematic, first, because it confuses epistemology with semantics. What the *creature* can distinguish is an epistemological question; what the *states* of the creature represent is a semantic/representational question. They are distinct. Second, this objection assumes meaning rationalism, from the point of view of the creature, even for the most basic representations: If the creature cannot know the difference (i.e., cannot distinguish), then there can be no difference in the content of its states. But why assume

meaning rationalism for the most basic level representations, even for the most basic creatures (in the paper under discussion, we attributed these representations to the nematode, *C. elegans*)?

We also mentioned that attributing contents was a way of carving up the world from the creature's point of view. I now take this to be mistaken. Attempting to carve up the world from the creature's point of view (i.e., distinguish things as the creature distinguishes them) may be why *we attribute* contents, instrumentally perhaps. But attribution of content, and why we do it, are each distinct from representational content itself. I assume that representation is a real, objective relation between two states in the world, which obtains whether or not we or anyone else *knows* that it does. The epistemological/metaphysical divide holds everywhere else. I see no reason why it should break down here. This is not to say that understanding what the creature distinguishes or can distinguish (as well as what that amounts to) is unimportant. It is however to say that that is distinct from understanding the objective nature of representational content. Finally, this objection also depends on an aversion to multiple contents, but again, we lack any argument for being so averse.

What we should conclude thus far is that there aren't any sound a priori arguments in favor of ruling any kind of structural preservation in or out. And this is as it should be. Understanding the nature of reality requires a combination of both conceptual and empirical research, working concurrently, where each serves to inform, correct, and constrain the other. We are approaching the point where we must turn to empirical evidence to arbitrate.

Our worry about multiple contents was partially based on the unsupported assumption that representations must be in the range, not the domain, of the representation function. Now let's consider whether there are independent reasons for choosing one over the other.

The guiding motivation behind all resemblance and isomorphism theories is that a representation represents what it does in virtue of either similarity or resemblance among representation and represented, or by sharing properties or relational structure. From this perspective, internal relations are "preserved" and structure is "mirrored" whether we consider representations to be in the domain or range. The basic motivation for resemblance theories does not tell one way or the other for this question, so we must get into the details.

Important disparities emerge when we consider the different kinds of structural preservation, especially isomorphism and homomorphism. If $\mathfrak{A}$ is isomorphic to $\mathfrak{B}$ then there exists a bijection $f: A \to B$. However, a function is invertible iff it is a bijection (Bond and Keane 1999, 116), hence there exists an inverse of f (denoted f'), from B to A. In the isomorphism case, it is arbitrary whether we place the representations in the domain or range of f. Let's say, for example, that A is the domain of a physiological relational system (hence, consists of the representations), while B is the domain of a stimulus relational system (hence, is the range of the function f), whose members are the representeds. If $\mathfrak{A}$ is isomorphic to $\mathfrak{B}$ then $f: A \to B$ is a bijection and is invertible. Let $f': B \to A$ be the inverse of f. Then we can say that the representations are in the range, rather than the domain, of the function f', which defines an isomorphism between $\mathfrak{A}$ and $\mathfrak{B}$ just as well as f does. The decision here is arbitrary because both f and f' are guaranteed to exist under the assumption of isomorphism, and the representation to represented mapping remains identical under both.

Consider homomorphism instead. If there exists a surjective but not injective function $f: A \to B$, then f does not have an inversion. Then it makes a difference to representational content whether we specify the function in such a way that the representations are in the domain or range. Suppose that $a_1, a_2 \in A$ both map, under f, to $b_1 \in B$, and f defines a homomorphism from $\mathfrak{A}$ to $\mathfrak{B}$. If the representations are in the domain of f then every representation must have a unique content (from the

definition of a function), although there will be some representations that represent the same thing (a_1 and a_2 both represent b_1, for example). If we take the representations to be in the range, then we get the possibility discussed above, of representations with multiple contents (b_1 will have both a_1 and a_2 as its contents). If we take multiple contents to be problematic, this would seem to provide argument that representations are in the domain, not the range, of f.

It is worth noting that from the perspective of measurement theory, the representations are taken to be in the range. Empirical relational systems map to numerical relational systems, and numbers represent empirical properties or relations. Suppes and Zinnes (1963, 6-7) explicitly note that homomorphism is generally the more appropriate kind of structural preservation because multiple distinct items may share the same magnitude, and hence should be represented with the same number. Thus, sometimes we want our representing numbers to have multiple contents; this is only possible if we take the representations to be in the range, not the domain.

We should note once again that there is nothing intrinsic to the notion of a mathematical function and its relationship to original representation that should lead us either way on this. Second, rejecting meaning rationalism, at the very least for the lowest-level, basic representations which are our target here, tells us that we do not have unerring introspective access to the contents of original representations. Hence, the intuitive oddity of representations with multiple contents, generated by folk psychological intuitions, is neither conclusive nor even applicable here. Third, failing the intuitive discomfort with multiple contents, there aren't any a priori reasons to reject multiple contents (so long as they are bounded and hence allow for error). So the multiple contents objection is inconclusive here. What further considerations are available to us?

Consider a stimulus relational system consisting of vibrotactile frequencies, and a physiological relational system consisting of neural firing rates. Clearly, we want to say that the firing rates represent the frequencies, but not the other way around[98]. The question presently under consideration reduces to how we define, or search for, the function connecting these empirical relational systems. Is it from firing rates to stimulus frequencies, or from frequencies to rates? Assume that we have some empirically plausible reasons to individuate the frequencies more finely than the set of firing rates. If this is a reasonable assumption, then the set of frequencies has more members than the set of firing rates. A bijection thus cannot exist between them, while a surjection can. The surjection can only be defined from the set of frequencies to the set of firing rates, with more than one frequency mapping to the same firing rate. But, since we've already concluded on independent teleological grounds that the firing rates, not the frequencies, are the representations, we thus have an argument for considering representations, in general, to be in the range. Further, we have an argument that concludes that sometimes representations have multiple contents.

This argument depends on individuation claims which need independent justification, and I won't attempt that here. Further, and critically, this argument relies on empirical evidence to help answer questions posed by conceptual research.

Thus if we want to avoid multiple contents we can always define the empirical function in such a way that the representations are in its domain. However, empirical considerations that might seem to justify individuating relational systems in such a way that there are more elements in the stimulus relational system than there are in the physiological system, weigh in favor of defining the function so that the representations are in its range. Since an aversion to multiple contents is ungrounded (at least

[98] It is worth noting that I use the word 'clearly' here *not* to signal a reliance on intuition, but because the teleological considerations, guided by Millikan's theory of proper functions, tell us that firing rates *are* representations (although it does not tell us *what* they represent), while vibrotactile stimulus frequencies are not.

for the basic representations at issue here), the weight of the reasons seems to fall on the side of representations being in the range. However, this is an issue that is probably best left open for now.

Now we have some meat on the bones of structural preservation theory, which goes like this. There are two questions to ask about representation: what is it to be a representation, and what determines the content of representations? Representation, like other aspects of cognition, is ultimately grounded in biology, thus something like Millikan's theory of proper functions which define biological categories is a good starting point for the first question. For the second question, we recognize that something like subject/predicate structure is what makes possible the having of a truth value. A naturalistic theory of representational content for original, basic representations should have two parts, one for f-predication and one for f-reference. F-predication is explained by structural preservation: assuming that *R* is a vehicle of representation, *R* has the f-predicative content (this is simply the predicative component of *R*'s content) that it does in virtue of being a member of an empirical relational system that structurally preserves a second empirical relational system. How to explain what *R* f-refers to, or, *R*'s f-referential content, has yet to be determined. Some questions that are left open thus far as are follows. What kind of structural preservation determines f-predicative content? Is it always the same to the exclusion of the others? Do different kinds of basic representations have their f-predicative content in virtue of different kinds of structural preservation? How do we solve the systemic component of the non-uniqueness problem? For this last question, one part of the answer is that we distinguish stipulating relational systems from discovering whether independently specified relational systems are in fact structurally preserved in each other. The other part of the answer involves causation, which will simultaneously provide the explanation of f-reference as well as solve the systemic component of the non-uniqueness problem. Finally, the mapping component of the non-uniqueness problem was addressed in 4.4.1, in the context of Millikan's theory.

Briefly, teleological considerations can help to determine, given two relational systems, which of the numerous structure preserving mappings between the two is the representation function.

Now that we have a clear understanding of structural preservation, I will conclude 5.4 with a no-miracle argument. The non-uniqueness of structural preservation *motivated* its use in a theory of representation; non-uniqueness does not guarantee that the theory constructed from that motivation will be any good. The no-miracle argument that follows concludes that, in fact, structural preservation *must* be involved in representation.

5.4.9 A New No-Miracle Argument

The success of the physical sciences demands a non-miraculous explanation. That success is explained by the assumption that our most successful physical theories are, at least approximately, true (granting the vagueness of approximate truth). But that in turn demands explanation: What is it about our best physical theories that make them true, or at least, capable of being true? The only answer here is that our theories *represent* the world (and further, they adequately or accurately represent the world). What is it about our theories that allow them to represent the world? To answer this, we look at our most successful physical theories, and we find that they share a very important commonality: measurement. Our most successful physical theories all involve the quantification and measurement of empirical phenomena, and the characterization of relations among various empirical quantities or magnitudes. What makes measurement possible, as we learned above, is the preservation of structure, from an empirical relational system to a numerical relational system. This suggests that what makes the truth of our theories possible is representation, and what makes representation possible is structural preservation. This suggests that some form of structural preservation lies at the heart of original

representation, and in fact that that is what representation *is*. Everything else (causation, teleology) is just gravy. The meat and potatoes are here.

This no-miracle argument isn't really new; it is very similar to something that Swoyer has argued. He says:

> Still, the point of much representation is to *mediate inferences* about the things in the world, and this raises what might be called the *applications problem* … How can *any* representational system – from rudimentary arithmetic to a complex natural language – be successfully applied to the world? How is such representation possible? I believe that the *best explanation* why a mathematical theory applies to the concrete phenomena it does is that it has many of the same *structural features* as those phenomena. It is a central thesis of this paper that *shared structure* of precisely this sort explains the applicability of a wide range of representational systems – including many *non*-mathematical ones – to the things they represent (Swoyer 1991, 451, emphases in the original).

Swoyer's applications problem seems to me to be a hybrid of two questions: (i) What is representation? (ii) What is it about representational systems (of all sorts) that gives them their utility in mediating truth-preserving inferences about the world? The first question is the topic of this dissertation. The second question is a question about the utility of representational systems, both neural and non-neural. Swoyer is correct that what answers the second part of the applications problem, what gives representations their utility, is that they allow for surrogative reasoning, and this is made possible through structural preservation. Further, from the no-miracle argument above, we

should conclude that, in addition to their utility in allowing surrogative reasoning, what allows representations *to represent*, that is, to apply to the world, is structural preservation.

With respect to *original* representation, we need more than simply a mathematical mapping from one relational system to another, for the various reasons outlined previously. What we need, in short, is a way of choosing among (i) the relational systems that are related via structural preservation (i.e. an answer to the systemic component of the non-uniqueness problem) as well as (ii) the functions that map the two systems in a structure-preserving way (i.e., an answer to the mapping component of the non-uniqueness problem). However, Swoyer's point still holds: the reason why we are able to use things like numerical relational systems and so forth to represent things is that they share structure with what we use them to represent. They just aren't original representations since they do not satisfy the further teleological and causal history requirements on original representation, to which we will turn shortly.

Measurement and quantification are at the heart of the success of the physical sciences. The application of numbers to empirical phenomena, that is, the use of measurement, *in every case*, makes use of structural preservation. We should conclude then that representation *is* structural preservation. While the non-uniqueness of isomorphism was a partial *motivator* for my theory, the new no-miracle argument *justifies* the use of structural preservation as the cornerstone of my theory of original representation.

5.5 Causal History and Nomically Grounded Causal Covariation

To understand original representation we must distinguish the metaphysical question from the content question. Representational content in turn must be understood in terms of its components, f-predication and f-reference. The theoretical framework presently under construction makes use of structural preservation to explain f-predication, and causation to explain f-reference. Now let's have a look at causation and the role that it plays in structural preservation theory. Afterwards, in section 5.6, I put all of the components together.

As argued in 5.2.2, causation is a viable candidate for explaining f-reference. While f-predication is multiply applicable, f-reference is specific. Further, what allows a representation to be specific is that it f-refers. Unlike structural preservation, causation is specific, and so is suited to play this role. Additionally, one of the major problems besetting causal-informational theories of content is that it does not seem possible to explain misrepresentation in terms of causal relations. However, f-reference is neither true nor false and so there is no need to explain "inaccurate f-reference". A state either f-refers or it does not. Finally, causation plays a role in the context of f-predication, by helping to solve the systemic component of the non-uniqueness problem, and thus selecting among a (potentially infinite) number of relational systems.

While causation is well-suited to play the roles I've mentioned above, we need to determine which kind of causal relations are of interest here: causal history or counterfactual causation? On the first option, we make use of simple etiology: if *a* caused *R* (and the teleological requirements on being a representation have been met), then *R* f-refers to *a*. On the second, we make use of counterfactual-supporting regularities: *a* reliably causes *R*, or, *a* would cause *R*, so (again granting further assumptions) *R* f-refers to *a*. I argue that causal history, not counterfactual causation, determines f-reference. However, counterfactual covariation plays an evidential role in structural preservation theory. First let's have a closer look at counterfactual causation, and more particularly, its relation to information.

I'm going to introduce a new term for a familiar concept, and the reason is that the familiar terms for that concept have some further connotations associated with them (namely, their conflation with information) that I'd like to avoid. The new term is *nomically grounded causal covariation*. Nomically grounded causal covariation is a species of counterfactual causation; it is a relation that obtains between two states of affairs, where one reliably causes the other. That reliable causation is grounded in the laws of the universe: because law *L* holds, *x* reliably causes *y*. I will neither quantify nor attempt to make more explicit the notion of "reliability", for reasons that will become clear momentarily. It is nothing other than lawful causal regularity, but I'll ask for my reader's indulgence in using my term so as not to confuse nomically grounded causal covariation with information. Let's look at why they are distinct.

First, keep in mind the five, possibly six (or more) different senses of 'information' that float around in the philosophical literature (see 2.1.2 for discussion of this). There is the colloquial, non-rigorous sense, which is associated with content or meaning. Then we have the concepts from engineering: mutual information, average information (entropy), and surprisal value, all of which are distinct quantities. We have Dretske's informational content, which is a semantic concept of information that is supposed to be a bridge between the colloquial and the technical concepts. Then we have 'indication' (i.e., Grice's 'natural meaning'), which is supposed to be synonymous with the fifth, semantic sense. Finally, Fodor sometimes uses the term 'information' to mean, essentially, lawful causal regularity. While there are several different senses of 'information', this last use is a conflation of two separate things; information is not a lawful causal regularity.

One fundamental claim that we must accept is that there *are* lawful regularities. That is, there are non-backtracking, counterfactual-supporting regularities in the world. Without that, inductive reasoning isn't possible, and without our ampliative forms of reasoning, the success of the sciences that

make ineliminable use of them would be miraculous. I take it as given that there are lawful regularities in the world, even without any precise explication of what that amounts to. Nomically grounded causal covariation is a kind of lawful regularity. It is an objective feature of the world, and the numerous nomic causal covariations that science finds are objective and to be discovered, not stipulated.

Information, on the other hand, is not objective, but is a product of our minds (this was argued in 2.4.2). In every case, the definition of 'information' depends on assumptions that make information dependent on minds. It depends on a difference between signal and channel conditions, which difference is relative to our interests and explanatory goals. Or it depends on background assumptions about which possible outcomes are relevant and which are not, to define frequency-based probabilities. Or in Millikan's local information, it depends on natural domains which constitute reference classes to which statistical frequencies are relative. In this last case, as with the previous two, the reference classes are themselves defined in terms of relevance, and hence, are mind-dependent (see 4.4.2 for discussion). I do not argue that any of this is a problem, unless we attempt to use information for a naturalistic reduction of mind (or in our case, original representation). More relevant to the present discussion, since surprisal value, average information, mutual information, local information, and informational content are all non-objective, but nomically grounded causal covariation is objective, it follows that nomically grounded causal covariation is not any one of these. Information is not mere causal regularity. On the ground of its objectivity at least, it would seem that covariation (though not any kind of information) is still a candidate for determining f-reference.

The relationship between nomic covariation and (the several kinds of) information is nonetheless a close one. That which makes information non-objective is the attempt to quantify it. Assigning numerical values to probabilities of events at a source is what allows for the quantification of bits of information, but it is also what makes information non-objective. Nomically grounded causal

covariation, by contrast, drops the quantification, and no attempt at numerical rigor is made. We should see the various forms of information, and in particular mutual information, as fallible epistemic guides to the underlying, objective, lawful regularities in the world, although not identical to those regularities. Further, while not identical, information is also not a quantification of the degree to which those regularities obtain. Or at least, information is not an objective quantification of the degree to which those regularities obtain, in the way that measurement of length is an objective quantification (up to a similarity transformation) of an objective, mind-independent property.

As an epistemic guide to the existence of nomically grounded causal covariation, mutual information has great utility for the practice of both science and philosophy. Ultimately, that fallible guide depends on our mental states, since it depends on relevancy assumptions. While representation cannot reduce to information, it does not follow that representation (or more specifically, f-reference) does not reduce to that to which information is a guide, namely, nomically grounded causal covariation. One these grounds at least, counterfactual covariation is still a candidate for determining f-reference.

Causal history is distinct from nomic covariation. The former only involves the actual causal etiology of some event or state, and does not deal with counterfactuals. Neither is it very closely related to information in the way nomically grounded causal covariation is. It is an objective feature of the universe, so no problems arise on that front.

The traditional reason why causal history was rejected in favor of some kind of a reliable covariation theory was to make room for error. If *R* represents everything that causes it, then *R* never misrepresents. On the other hand, suppose we determine that *R* lawfully covaries with *a* but not *b*. In the case of "wild" tokenings of *R*, caused by *b*, it was hoped that we could explain *R*'s misrepresentation of *b* in terms of reliable covariation. This consideration does not apply under the present circumstances.

Error is explained in terms of both f-reference and f-predication, but no theory of "inaccurate f-reference" is necessary.

I claim that causal history, not covariation, determines f-reference. Suppose that *a* reliably causes (i.e. nomically covaries with) *R*, but on one occasion, *b* causes *R*. Should we say that *R* f-refers to *a* or to *b* (on that occasion)? If *R* f-refers to *b*, then we have an explanation of error that goes like this. In virtue of the structural preservation between a relational system of which states of *R* are members, and a relational system of which states of *b* are members, *R* f-predicates (say) *F*-ness of *b*. *R* f-refers to *b* in virtue of etiology. If *b* is *F*, then *R* is true; if *b* is not *F*, then *R* is false. That gives us a fairly neat and compact explanation of error. On the other hand, covariation seems to introduce two different potential sources of error, resulting in indeterminate content.

Let's suppose that *R* f-refers to *a*, even though in this case, *b* caused *R*. The f-predication story from above is the same. But in this instance, *R* f-refers to *a*, not *b*. Does *R* f-refer in "error"? That can't be right since we've assumed, somewhat axiomatically, that there is no such thing as erroneous f-reference. But further, we've learned from the various extended attempts to make sense out of an error-theory based solely on covariation, that it doesn't work, and we cannot make that story work. In that case, we'll have something like a disjunction problem for f-reference: *R* would have to f-refer to the disjunction, (*a* or *b*), but to neither *a* nor *b* individually. That also doesn't work, because it is the singularity of causation that we need in order to help deal with the systemic component of the non-uniqueness problem, and in order to make sense out of error and f-reference. On these grounds then, we should conclude that causal history but not nomically grounded causal covariation, is what determines f-reference.

Millikan (1984, 2004) makes a distinction between indicative and imperative representations, as well a combination of both, which she calls pushmi-pullyus. Indicative representations are analogous to indicative sentences, or to beliefs, in that they say how the world is. Imperative representations are like imperative sentences or desires, in that they represent a desired state of affairs. This is a common distinction, which maps onto something like the distinction between sensory and motor representations. It is an important distinction, and any theory of original representation should be able to account for both.

My theory accounts for them as follows. For indicative representations, *R* f-refers to whatever caused it, and f-predicates whatever property the representation function maps to *R*. For imperative/motor representations, *R* f-refers to whatever it caused, and f-predicates some property in similar fashion to its indicative cousin. How then do we determine whether *R* f-refers to what it caused, thereby making it an imperative, or whether *R* f-refers to what caused it, thereby making it an indicative?

To answer this question, let's return to the basic story. The first representation question is the question of what it is to be a representation, and something like Millikan's teleological theory of proper functions gives us the answer to that question. However, at the same time, it also gives us the answer to whether a given representation *R* is either indicative or imperative (see 4.2, and especially 4.2.1). To review Millikan's theory, in the case of imperative representations, the consumer/interpreter device of that representation has as a proper function the production of conditions in the world onto which the representation maps in accordance with the semantic mapping function. For indicative representations, the Normal explanation of how the consumer device performs its proper functions (whatever they happen to be), makes reference to the fact that the representation used by the consumer maps, in accordance with the semantic mapping function, to some state of affairs in the world. Determining *that*

R is a representation thus simultaneously determines *what kind* of representation it is. Then we need to determine the semantic mapping function, or, we need to answer the second representational question, which is the question of content. Given that *R* is a representation, and given what kind, independent typing considerations and causal history determine relational systems, and selects from among them in order to solve the systemic component of the non-uniqueness problem. Teleology returns to the story to determine which of the numerous mapping functions between the given relational systems is the representation function. Assuming that it obtains, structural preservation determines what *R* f-predicates of whatever it f-refers to, and causal history determines what *R* f-refers to. If *R* is a motor/imperative representation, then *R* f-refers to what it caused. If *R* is an indicative representation, then *R* f-refers to whatever caused it.

To conclude this chapter, I will provide an overview of structural preservation theory, in a simplified format that I hope will prove helpful.

5.6 The Structural Preservation Theory of Original Representation

Q: What makes something a representation?

A: A representation is a state of a biological organism, and states of biological organisms fall into biological categories in virtue of their teleofunction. We distinguish two kinds of representations: indicative and imperative. Both indicative and imperative representations have the teleofunction of bearing some correspondence relation to a state of affairs. The difference between indicative and imperative representations involves their relations to the biological devices that use them. The consumer of an imperative representation has the teleofunction of producing a state of affairs such that

the representation corresponds to it. Regarding the consumer of an indicative representation, the Normal explanation of its teleofunction (whatever it is), makes reference to the fact that the representation bears some correspondence relation to the world[99].

Q: Does teleofunction determine content?

A: No. Teleofunction determines *that* a thing is a representation as well as what kind of representation it is, but not, say, that *R* represents that *a* is *F*. Teleofunction however plays a role in selecting among various mappings (see below).

Q: Given that *R* is a representation, in virtue of what does *R* represent what it does?

A: To have a content, to be true or false, *R* both f-refers to something and f-predicates some property of that thing. *R* has the content that *a* is *F* if *R* f-refers to *a* and f-predicates *F*-ness of *a*.

Q: Given that *R* is a representation, in virtue of what does *R* f-refer to *a*?

A: Depending on what kind of representation *R* is (indicative or imperative), *R* f-refers to what caused it for indicatives, or what it caused for imperatives. Note that this is actual causal history, not counterfactual causation.

Q: Given that *R* is a representation and that *R* f-refers to *a*, in virtue of what does *R* f-predicate *F*-ness of *a*?

A: If *R* is a member of the domain of a relational system *RS-1* , *F* is a member of the domain of a relational system *RS-2*, *RS-1* and *RS-2* are structurally preserved in each other, and the representation function f maps *F* to *R* (or *R* to *F*), then *R* f-predicates *F*-ness of *a*.

[99] This is Millikan's work, not mine. See (Millikan 1984, 100) and my section 4.2.1.

Q: In virtue of what is f a representation function?

A: A mapping function f from the domain of one empirical relational system to the domain of another empirical relational system is a representation function if the members of one of those relational systems' domains are representations, and those representations have the teleofunction of being mapped to *thusly* (that is, according to f) from the members of the domain of the other relational system. More colloquially, if the representations have the teleofunction of bearing the f correspondence relation to some states in the world, then f is a representation function[100].

Q: How does error fit in to this theory?

A: *R* is in error if it fails to map according to f, even though it has that as a teleofunction. This is the kind of error I've called *failing to represent*. Second, *R* is in error if it predicates *F*-ness of *a*, but *a* is not *F*. This is what I've called *representing falsely*.

Q: Can you give an example?

A: Chapters 7 and 8 include several detailed examples of how the brain realizes representations as defined by structural preservation theory.

Resemblance and isomorphism are widely considered to be hopeless for a theory of representation. In the next chapter I consider several objections to these kinds of theories, and I argue

[100] Recall that I've argued that, provisionally at least, we should consider the representation function to run from represented to representation. Hence, f maps *F* to *R*, yet we say that *R* has the teleofunction of "mapping to" *F*, or of bearing a correspondence relation to *F*. The correspondence relation is the "inverse" of f. I use the scare quotes because a function is invertible iff it is bijective, but not every representation function will be bijective. If f is not injective, then inverting it will result in one item mapping to multiple things, and not being a function. In terms of speaking loosely of a correspondence relation, that is fine, even though it is not a function.

that structural preservation theory has the resources to account for each one of them. Through a discussion of these objections I will further clarify and refine the theory.

Chapter 6: The Nature of Representation II – Traditional Objections to Resemblance Theories

6.0 Introduction

Resemblance theories, picture theories, map theories, image theories, and isomorphism theories of representation are all related, and each is widely considered to be hopeless. Fodor tells us "I do think that this is the closest to solving a mind/body problem that we've ever gotten. And I guess that these days everybody understands how much it depends on rejecting the idea that mental representation is a species of resemblance" (Fodor 1994, 86). Godfrey-Smith (1994, 268), discussing Millikan, says that "[her] presentations sometimes draw also on the vocabulary of another, largely discredited, approach to truth and meaning, the 'picture' theory". Cummins (in his 1989, 87) says "[a] central insight of the seventeenth century was that mental meaning cannot be understood in terms of resemblance"[101]. In a discussion of correspondence truth, which ultimately is to be understood in terms of representation, Devitt writes, "the crude notion of resemblance (or mirroring) has no place in contemporary correspondence theories" (Devitt 1997, 50).

While I do not defend a resemblance, picture, map, or image theory of representation, I do rely very heavily on structural preservation. Like the others, this is considered to be completely discredited. I intend to challenge the orthodoxy.

Since the non-viability of this sort of theory is so widely accepted, it would seem that discussion of the various objections deserves its own chapter, albeit a shorter one. In this chapter I address as many of the objections to resemblance or image theories as I have found. I argue that structural

[101] Cummins has subsequently changed his view on this, and indeed argues that isomorphism does underwrite representation (Cummins 1996).

preservation theory has the resources to account for each one of them, and further, most of the objections are a species of one underlying worry: isomorphisms are non-unique. However, that underlying concern is actually the motivation for constructing my theory as I did; thus it is not a liability but an advantage of the theory. Finally, I conclude with a further worry about the causal etiology component of the theory.

Most of the objections to resemblance/similarity/isomorphism as the ground of representation were either explicitly stated in, or can be culled from, Nelson Goodman's *Languages of Art* (1968; I will focus on the 2nd edition from 1976). Fodor, especially in chapter 4 of *The Language of Thought*, adds some additional objections. I begin with Goodman.

6.1 Goodman's *Languages of Art*

Goodman did not object to resemblance specifically as the ground of *mental* representation; his work was intended, at least partially, as a work in aesthetics. It is widely assumed that his general objections transfer over to the specific case of mental representation. The first two objections are that, while resemblance is reflexive and symmetric, representation is neither of these:

> Some of the faults [of resemblance] are obvious enough. An object resembles itself to the maximum degree but rarely represents itself; resemblance, unlike representation, is reflexive. Again, unlike representation, resemblance is symmetric: … while a painting may represent the Duke of Wellington, the Duke doesn't represent the painting (1976, 4).

Let us begin with reflexivity. Any relational system maps to itself in a structure-preserving way: Simply map each member of the domain of that system to itself. This is known as the *identity morphism*. If structural preservation alone were sufficient for original representation, this would be a troubling objection. However, first, my theory splits into two parts: what makes something a representation is distinct from what makes representations represent what they do. If an object or device fails to satisfy the teleological criteria on being a representation, then regardless of what it is isomorphic to, it has no representational content because it is not a representation. Of the biological mechanisms that do satisfy those teleological criteria for being a representation, structural preservation alone does not determine representational content. Two elements are required: structural preservation and causal history. Things do not cause themselves; hence, even though relational systems do trivially map to themselves, they do not have themselves as their representational content. The causal requirement rules out this objection.

Symmetry is also ruled out by the causal requirement: if *x* caused *y* then *y* did not cause *x*. The teleological requirements break the symmetry as well. If a stimulus relational system is isomorphic to a physiological relational system and vice versa, it does not follow that the elements in the stimulus relational system satisfy the teleological requirements on being a representation. In fact, they most likely do not. Finally, just because $\mathfrak{A}$ is structurally preserved in $\mathfrak{B}$ does not imply that $\mathfrak{B}$ is structurally preserved in $\mathfrak{A}$. As discussed above, a function is invertible iff it is a bijection; hence, only in the case of isomorphism will there be a straightforward inversion of structural preservation. Nonetheless, just because there is no inversion of some homomorphism determining function from A to B does not imply that there isn't a distinct function from B to A that also determines a structure preserving morphism from $\mathfrak{B}$ to $\mathfrak{A}$. My reply to this objection is, in a sense, overdetermined: Both causation and teleology

break the symmetry, so there is no need to mention this last point regarding the asymmetry of structural preservation. However, I mention it because it is important to see the distinction between resemblance and structural preservation, and not to think of structural preservation metaphorically in terms of resemblance. Some of the objections to resemblance carry over to structural preservation, while some do not.

Goodman continues, arguing that resemblance is not sufficient for representation: "[I]n many cases neither one of a pair of very like objects represents the other: none of the automobiles off an assembly line is a picture of any of the rest; and a man is not normally a representation of another man, even his twin brother" (1976, 4). No kind of structural preservation is sufficient, either for being a representation or for having representational content, on my theory. So this objection does not apply.

He also argues that resemblance is not necessary for representation: "The plain fact is that … no degree of resemblance is sufficient to establish the requisite relationship of reference. Nor is resemblance *necessary* for reference; almost anything may stand for almost anything else" (1976, 5). Reference, for Goodman, is "the core of representation" (p. 5). We don't need to delve into Goodman's theory of the relations between reference and representation. All that matters here is that, if resemblance is neither necessary nor sufficient for (Goodman's) reference, then it is neither necessary nor sufficient for representation. This would be a challenging objection, but Goodman supplies no argument for it. To be fair, Goodman is not concerned with a naturalistic reduction of mental representation (he probably would think it a hopeless project from the start since he does not share my realist presuppositions). Thus, in a certain sense he may be right: consider Dretske's coins and popcorn pieces that stand for basketball players as a result of stipulation (Dretske 1988, 52). In the sense in which stipulation counts for reference, Goodman is right that anything can be used to refer to anything else. But that's not the sense that matters here. What would be a counterexample to my theory would

be a case of original representation in which structural preservation does not obtain. No argument has been provided that such a counterexample exists, so we can leave this objection aside.

Goodman makes the above four complaints explicitly. There are two more objections to a naturalistic theory of mental representation that are implicit in his work. The first is the singularity of mental representation: a mental representation, such as a thought or a concept, can be about a single item, while not being about all the members of some given class. However, Goodman writes,

> a picture, like a predicate, may denote severally the members of a given class. A picture accompanying a definition in a dictionary is often such a representation, not denoting uniquely some one eagle, say, or collectively the class of eagles, but distributively eagles in general (Goodman 1976, 21).

Fodor explicitly makes the connection to mental representations:

> The concept *tiger* represents *all tigers*; but the concept *this tiger* represents only this one. There must be (possible) tigers that resemble this tiger to any extent that you like, and if resemblance is sufficient for representation, you'd think the concept *this tiger* should represent those tigers too. But it doesn't (Fodor 1984, reprinted in his 1990, 33).

The multiplicity of things to which predicates apply and the singularity of causal relations are both motivators for my theory. To have a truth value, an original representation must both predicate some property of an object and specify which object it predicates that property of. The f-referential component of original representation is what accounts for the singularity of representations, and the singularity of the causal relation is what accounts for f-reference. While resemblance, in general, is not a specific relation (as Goodman correctly notes), neither is predication. By requiring both f-reference and f-predication, and providing different theories for each, we are able to account for both the singularity of f-reference (with causation) and the multiplicity of f-predication (with structural preservation). So this objection is avoided.

The final objection that is implicit in Goodman's work is that, at bottom, resemblance is really an intentional notion and so cannot be used as a ground for naturalizing representation. In the context of Goodman's work, he uses the claim that resemblance is an intentional notion to reinforce his claims against realism. Goodman would not make this particular objection against my work since he wouldn't accept the realist, naturalist presuppositions of my project in the first place. It is a serious objection nonetheless, and so we should have a look at it[102].

Goodman discusses options for characterizing pictures as realistic or not. The first, intuitive option that he dismisses is that a picture is realistic if it copies, as best as possible, the item it pictures. This can't be right, he argues, because there are potentially an unlimited number of "ways" that the item is (that is, it has a potentially unlimited number of properties). But we cannot characterize which

[102] That resemblance and similarity are notoriously difficult to characterize is well-known (Lewis 1973 for example simply takes global similarity as fundamental and unanalyzed, then builds his metaphysics on top of that). But the specific application of this well-known fact in an argument against resemblance as mental representation is a distinct point. David Pereplyotchik once pressed me on something like this at a conference.

are the ways that are relevant for accurate copying[103]. After rejecting imitation as the ground for realistic pictures, he concludes that representing is not a matter of imitating, but of classifying or characterizing. This is not a passive activity, Goodman claims, for every object is a member of countless classifications, but "to admit all classifications on equal footing amounts to making no classification at all" (Goodman 1976, 32). Thus, classifying anything (and hence, representing), is ultimately, in some sense or another, a matter of preference: "Classification involves preferment" (1976, 32). Finally to draw this back to resemblance, he argues that, rather than resemblance being a criterion for representational accuracy (or, for a painting's being realistic), resemblance is itself a product of representation[104]. Since representation involves preference in classification (hence, judgment), so does resemblance.

We can get at the source of the objection through Goodman's reasoning, but we can also take a more direct route. Any object instantiates an unlimited number of properties, and shares at least some properties with (perhaps) everything else. Resemblance involves the sharing of properties. If every object shares properties with everything else, and if, for purposes of explaining representation, we want to say it's not the case that every item resembles everything else, then we need to judge which of the properties of an item are the relevant ones for ascriptions of resemblance. This will involve a judgment of salience. And this leads us to our objection: if resemblance ultimately depends on the preference choices that cognitive agents make (Goodman's argument), or if resemblance depends on a salience

[103] "This simple-minded injunction [that, to make a faithful picture, copy the object as it is] baffles me; for the object before me is a man, a swarm of atoms, a complex of cells, a fiddler, a friend, a fool, and much more. If none of these constitute the object as it is, what else might?" (Goodman 1976, 6).

[104] "Resemblance and deceptiveness, far from being constant and independent sources and criteria of representational practice are in some degree products of it", he writes on p. 39. In the footnote (fn. 31) on that page, he writes, "[J]udgments of complex overall resemblance are [not objective]. In the first place, they depend upon the aspects or factors in terms of which the objects in question are compared; and this depends heavily on conceptual and perceptual habit" (Goodman 1976, 39). While not relevant to the point I make in the text, for the curious, Goodman concludes that what constitutes a realistic painting is (unsurprisingly) relative to the standard conventions of the time; it is ultimately a matter of habit.

judgment (direct route), then resemblance is, at bottom, an intentional notion. It therefore cannot ground original representation.

If this argument applied to my theory, it would be a fatal objection because it undercuts the central, naturalistic motivation for my project in the first place. It does not apply. The key to seeing why it doesn't apply is to keep clear the distinction between discovering whether independently specified relational systems bear structural preservation to each other, and stipulating relational systems in such a way that they do (or don't) bear structural preservation to each other. For example, let the domain of $\mathfrak{B}$ include firing rates, and let the greater-firing-rate relation order them. Suppose further that I would like to argue that the firing rate of some particular neuron in somatosensory cortex represents a particular frequency of a stimulus at the fingertip, partially in virtue of the structural preservation of the empirical relational system $\mathfrak{A}$ composed of frequencies, in $\mathfrak{B}$.

My opponent may object as follows: It is insignificant that $\mathfrak{A}$ is structurally preserved in $\mathfrak{B}$. Everything resembles everything else to some degree or another, and structural preservation, while a refined version of resemblance, is resemblance nonetheless. Give me any relational system, including empirical ones, and I'll give you another empirical relational system to which it is isomorphic (give me a finite one, and you've made my task easy). Ultimately what matters here is that you, the cognitive agent, have determined that certain of the properties of frequencies, to the exclusion of others, are salient for the purposes for which you ascribe resemblance (i.e., structural preservation). Your theory is vacuous since resemblance, and hence structural preservation, is trivial. The only way to make your theory non-vacuous is to choose some properties as salient and hence relevant for ascribing resemblance. But that makes it a non-reductive theory.

The confusion in the above objection should, I hope, be apparent. The objector has failed to distinguish my claim *that* $\mathfrak{A}$ is structurally preserved in $\mathfrak{B}$ from the distinct claim that *there exists* some empirical relational system (not necessarily $\mathfrak{A}$) that is structurally preserved in $\mathfrak{B}$. I do not deny that there exist infinitely many empirical relational systems that are structurally preserved in $\mathfrak{B}$. But, to repeat, I also do not claim that structural preservation alone is sufficient to determine representational content. For that, we need (i) teleology to determine that the states of a physiological mechanism or device are representations, (ii) causation to help pare down the infinite number of relational systems that it structurally preserves, (iii) independent considerations for specifying the empirical relational systems (and hence, continuing to pare down the number of relational systems that $\mathfrak{A}$ structurally preserves) and (iv) teleology, again, to select one among the various mappings as the representation function. Perhaps even with all this machinery, further work needs to be done. Presently however, all that I aim to show is that the objection under discussion does not apply to my theory.

There are two different confusions that we must sort out. The first I have discussed above: stipulating relational systems to be isomorphic, and discovering whether relational systems are isomorphic, are distinct endeavors. That a relational system may always be stipulated in such a way that it is structurally preserved in another, is both true and irrelevant. The second confusion is based on a notion that we can quickly disabuse ourselves of. Resemblance and structural preservation are not the same thing. It is, to some extent at least, true that everything resembles everything else, so long as we suppose that resemblance is the sharing of properties. But it is patently not true that everything is isomorphic to everything else, or that everything is structurally preserved in everything else. Structural preservation is a relation that obtains among two relational systems. It is an entirely objective relation that is not dependent on any cognitive agent or her judgments. Further, proving that isomorphism or other kinds of structural preservation obtains is a difficult and non-trivial endeavor.

Thus, structural preservation is an entirely objective matter. Further, it is not true that everything is structurally preserved in everything else. Finally, that there exist a potentially infinite number of relational systems that are structurally preserved in any given relational system implies neither that structural preservation is trivial, nor that structural preservation is not objective, nor that my theory is non-reductive.

6.2 Fodor's *Language of Thought*

In his (1975), Fodor argues that there must be a language of thought; or, human psychological states and processes are constituted by a system of mental representations that exhibit syntactic structure. In chapter 4 he argues that the internal code cannot be comprised of representations that represent or refer in virtue of resemblance. The symbols in the language of thought cannot be imagistic, iconic, or non-discursive (these are, for the most part, synonyms in Fodor's vocabulary). At the time of writing, the predominant view among psychologists was that thoughts refer in virtue of resemblance: "The ur-doctrine in this field is inherited from the British empiricist tradition in philosophy: Thoughts are mental images and they refer to their objects just insofar as (and just by virtue of the fact that) they resemble them" (Fodor 1975, 174). This cannot be right, Fodor argues, for the following reasons.

First, icons cannot have truth conditions. To see this, he asks us to imagine a language much like English, except referring expressions like 'John' and 'green' are replaced with pictures. Icons play the role that words do in English. The problem is that, while in Iconic English icons resemble what they refer to, sentences in Iconic English don't resemble what makes them true. In order to say, for example,

"John is green", we may place the icon that resembles John to the left of the icon that resembles greenness:

> But *that* doesn't look like being green; it doesn't look much like anything. Iconic English provides a construal of the notion of a representational system in which (what corresponds to) *words* are icons, but it provides no construal of the notion of a representational system in which (what corresponds to) *sentences* are (Fodor 1975, 179).

What allows Iconic English to have truth conditions is essentially that it retains the syntactic structure of English. But, Fodor argues, "thoughts are the kinds of things that can be true or false. They are thus the kinds of things that are expressed by *sentences*, not words". Further, while we can at least imagine a representational system that has *words* that refer in virtue of resemblance, "it makes no sense at all to imagine a representational system in which the counterparts of sentences do" (1975, 179).

Let us, however, try to imagine such a system. In trying to have icons that act as sentences (and thus have truth conditions), the problem is that icons have indeterminate content. Suppose, for example, that we wished to have an icon that means the same as the sentence, "John is fat". We may perhaps have a picture of John with a large belly. But then how do we picture "John is tall"? If it is the same picture, then the representational system cannot distinguish among them. Further, John must, of necessity, be pictured as sitting, or standing, or lying down, or in some posture. How can the representational system distinguish among the contents, "John is sitting", "John is fat", "John is tall", etc.? According to Fodor, it cannot. There is simply no way that there is a language in which truth is

defined for icons; "the trouble is *precisely* that icons are insufficiently abstract to be the vehicles of truth" (1975, 180).

Fodor correctly notes that "the kind of thing that can get a truth value is an assignment of some property to some object" (1975, 181), but, he argues, "the trouble with trying to truth-value icons is that they provide no way of [specifying *which* property is being assigned]. Any picture of a thing will, of necessity, display that thing as having indefinitely many properties" (1975, 181). Icons are insufficiently abstract, but the trouble extends even further. They are insufficiently abstract in two ways. First, "they correspond to the same world in too many different ways [as above: a picture of fat John lying down while eating Doritos corresponds in various ways to the same scenario], [but] they also correspond in the same way to too many different worlds" (1975, 181). In this latter case, while a picture of John with a large round belly corresponds to (among other things) John's being fat, it also corresponds to John's being pregnant, in another possible world. Assume, then, that John is both fat and not pregnant. The same reason that we have to say the icon is true provides as much reason to say that it is false. It resembles John's being fat and John's being pregnant equally well[105].

The underlying problem is that icons are insufficiently abstract to have truth conditions. Because of this, they cannot *represent*. However, we've been assuming that icons can *refer* (that is, pictures can act like words in Iconic English, and refer in virtue of resemblance). He argues that this is not correct either: "In natural languages, to put it succinctly, the vehicles of reference are utterances that are taken under (i.e., *intended* to satisfy) descriptions" (Fodor 1975, 182, my emphasis). Sometimes

[105] This is the worry that Wittgenstein (1953) had: a picture which corresponds to a man walking uphill equally resembles a man sliding downhill. Dennett (1969, 136-137) has a related worry about tigers and their stripes. While he takes this concern to show that seeing or imagining has a descriptional (rather than imaging) character, the worry nonetheless seems to apply here. An image of a striped tiger must have some definite number of stripes. But a description of a striped tiger need not. When I imagine a striped tiger, there does not seem to be, or at least need not be, any answer to the question, "how many stripes?". Should we take this to imply that, like descriptions, thoughts abstract away from such detail? If so, this would seem to imply that pictures, icons, or images, cannot underwrite thoughts, and thus mental representations do not refer in virtue of resemblance.

I may entertain an image and take it to satisfy one description, other times I may take the same image to satisfy another. But it is not resemblance that is sufficient for reference; it is *my intention that it satisfy some particular description*. Fodor writes, "Images usually do not *refer* at all. But when they do – as, e.g., in Iconic English – they do so in basically the same way that words and phrases do: viz., by satisfying, *and by being taken to satisfy*, certain descriptions" (Fodor 1975, 183, second emphasis mine).

Fodor here uses the phrase 'being taken to satisfy' as synonymous with 'intended to satisfy' (see previous quote). Thus, this particular usage of 'reference' is not the usage that we are currently interested in. It cannot be the level at which intentional or representational states bottom out, because it crucially depends on the intentions of cognitive agents. At this point in his career Fodor had not yet taken up the question of a naturalistic reduction of representation, so this is not a criticism on that score. However the claim that my mental images only refer in virtue of my intending them to refer is not an objection to my theory because at this point we are just talking past each other. Partially this is because I have not proposed an image theory, and more substantially because the sense of 'reference' that Fodor uses is different than the way I use it (or at least, Fodor's *reference* and my *f-reference* are not the same thing). I'll thus ignore Fodor's criticisms of reference in virtue of resemblance, and focus instead on the more germane objection that icons are insufficiently abstract to have truth conditions, and thus resemblance cannot underwrite *representation*.

To see if and how these objections apply to my theory, we'll need to translate them away from talk of mental images and pictures. I do not propose a mental image or picture theory of representation. I propose a two-part theory of representational content, one of whose parts makes use of structural preservation. As a first step then, we must clarify the distinction between picture theories and structural preservation theory. According to first-order resemblance or picture theories, the individual vehicles of representation share properties with their contents, in the way that Fodor's icons

of John and fat John resemble or share properties with John, or in the way a portrait of President Obama resembles President Obama. According to structural preservation theory, individual vehicles of representation are elements in the domain of a relational system. They represent partially in virtue of being a member of the domain of a relational system whose structure is preserved in another relational system. No claim is made about the structure of the individual items in the domain of the relational system which are the vehicles of representation. While this is a critical difference, the worries about indeterminacy do, to some extent, transfer over into structural preservation theory.

The underlying worry for Fodor is that icons are insufficiently abstract to have truth conditions. This results in two separate objections. First, while it at least makes sense to imagine a language that has icons playing the role of words, where they resemble what they *refer* to, it does not make sense to imagine such a language where icons resemble what they *represent*. That is, icons do not resemble states of affairs that make them true. The second worry is that, insofar as we can imagine such a purportedly impossible system, the content of icons (now considered as playing the role of sentences) is indeterminate. Any given icon resembles a potentially unlimited number of states of affairs. Because of this, icons cannot have truth values: Whatever reason we have for claiming that icon *I* is true, we will always have just as much reason to claim that it is false (recall John's being both fat and not pregnant).

These worries transfer over into structural preservation theory in the following way. One of the major hurdles is the non-uniqueness problem. Any given relational system is structurally preserved in a potentially unlimited number of distinct relational systems. Further, even given a second relational system, the manner in which the first relational system of interest is structurally preserved in the second is overdetermined. That is, there will generally be several, and perhaps a potentially unlimited number of mappings from the domain of one relational system to the domain of the other that preserve structure equally well. Following Swoyer's discussion of conventionalities in measurement theory, I

have dubbed the first aspect of this problem the *systemic component* of the non-uniqueness problem, and the second, the *mapping component* of the non-uniqueness problem.

Consider Fodor's second concern: any icon maps to the same world in too many different ways (does this picture represent John's being fat, or John's being tall?), and further, any icon maps in the same way to too many different worlds (is this a true representation of John's being fat, or a false representation of John's being pregnant?). For our purposes we need not worry about the first-order, picture theory version of this objection. But we do need to worry about the structural preservation version of this objection, and quite clearly these worries correspond to, respectively, the mapping component of the non-uniqueness problem and the systemic component of the non-uniqueness problem. And to these, we have answers.

I've discussed this several times previously so I will keep it brief. We begin with the systemic component. First, teleology allows us to determine which biological devices or mechanisms have states that are representations. Second, independent considerations drawn from the relevant biological sciences help us determine how best to type the states of that mechanism in order to specify a relational system. Third, causal history helps to select among the numerous relational systems that any given physiological relational system might structurally preserve. Fourth, independent typing considerations, generated by the physical sciences relevant to understanding the stimulus relational system of interest, help us determine how best to type this second empirical relational system. All of this is preliminary to investigating if, and if so what kind of, structural preservation obtains. Once we've got independently specified empirical relational systems, we then need to determine the properties of those systems (weak order, total order, etc.) in order to determine whether one system is structurally preserved in the other. If so, *only then* must we tackle the mapping component of the non-uniqueness problem.

Given independently specified empirical relational systems and a determination of which kind of structural preservation obtains, we need to determine which of the numerous mappings is the *representation function*. Recall from our discussion in 5.4.7.1 that, assuming that we can determine all of the above, this will only prove *that* structural preservation exists. It will not tell us *which* of the potentially infinite mappings determines representational content. For this, we return to teleology. I argued in 4.4.1 that Millikan has provided us with the conceptual machinery necessary to rule *in* which of the numerous mapping functions is the representation function, and to rule out all of the rest, by appealing to stabilizing and standardizing teleofunctions. Finally, I reiterate that we must be vigilant in not confusing the ability to stipulate structurally preserved relational systems, which is irrelevant, with discovering whether independently specified relational systems bear structural preservation to each other. Thus, Fodor's concerns from considerations of mental imagery and icons do transfer over into structural preservation theory, but once we have an adequate grasp of how they apply, we can see that my theory has the resources to account for this worry. In addition, this helps to underscore the motivation for the theory in the first place: that structural preservation applies to so many things is what makes it a viable candidate for determining f-predication, and that causation is specific and singular is what makes it a viable candidate for determining f-reference.

Let's return to Fodor's first worry. He argued that icons cannot have truth conditions because "it makes no sense at all to imagine a representational system in which the counterparts of sentences [represent in virtue of resemblance]" (1975, 179). However, Fodor continues in the next paragraph to consider just what he claims he cannot imagine: icons that act as the counterparts of sentences in that they have truth conditions. He then argues that, because of insufficient abstractness, icons cannot have truth conditions. That is the real argument, which I've dealt with above. The "it makes no sense to imagine it" claim is, I think, a rhetorical flourish.

Before we move on, I'd like to note the commonality of each of the above objections. We've so far discussed objections that resemblance is reflexive, symmetric, not sufficient for representation, and not singular. We've discussed the claim that resemblance is at bottom an intentional notion, and we've discussed the claim that icons/non-discursive representations are insufficiently abstract to have truth conditions. Each of these objections is ultimately the same: They are all versions of the non-uniqueness problem. Even the claim that resemblance/isomorphism is an intentional notion is a version of it, since the worry there is that "everything is isomorphic to everything else", or, we must select among an unlimited number of isomorphic relational systems, and doing so presupposes a cognitive agent. At bottom, all of these objections worry over the fact that structural preservation is non-unique. I have, however, provided an analysis of the components of that problem, and provided the tools for dealing with each component. And to repeat one last time, it is the very non-uniqueness of structural preservation that suits it for the work I ask it to do in my theory.

6.3 Structural Preservation and Syntactic Structure

Resemblance theories attempt to explain representation in terms of a sharing of properties between the vehicle of representation and its content. In this respect, they are typically thought of by analogy with images, pictures, maps, scale drawings, diagrams, and so forth. By contrast, the language of thought hypothesis presumes that thoughts are like sentences in that they have syntactic structure. Or, thoughts have a combinatorial semantics, where meaningful parts can be shifted around into other complexes while retaining their meaning, thus allowing for new meaningful complexes. However, the language of thought hypothesis does not include an explanation of the meanings of the constituent parts. It only says *that* they have meaning (however it is determined), and those parts can be

rearranged the way words in sentences can be rearranged. It has been argued (by Fodor, for one) that only if we assume the language of thought hypothesis can we explain such things as mental *processes*. It has been further argued that maplike theories (and hence, resemblance theories) are incompatible with the language of thought hypothesis. Thus, since we need the language of thought, and resemblance theories are incompatible with that hypothesis, it follows that resemblance theories are not viable. Let's have a closer look at this objection.

In the postscript to his (1987) Fodor revisits the motivation for assuming the language of thought hypothesis, and provides three arguments in its favor. Most relevant for us here is his second: only by assuming the language of thought hypothesis can we explain *thinking* (considered as a *process*) (see Fodor 1987, 143-147). In the text he makes heavy use of an example from psycholinguistics. When we understand a sentence, we construct a mental representation of that sentence, which is essentially a parsing tree. This tree specifies the constituent grammatical structure of the heard sentence. In order to explain the mental process of, say, answering a *wh*-question in English (i.e., who did John bite?), the psycholinguistic story goes something like this. The hearer must first construct a mental linguistic parsing tree. Then the answer is generated by, essentially, moving pieces of the parsing tree into something like reverse order. That is, to get from the one representational state of understanding the question to another, which is a formulated answer, the cognitive agent must *move a piece of the parsing tree itself*. Thus, the psycholinguistic theory explicitly quantifies over mental representations that have constituent structure, whose parts can be "moved" about while retaining their meaning in order to construct new, syntactically structured representations. Psycholinguistics is thus committed to the language of thought hypothesis, Fodor argues.

So far this is not an argument against either the picture theory or the structural preservation theory of representation. To get that we need an additional step, which is the claim that the language

of thought hypothesis is incompatible with this kind of theory. Devitt and Sterelny (1999, 138-140) make that argument as follows. First, following Fodor, they argue that only by assuming the language of thought hypothesis do we get an explanation of thinking, and in particular, of the making of inferences. For example, for Oscar to come to believe that Reagan is dangerous, perhaps he inferred in the following way: All Christians are dangerous; Reagan is a Christian; therefore, Reagan is dangerous. The explanation of this process of inference adverts to the form of that argument: All *F*s are *G*s; *a* is *F*; so, *a* is *G*. But that explanation makes use of syntactically structured complexes, and the explanation of the inference-process relies on the ability to move the constituent parts of the complex (*F*, *G*, *a*), while retaining their meaning. Like Fodor, Devitt and Sterelny argue that psychological processes require the language of thought hypothesis. These authors provide an additional step however, by arguing that maplike theories cannot account for psychological processes, and so must be mistaken:

> [I]t is difficult to see how [the view that thoughts are like maps, images, or diagrams] could account for thinking. Formal logic gives us a very good idea of how an inference like Oscar's might proceed if the steps are represented linguistically … Despite the success of [connectionist machines, which apparently use non-discursive representations] with some forms of problem solving, connectionist processes seem rather far from capturing anything like human inference (1999, 139-140).

There are several replies to this objection. First, it is not obvious that this objection applies to my theory, and we can see this in at least two ways. In chapter 1 I attempted to carefully draw a line around my explanandum. That is, I attempted to constrain the questions that I would ask in such a way

that they are both motivated by earlier ways of thinking about intentionality, thought, and thinking, but also posed in a careful enough way that they would be answerable. In so doing, we examined several theoretical endeavors, including folk psychological characterizations of thought and thinking and the classical computational theory of cognition, from which both of the above examples are drawn. We also examined neuroscience and ethology, connectionism, and semantics. What we found was that there is a common core to the various uses of 'representation', and I set out to explain that. I did *not* set out to explain "thought" or "thinking", but rather, *original representation*. Additionally, in 5.1.2, I explicitly repudiated linguistic constraints as adequacy conditions on my theory. I did this because, while language understanding and explicit, linguistically structured inferences are legitimate explanatory targets, they are not my targets. It is not an objection to my theory which sets out to explain *x*, that it does not explain *y* ($y \neq x$).

A distinct objection might go like this: Your theory does not explain *y*, and, though you did not set out to explain *y*, *you should have*. In other words, your explanatory target is not a worthy or interesting one. Because you don't explain what *really* needs explaining, your theory is not very valuable.

This objection objects to my choice of explananda, *not* to the theory I proposed as an explanans. It is thus not an objection to structural preservation theory as a theory of original representation. Since my explanandum is something that is common to several theoretical and commonsense explanatory endeavors, and since it forms the core of intentionality, explanation of which is a more traditional project, I take it that my explanatory target is indeed a worthy one. In section 8.6 I will further address this worry.

The second way to see that Devitt and Sterelny's objection may not apply to my theory is as follows. The claim that "it is hard to see how the view that thoughts are like maps, images or diagrams could account for thinking" is based on a metaphor. While that metaphor does seem applicable to a first-order resemblance or picture theory of representation, it is not clear that it applies to structural preservation theory, which has several elements to it. In order to demonstrate that the objection applies to my theory would take additional argument. Unlike the worries about indeterminacy of content for picture theories that do translate into the non-uniqueness problem for my theory, it is not immediately obvious how this metaphor or its related objection applies to my theory, especially considering my explanatory target.

While I stand by my initial reply that this objection probably doesn't apply to my theory, I understand that this is not a very satisfying reply. So let's assume it does apply to my theory and examine the objection once again. It has two parts: The language of thought hypothesis, and the claim that my theory is inconsistent with the language of thought hypothesis. Both parts must be the case if the argument is sound.

The debate over the language of thought hypothesis is detailed and involved, and this is not an appropriate place to get embroiled in it. However, I will simply present some very brief considerations to show that, at the least, the arguments in favor of the hypothesis are not decisive. It is correct that classical computer science, or GOFAI[106], models some aspects of human inference better than connectionism. It is also true that connectionism models other aspects of human cognitive capacities, such as perceptual discrimination and categorization, better than GOFAI. It doesn't seem that the things more easily modeled/explained by GOFAI are any more or less important or relevant to understanding

[106] Haugeland introduced this term in his (1985, 112) as an acronym for 'Good Old Fashioned Artificial Intelligence', and it has stuck.

the human mind, than the things better modeled by connectionist machines. Both are viable research programs. Further, and importantly, connectionism at least aims for (but does not usually achieve) biological plausibility, whereas GOFAI does not even aim for it. Whatever the specifics, our minds are intimately related to our brains, whose computational elements quite clearly are neurons or sub-neural elements. Without overwhelming evidence to the contrary, the only reasonable position to take is this: However *neurons* compute, that is how *our minds compute*. Even if GOFAI can model inference better than connectionism, that would seem to be an artifact of the paucity of our current connectionist machines' computing power, due to technical constraints. It is entirely relevant to keep in mind that the brain has over 10^{10} neurons and at least 10^{13} synapses. No connectionist machine is anywhere near that level of complexity.

What really matters is how our brains compute, and if connectionist models have a better way of getting at that, then that is a significant point in their favor. Additionally, this debate is not exclusively connectionism vs. GOFAI. There are further, biologically plausible models of computation in addition to connectionist models (Rieke 1997; Gerstner and Kistler 2002; Eliasmith and Anderson 2003; Koch 1999), which, like GOFAI, ultimately aim to explain human cognitive capacities (among other things). Anyway, all of this is only to say that there are several viable research programs under way, and Fodor's "in principle that couldn't work" objections don't cut it. We have reason to continue all of these research programs, and (perhaps) eventually seek a synthesis of them.

However, in addition to assuming that the objection applies to my theory, let us also assume that, as Fodor says, the language of thought hypothesis is "the only game in town". We have no other options and must make do with it. Even granting these two assumptions, the objection still doesn't work, because my theory is compatible with the language of thought hypothesis.

For the language of thought hypothesis to be true, the system of mental representations with which humans perform their cognitive computations must be syntactically structured in a quasi-linguistic way; the parts must be both meaningful and transportable. This does not say anything about how the parts get their meanings. Let's assume the following story. My theory of original representation, which is a theory of the lowest-level representations, is true. The basic representations have representational content in virtue of both structural preservation and causal history. However, the basic, lowest-level representations are *not* the subsentential parts described by the language of thought hypothesis. The subsentential parts of sentence-like thoughts, those parts which must be both meaningful and transportable, and which make possible the syntactic structure of thoughts and hence the process of thinking, are *derived* representations. They have their representational content in virtue of their bearing some suitable relation to other states which are representations (i.e., the basic representations that my theory explains). The language of thought hypothesis does not need, and does not state, that its subsentential parts are the most basic representational states in the universe, and it does not provide any theory of how those subsentential states get their meaning. It is thus compatible with my theory.

Certainly there is a lacuna in this story: What is this suitable relation that accounts for the representational content or status of non-basic representations? I do not know, nor, for the purposes of replying to this objection, need I supply an answer. All that matters is that the language of thought hypothesis does not imply that my theory is false. Thus, even with the two assumptions made above, the objection that my view cannot account for psychological processes is illegitimate. My view has just as many resources available to it for accounting for psychological processes as, say, Fodor's asymmetric dependence theory.

It is worth noting, however, that no one has a story about the "suitable relation" in question. Dretske, Fodor, and many others have attempted to provide a theory of representation for the most basic kinds of representations. They each claim, as I have, that their goal is to explain basic representation, and then to extend the theory from there. But the claim, for example, that language "expresses" thought, and this is what accounts for the intentional content or meaning of language tokens, is merely a label for the solution. It is not the solution itself. Additionally, even the language of thought hypothesis suffers from this same gap. How, precisely, does the meaning of the whole sentence-like complex derive from the meanings of its parts? At a certain level, a description of syntactic structure is only a description *that*, not an explanation of *how* or *why*. It is not an explanation of, say, how the parts retain their meaning, or what the actual "moving about" of parts is constituted by (since this is only a metaphor). It is also not an explanation of how the parts form together into an aggregate in such a way that their new "shape" constitutes a new semantically evaluable complex. Answering each of the above questions with, "in the same way that a sentence does", is not explanatory since the explanation of the meaning of the sentence derives from the explanation of the meaning of the thought it expresses. Thus, while I need not answer the question of what the suitable relation under consideration amounts to in order to reply to the present objection, it is worth noting that no one else has an answer for that question either.

There is one final point that I would mention in connection with the language of thought hypothesis and its relation to a theory of representation. The intelligibility of the language of thought hypothesis depends on there being a distinction between quasi-linguistic, or discursive, representations, and non-discursive representations. Otherwise, we might as well call it the "system of representations hypothesis". But wherein does this distinction lie?

Fodor has the following three things to say on this. First, he doesn't know how to characterize the distinction. Second, under close scrutiny the distinction may collapse. Third, it doesn't make any sense to even talk about cognitive processes being carried out in a medium of non-discursive representations[107]. If the *language* of thought hypothesis is a substantive claim over and above representationalism, which is the claim that cognitive processes are representational, then it stands to reason that there is another side of the distinction, and hence, it is at least intelligible that there is a non-linguistic medium of representations. Without that other side, there is nothing his claim is distinct from and hence, there is no distinction. If there is no distinction, then perhaps the claim that there is a language of thought is just a tendentious way of saying that human cognitive capacities depend on a representational system that has certain properties (such as compositionality, productivity, and systematicity). But languages are not the only systems that have these properties; Millikan's (1984, 2004) system of structured representations is not a "language" of thought in the sense that Fodor advocates, but it does allow for productivity, systematicity, and compositionality (see my 4.2 for discussion). It seems that, perhaps, talk of "imagistic", "non-discursive", and "discursive" representations is metaphorical and adds no substance to the debate, either over the language of thought hypothesis or the nature of representation. At the very least, we should be clear about what these terms amount to if we want to make use of them. If Fodor readily admits that he not only doesn't know how to characterize the difference but also agrees that, upon examination, there may be no

[107] First: "Bruner, like most other writers who have concerned themselves with the nature of symbolism, assumes that there is a principled distinction between 'iconic' symbols (viz., images) and 'discursive' symbols (viz., words or descriptions). I'm inclined to consider that reasonable though, notoriously, it is extremely difficult to say what the distinction consists in" (Fodor 1975, 175, fn. 10). Second: "Such cases suggest how rough-and-ready the unanalyzed contrast between images and descriptions really is. For present purposes I am using the materials at hand, but serious work in this area would require sharpening (*and perhaps ultimately abandoning*) the framework of distinctions I have been assuming" (1975, 190, fn. 24, my emphasis). Third: "I am, in fact, strongly inclined to doubt the very *intelligibility* of the suggestion that there is a stage at which cognitive processes are carried out in a medium which is fundamentally nondiscursive" (1975, 177, Fodor's emphasis).

difference, then it becomes difficult to see the import of any of his arguments against non-discursive representations.

None of this should be construed as a serious analysis of, or argument against, the language of thought hypothesis. Rather, I only seek to sow doubt against the claim that "the language of thought hypothesis is the only game in town". It is not, and in fact it may not even be a substantial hypothesis to begin with. Hence, even if there is a principled distinction in just the way Fodor needs there to be, and even if the "psychological processes" objection applies to my theory, and even if we must make do with the language of thought hypothesis, still, the objection doesn't work. At the end of the day, if my objector gets past all of my "even-ifs", my theory is still compatible with the language of thought hypothesis, and thus can explain psychological processes as well as anyone else's.

Let's conclude this section with a classic objection: As Berkeley has argued, images cannot represent abstract ideas. "Nothing could look like, say, virtue since virtue doesn't itself look like anything. I take it that the arguments against the identification of abstract ideas with images are sufficiently familiar from Berkeley" (Fodor 1975, 174-175, fn. 7). I imagine that by now my reader can anticipate what I would say to this argument. Rather than saying, "that's not what I was trying to explain in the first place" once again (even though, that's not what I was trying to explain in the first place), let's take a "big picture" look, at representation, thought, intentionality, and our strategy for making sense of them in the physical world.

I began this dissertation by mentioning that my project is similar to but distinct from the project of naturalizing intentionality. The motivation for my choice of explanandum is partially that I wanted to constrain the questions asked in such a way that they would at least be answerable. The way I see it, this allows for progress. Perhaps my answers don't work, and perhaps my explanandum is too "simple"

or is not “interesting” enough. But my question *is* answerable, and I have a theory that explains the most rudimentary kind of aboutness, with the capacity for error, in biological systems. If it pushes the literature forward, great; if not, at least we will have discovered a route *not* to take. However, a second motivation, besides that I wanted to be able to answer my own question, is this.

Dretske, Fodor, in fact almost everyone writing on intentionality, and certainly the philosopher who objects, “your theory doesn’t explain how we represent virtue therefore none of it works”, are all trying to run before we can crawl. Everyone wants a theory that explains aboutness, error, the generation of referential opacity, the representation of abstract concepts, the nature of thought and how representations underwrite thinking processes, the relation between thought and language, and perhaps as well, an account of introspection and an explanation of why meaning rationalism at least seems to be true. And in trying to provide a theory of all of that, we end up biting off more than we can chew. Rather than continuing to tread these well-worn paths, what I have proposed here is to take the tools provided us by previous work on intentionality and the foundations of measurement, and apply them to a different, and closer, target. Rather than trying to crawl, walk, run, and fly all at the same time, I’ve tried to bring it down to earth and only attempt to crawl. If my theory does not explain how also to run and fly, that is because I deliberately bracketed that off right from the start. We may discover that my tactic of bracketing off these further questions in order to focus on those I have asked is what allows for an answer to the more fundamental, “how-to-crawl” questions. Additionally, if deliberately ignoring the rooftop is what allows for construction of the foundation, it may also turn out that, once the foundation is laid, we may begin to be able to see the rooftops as well. Anyway I’ve run

out of metaphors. My reply to Berkeley's objection is: that is far outside the scope of what I set out to explain[108].

Resemblance and, by extension, isomorphism, are widely considered as discredited resources for understanding representation. I hope that I have made it clear that the majority of the worries center around the non-uniqueness problem for structural preservation theories. I have analyzed that problem into its components and provided the resources handling both components. While my theory does have the resources to handle the non-uniqueness problem, I hope that I have also made it clear that it is the non-uniqueness of structural preservation that suits it for explaining f-predication. Thus, the traditional worry about non-uniqueness is not a liability, but an asset for my theory.

6.4 The Causal Chain Problem

In almost any causal exchange, and especially those involving physiological changes as a result of changes in ambient energy, there are intermediate causal steps. There is a "chain" of causation. For example, when I have a visual perception of a red apple, electromagnetic energy reflects off the skin of the apple, at various wavelengths. Photons arrive at the transducer cells in the retina, causing a chain of biochemical events which result in a change in the voltage across the membranes of the cells. This results in further changes in other cells, waves of depolarization transfer down axons, neurotransmitter is released in various places, and so forth. The causal chain that begins with the apple and ends with a

[108] I guess I just can't resist a slightly more substantial reply. Set membership is essentially unconstrained, therefore relational system membership is essentially unconstrained. There is no reason in principle why a structural preservation theory could not explain the representation of abstracta like virtue, since virtue, goodness, number, etc., could each be a member of some relational system. Recall the difference between picture theories and structural preservation theory: it is irrelevant that virtue doesn't "look like anything". What, exactly, does a relational system look like?

mental perception of the apple has a very large number of intermediate steps. The causal chain problem is: Which "link" in the causal chain is the appropriate place to confer content? Does my perception of the apple represent changes in my primary visual cortex, or my thalamus or optic nerve, or retina, or what? Intuitively we want to say that my perception represents the apple and not any of my brain states, but on what theoretical grounds may we say this?

To answer this, first, let us drop the "perception of an apple" talk. I used that language only to make clear the idea of the causal chain problem. I am only interested in the most fundamental kind of representation in the universe, and talk of perception will only serve to obfuscate that by importing intuitions about consciousness and introspection. Additionally, recall our discussion from 5.2.2 on the dangers of assuming that one representational system (i.e., English) can adequately capture the representational content of another representational system (i.e., the basic, non-derived neural representations of interest here). We should not assume that there is a straightforward translation from the basic representations to English. With that said, the causal chain problem does transfer over to my theory, because there is still a causal chain with intermediate links, from content to representation (or representation to content).

While there will typically be intermediate links to the causal chain of interest, this does not guarantee that there will be structurally preserved relational systems at every link as well. As we learned above, the claim that "everything is isomorphic to everything else" is false, and structural preservation, though it is a broad-ranging category compared to isomorphism, is still a fairly narrowly restricted relation. So, first, we have no reason to believe that it will obtain at every link in the chain. However we may also safely assume that, at some links at least, it does obtain. In that case, teleology arbitrates. Biological devices have proper functions in virtue of their being tokens of a type of device, previous tokens of which have proven to have survival or reproductive value. The adaptive value of

representational devices is not in their ability to get the organism into states that are adapted or shaped to other states of itself (not usually, I suppose), but is in their ability to get the organism into states that are shaped to conditions in the world.

The causal chain problem is really another instance of the systemic component of the non-uniqueness problem. What it asks is, which relational system has members of its domain that are the contents of the representations? By making use of teleology to help with this question, I've introduced yet another role for teleology, as well as another way of helping to solve the systemic component of the non-uniqueness problem. Teleology, causal etiology, and independent typing considerations all work together to answer that question. I readily admit that I have only waved my hand at a solution to the causal chain problem, but I don't take that as a strike against my theory since the objection has only been made in the abstract. It suffices to note that it is not an insoluble problem. More specific articulations of this solution will be given in the context of a consideration of actual cases, especially in 7.5.2.3.

By way of conclusion for this chapter, I would emphasize that what I propose here is not intended to be a completed theory. Rather, it is a theoretical *framework*, whose details bear filling in, and which is open to revision based on further conceptual work as well as empirical data. I do however contend that it is conceptually coherent, consistent with a larger body of physical theory, and satisfies our explanatory adequacy conditions. It is naturalistic in the sense that it does not make illicit use of intentional or semantic concepts. It shows how representation fits into the natural world, and it is consistent with a wider body of physical theory. It explains what it is to be a representation, as well as representational content. It has an explanation of aboutness and error. It is, at least in principle, implementable in the nervous system, and in the next chapter we will remove the "in principle" clause. We will examine the causal efficacy requirement in chapter 8.

Given the work completed thus far, we have a provisional, workable framework for thinking about naturalistic, original representation. In the final two chapters, we'll take a completely different approach to understanding representation. We're going to look at single-cell intracortical recordings of neural activity, taken from an awake Macaque monkey performing a cognitive task. The empirical research will provide evidence for as well as illustration, clarification, and refinement of, structural preservation theory.

Chapter 7: The Implementation of Representation I - Evidence and Structural Preservation Theory

7.0 Introduction

Any responsible approach to understanding the nature of mind in the physical world will take into account the physical systems that implement the mind or its states. To the best of our knowledge, only the active nervous systems of living biological organisms do this. Perhaps it is possible, in principle, that space aliens and futuristic robots have minds, but that is immaterial to the present discussion. A working brain is the seat of a mind, and so to understand the mind, we should look to the brain.

Beneath this seemingly banal observation lie deep conceptual difficulties. Specifically, how will looking at the brain help to understand the mind? What are we looking for? How will we know that we've found it? In what sense are my questions empirical at all? In this and the following chapter I will address these and related questions, in an attempt to demonstrate a convergence of conceptual and empirical support for the structural preservation theory of representation.

The structure of this chapter is as follows. I begin with some background assumptions and two hypotheses that structure further discussion, and then clarify the logic of the claims I make in these final two chapters. Then in 7.3 I present an overview of some recent experimental work involving electrophysiological recordings taken from an awake Macaque monkey performing a sensory discrimination task. In 7.4 I defend the claim that the brain states that I mention are representations and do so without presupposing structural preservation theory. In 7.5 I apply structural preservation

theory to the findings, and argue that the empirical work provides evidence for structural preservation theory.

7.1 Background: Assumptions and Hypotheses

I assume the following: (i) scientific realism[109], (ii) confirmational holism[110], (iii) the legitimacy of inference to the best explanation[111]. Additionally, I take the relation of science to philosophy to be roughly as Quine has described it: There is no first philosophy, but philosophy and science are continuous and on a par. I will not defend these assumptions here; however, they do operate in the background and sometimes the foreground of these final two chapters, so I shall make them explicit.

The doctrines of multiple realizability and the autonomy of psychology might be considered deterrents to the implementation component of my project. Fodor has argued in the preface to his (1975) that money, for example, can be physically instantiated in an indefinite number of ways, hence the study of economics is autonomous from physics and chemistry. Similarly, he argues, minds can be instantiated in an indefinite number of ways, so that psychology is autonomous with respect to neuroscience. Thus, if representation can be instantiated in many different ways, then studying its neural implementation does not help to understand representation, or the mind, itself.

[109] I accept scientific realism in Devitt's (1997, 24) sense: Tokens of both the observable and unobservable scientific types exist and do so mind-independently. As it applies here, I assume that representations exist, and that their nature is to be *discovered*, not stipulated.

[110] I take confirmational holism in the standard sense of the Duhem-Quine thesis (Quine 1951). Hypotheses can only be tested from within the context of a set of auxiliary hypotheses, but not in isolation from their background conceptual framework. In other words, conceptual frameworks are tested as a unit each time an experiment is performed, but, given our degree of confidence with respect to various background assumptions, we can consider one or several hypotheses to be those that are under test, even though, in principle at least, all hypotheses are open to revision.

[111] Not everyone accepts abduction. Van Fraasen (1980, 1985), for example, does not.

I do not propose to study the mind by studying collapses of wave functions or superstring theory. I propose to study the mind partially by studying the brain. A more relevant analogy would be between economics and social psychology, not economics and physics. The appeal to the multiple realizability of money by entities from physics is a gross exaggeration of a sound point, which is that different sciences have their own proprietary language. But that basic point does not support Fodor's radical, dualist-leaning methodological conclusion. The autonomy of the special sciences is a matter of degree: The sciences intersect and cross-cut each other constantly. Even the science of economics draws on phenomena from presumably more "basic" sciences, by explaining market changes in terms of fears or hopes of consumers, or at the very least, of consumer *behavior*. The special sciences can be, for some practical purposes, considered to be relatively autonomous, but the universe doesn't have to cut itself up quite so neatly just so that biologists can ignore chemists and psychologists can ignore biologists. Each science has its own field of study, but that study is both constrained and enriched by some of the findings in closely related fields of study.

The force of the standard arguments for multiple realizability is similarly exaggerated. While it can certainly be *imagined* that any number of physical substrates can instantiate the functional architecture of the mind, this does not imply that in our physical universe, they *actually do*, or *even can*. Nervous systems are the only known physical systems that instantiate minds, and this makes studying them with an eye to the psychological states that they implement crucially relevant to any responsible naturalistic approach to the mind. And this follows *even if multiple realizability is true*. Thus, neither doctrine should be considered a deterrent to my project here.

I advocate two distinct hypotheses. The first is the *representation hypothesis*. I will describe an experimental paradigm in which a Macaque monkey is trained to perform a vibrotactile discrimination task. The monkeys accurately discriminate the frequency of two vibrating tactile stimuli at levels far

greater than chance, and for correct discriminations they are rewarded with juice. The task is designed so that the monkeys form a representation of the initial stimulus, hold it in working memory, then form a representation of the second stimulus, compare the two and form a decision on which is greater, then output a motor plan to behaviorally signal its choice. The representation hypothesis is simply *that* the monkey has representational states. It is essentially the brief story just told, involving working memory, a decision process, sensory representations and motor plans.

It is important to emphasize that this is not to say that Dennett's intentional stance ought to be taken with regard to the animals. It is not to say that there should be a presumption of minimal rationality, nor that descriptions of these states generate an intensional context. The representation hypothesis is silent on whether the representational states of these animals are properly described as the folk psychological propositional attitudes, or are amenable to intentional description of the sort that our folk psychology would have us apply to each other.

Recall from 1.2.1, 1.2.2, and 1.3.1 that *representation* is a theoretical construct common to at least five explanatory approaches, one of which is folk psychology, while *intentionality* involves aboutness and other properties associated with the propositional attitudes of folk psychology (even though 'intentionality' is a technical term and 'representation' has ordinary currency). When we examine the different ways that 'representation' is used in each of these theoretical endeavors, we find a common core, which involves aboutness, the capacity for error, and causal efficacy. Additionally, when we look at the ordinary language use of 'representation', we find two recurring themes: Representations are proxies or surrogates, or, representations are things that point to, or are about, other things.

While misguided in some of his arguments, Ramsey (2007) is right that, for a theoretical posit to be a *representational* posit, it should accord to some unspecified degree with the ordinary language use of 'representation'. I argued in 1.3.1 that the representations of cognitive science and those of neuroscience both provisionally deserve the name, even though they differ from each other.

In the experimental paradigm under consideration, I will argue that the best explanation of the monkeys' behavior adverts to representational states. Notions like working memory and sensory representations are part and parcel of cognitive psychology and cognitive science. These concepts involve pointing to or surrogacy, so if the monkey has working memory or sensory representations, this justifies saying that the monkey has *representations*. Since there is a common core to the various theoretical concepts, if a state is representational in cognitive psychology's sense, then it is representational in the core sense that I've outlined.

I will call the second hypothesis the *vehicle hypothesis*. This is the hypothesis that the particular brain states to be discussed shortly are vehicles of representation. Notice that this is distinct from the representation hypothesis: The representation hypothesis does not tell us which states of the animal's nervous system are the vehicles of representation, if any. (Dualism is consistent with the representation hypothesis.) While the representation hypothesis explains the animal's behavior, the vehicle hypothesis explains certain empirical findings to be discussed shortly, involving correlations between brain states, behavioral outputs, and peripheral energy.

While both the representation and vehicle hypotheses advert to representational states, neither is an explanation *of representation*. Rather, they each use the reasonably well-understood yet importantly vague concept of representation in their respective explanations. The structural

preservation theory of representation, by contrast, explains representation by explaining what it is to be a representation as well as what determines representational content and error.

7.2 The Relation of Theory to Evidence and the Dual-Approach Strategy

I contend that empirical findings from neuroscience can be fruitfully brought to bear on the traditional philosophical question addressed in this dissertation. Further, we can find evidence for or against the structural preservation theory of representation (henceforth *SPT*). However, the relation between theory and evidence is two-fold, and can be described by two fundamental claims.

Claim 1: If we find representations in the brain, then they will have the properties attributed to them by SPT.

Claim 2: SPT is the best theory of representation. Therefore, if SPT determines that *X* is a representation, then we should conclude that *X* is a representation, *on the grounds that SPT implies this*.

From the perspective of claim 2, any empirical work I appeal to can illustrate, but not confirm, SPT. SPT is not amenable to empirical test from this perspective. However, appealing to empirical research on the brain is still a conceptually useful exercise. First, one of the adequacy conditions on a theory of representation is that it be implementable in the brain, not only "in principle", but that we have some reasonably specific suggestions on how, if SPT were true, the brain would implement it. To

satisfy this condition, we need to look at the brain and show how representations would be implemented, if SPT were true. Second, there can be no better illustration of the theory than an empirically real example, rather than something contrived. Third, appealing to neuroscience allows for the back-and-forth, conceptual-empirical interplay which I first described in 1.3.4.

With the conception of representation defined by SPT in hand, we approach the brain, and see if we can find states that, according to SPT, are representations. We look at those states provisionally identified as representations, analyze them and the relations between those states and what they purportedly represent, and then sharpen our conception of representation. This allows for a back-and-forth, incremental approach to understanding representation, which takes into account the conceptual/empirical nature of the question.

Claim 1 is distinct. It has the standard form of an inference to the best explanation: If SPT is true, then if we find representations in the brain, we would expect them to have the properties attributed to them by SPT. That is, we should expect representations to be related to their contents by causal history and structural preservation. If we do find this, then we can say that the best explanation for this finding is that SPT is true. Notice that claim 2 allows something that claim 1 does not: From the perspective of claim 2 we assume SPT, but from the perspective of claim 1 we may not.

To make good on claim 1, we need an independent way to identify a thing as a representation with content *C*. I expect this to be controversial: How can we identify a thing as a representation with content *C*, without presupposing a theory? If we identify *R* as a representation, then we have already assumed a theory of representation, and thus, we are now working from the perspective of claim 2, and therefore cannot provide empirical confirmation of the assumed theory. There is a way around this worry, however, which can be seen with an analogy.

We can identify water as such without assuming a theory of water. By perceiving its superficial properties, such as its being colorless, odorless, and found in our lakes and streams, we can identify a thing as water. However, to get a better understanding of the nature of water, of what water really *is*, we need the tools of modern chemistry. It is only once we have those tools at hand that we are able to understand that water is constituted by H_2O molecules.

In the case at hand, the situation is more complicated, because it is not clear what the "superficial properties" of representations are, by which we can identify them as such. But there is an important distinction here: First, there is the claim that I have *general* diagnostic criteria, or an algorithm by which we can take any particular thing, apply the diagnostic algorithm, and have an answer as to whether or not it is a representation and if so, what its content is. Second, there is the claim that, in some *particular* case, I can provide reasons to believe that that thing is a representation with content *C*. I do not make the former claim, but I do make the latter. I defend the latter claim by defending the vehicle hypothesis.

Suppose I claim that *R* is a representation with content *C*, on the grounds that *R* is *G*, where *G* involves some reasons independent of SPT. Note that *G* simply is the reasons that I will presently give in support of the vehicle hypothesis, in 7.4. Suppose further that my objector denies this: No, you are not justified in claiming that *R* has content *C*, or, no, *R* is not a representation. Notice the course of the dialectic. All that my objector disagrees with is whether or not I have identified a representation. My objector has *not* disagreed with the more general claim that SPT (or any other theory of representation) may enjoy empirical support. To make good on that claim, my objector would need to establish that it is *impossible* to identify a representation as such without presupposing a substantive theory of representation. But how could someone establish such a claim? Especially in light of the fact that we

can identify water and fish without a theory of chemistry or evolution, the objector's impossibility claim becomes very hard to swallow.

Here's another way of describing claim 1. I assume that there are such things as representations, although I don't know what they are in any deep sense, nor do I know how they cohere with the rest of our physical theory. I also assume that, whatever representations are, it is possible to identify them as such even though I don't have a deep theory of what they are, in the same way that I can identify water as such without having to know that water is H_2O. Thus, we identify states as representations, a claim which I'll support in 7.4, without assuming any theory about the nature of representation, in the same way that we identify water without a theory of chemistry. Then we see what SPT says about those states. Does SPT claim that those states are representations? Does it claim that those states have the same content that the "pre-theoretic identification" [112] does? If yes, then we have confirmation. If no, then we have disconfirmation.

There are three different ways of getting at the concept of representation. First (in no particular order), there is a theory of representation, or SPT. This is a theoretical conception of what representation *is*, of the nature of representation. It is analogous to the modern chemical theory of the nature of water. Second, we will identify representations as such without presupposing a theory of representation, in 7.4 and after. This is analogous to perceiving the superficial properties of water and identifying water as water without needing a theory of chemistry to do so. A relevant dis-analogy is that, while being an odorless, colorless liquid in our lakes and streams is a reasonably general diagnostic indicator by which we can identify stuff as water, my arguments for the vehicle hypothesis are specific to this case.

[112] I will use the phrase 'pre-theoretic' frequently in the following discussion. By it I mean, "prior to a substantive theory about the nature of representation", not "prior to *all* theory".

Third, we have the preliminary identification of my explanandum in terms of aboutness, the capacity for error, and causal efficacy, which was established in chapter 1 as a result of analysis of the variations in the concept of representation across theoretical and commonsense domains. This does not share an analogous counterpart to the water example. The role of this preliminary identification is as follows. First, it establishes an explanandum; it provides something like a conceptual target, or a conceptual way of identifying what needs explaining. Second, this preliminary identification in part helps to establish a usage of 'representation', which accords to some degree with common use as well as with the different theoretical uses.

Thus, we actually have *two* preliminary ways of identifying representations in the absence of a substantive theory about the nature of representation. We have the preliminary identification of the explanandum, which provides something like a conceptual identification of what needs explaining. I will also propose a second method of identifying at least some physical vehicles of representation as such, in a way analogous to the identification of water via its superficial properties, via the vehicle hypothesis.

7.3 The Neurobiological Mechanisms of Vibrotactile Discrimination

In this section I briefly review an experimental paradigm in which trained Macaque monkeys (henceforth simply *monkeys*) discriminate two vibrating tactile stimuli, in what's known as the *flutter range* of 5-50 Hz, to the fingertips. The monkeys press one of two buttons to signal that either the first

or the second stimulus was of a higher frequency. While the monkey performs this task, microelectrodes implanted in its brain record the activity of single cells[113].

7.3.1 The Implicit Theory and its Critique

There is an implicit theory of representation operating in the background of this literature. It is untenable, and in order to avoid any confusion or guilt by association on my part, I begin with some brief remarks on that. The primary tools used to discover neural "codes" are regression analyses and information theory[114]. The concept underlying both of these theoretical tools is that, if the regression or information analysis finds some relationship between two quantities, then it is very likely that a reliable covariation exists between them. Additionally, standardized measures of neural behavior are compared with animal behavior. The underlying assumption here is that, if (say) firing rate in primary somatosensory cortex (or *S1*) is a representation of stimulus frequency, then the neurons downstream of S1 will *use* that coding mechanism in their computations, and eventually some measurable effect, a result of that use, will show up at the behavioral level.

The implicit theory is something like a combination covariation/use theory. That is, it is in virtue of (i) the correlation between firing rate and stimulus frequency, as well as (ii) the fact that neurons downstream of the particular brain states at issue appear to use those states in further computation and control of behavior, that makes those states representations, and determines their content.

This combination of covariation and use is very much like what Bechtel (2001) has proposed. He proposes what he calls a "minimal notion of representation, wherein a representation is an information-

[113] In this section I provide only the briefest review necessary to make my arguments. For a more complete literature review, see Appendix B.

[114] I discuss regression analysis and information theory more fully in Appendix B. Recall that I have argued that mutual information can be legitimately used as a fallible epistemic guide to the existence of an underlying, objective regularity, even though mutual information is itself not an objective quantity.

bearing state or event which stands in for what it represents and enables the system in which it operates to utilize this information in coordinating its behavior" (Bechtel 2001, 347). He also argues that this is in fact the theory that appears to be implicit in much neuroscientific research, including Hubel and Weisel's landmark work on the visual system (Hubel and Wiesel 1962, 1968) and their discovery of "edge detectors". Bechtel is spot-on with his analysis that this is *in fact* the theory that neuroscientists appear to be assuming. Of course, it does not follow from this that that is what they *ought* to be assuming.

I've made these critiques previously and in detail, so here I will be brief. First, we'll need to be clearer about covariation. If Dretske's (1981) informational content, Millikan's (2004) local information, or Shannon's (Shannon and Weaver 1949) mutual information is appealed to, then the analysis of representation is non-reductive. If it is simply causal covariation, then we get the disjunction problem and no possibility of error. If by 'information', we mean something like the colloquial concept, then we have a non-reductive and non-explanatory account. Bechtel explicitly cites Dretske (1981), so for exegetical purposes at least, we know where Bechtel stands, and thus why he is wrong in that stance. Second, Bechtel's notion of use may diverge from the implicit concept of use assumed in the work described here. Bechtel cites Millikan (1984), and makes clear that he is focused on teleofunction, not actual use. Bechtel seeks to use teleology to make room for error: a state represents only that which it has the function of covarying with (Bechtel 2001, 336). On the other hand, the concept of use implicit in the empirical literature to be discussed is something more along the lines of *actual* use: because there is or is not a correlation between behavior and some pattern of neural activity, we are asked to reach conclusions about which pattern of activity is or is not a representation.

Regardless of whether we choose teleology or actual use, we'll ultimately reach the same two problems. First, the informational/covariation component of the analysis doesn't work for the reasons

briefly mentioned above. Second, to get either truth or error, we need structure. The distinction between what a representation is about or refers to, and what it says or predicates of that thing, is absolutely crucial. This is what makes possible both truth and error and hence, content. The failure to recognize this is perhaps *the* fundamental flaw in conceptual work on representation until now. So neither the implicit theory nor Bechtel's philosophical reconstruction of it are tenable theories of representation.

This does *not* imply that the states I discuss below are not representations: I've given an argument against a theoretical conception of what representation is. This is independent of the representation and vehicle hypotheses, which do not make any claims in that regard.

7.3.2 Review of Empirical Literature

The basic, classical task (LaMotte and Mountcastle 1975; Mountcastle, Steinmetz, and Romo 1990) is as follows. A seated monkey has its left hand secured, palm up. A stimulator tip is lowered, indenting the skin of one of the monkey's fingertips; it is not vibrating at this point. The monkey then presses a key with its free right hand, and holds the key down. The stimulator then produces a sinusoidal vibration, between 5 and 50 Hz, to the left hand fingertip (this is the *base stimulus*, or f_1 for first frequency), followed by a delay period (or *interstimulus interval*), followed again by a second stimulation (the *comparison* or f_2), also between 5 and 50 Hz. At the offset of the comparison stimulus, the monkey releases the key with its right hand, and signals its choice on which frequency was faster by pressing one of two push buttons located at eye level. The monkey is rewarded with a drop of juice for correct discrimination.

A schematic of the neural events that occur during this task is as follows. Rapidly adapting, superficially located mechanoreceptors in the finger known as *Meissner's corpuscles* transduce the

mechanical energy into action potentials, which travel up the spinal cord, through the thalamus, into S1, and thence to the secondary somatosensory cortex, or S2 (Gardner and Kandel 2000; Gardner, Martin, and Jessell 2000; Vallbo 1995). The outgoing signal from S2 then gets widely distributed, to at least the prefrontal cortex (PFC), the ventral premotor cortex (VPC), and medial premotor cortex (MPC); PFC and VPC both appear to be serially connected to MPC. Then MPC transmits activity to the primary motor cortex (M1), whose activity ultimately results in the monkey's button-pressing behavior signaling its choice (Romo, DeLafuente, and Hernandez 2004). These cortical areas are typically associated with cognitive activities in the following way. Primary and secondary sensory areas are involved in sensory processing. PFC is widely implicated in short-term or working memory processes, and MPC/VPC are considered to be pre-motor areas, which begin the transformation of signals from sensory and memory processes into motor plans. Primary motor areas are associated with the implementation of generalized motor plans, which then get refined into more specific muscle commands, taking into account various feedback mechanisms and so forth by the basal ganglia, cerebellum, and spinal cord.

The neural activity that occurs during the presentation of the stimulus is as follows. In the periphery, neural firing is phase-locked to the stimulus, where the neuron fires a spike or burst of spikes for each amplitude peak of the sinusoidal stimulus (Mountcastle et al. 1969; Mountcastle, Steinmetz, and Romo 1990; Salinas et al. 2000). Traveling into the cortex, there appear to be two subpopulations in S1. In the first, subpopulation-1[115], neural activity is no longer phase-locked to the stimulus, but the temporal structure of neural firing correlates with the stimulus frequency, in the following way. Periodicity is the property of exhibiting regular, repeating characteristics. Using a Fourier decomposition of the firing pattern, it is possible to deconstruct the function describing that pattern into its component sine and cosine functions, as well as determine their "power", or, determine which of them contributes

[115] I invented this terminology, so it won't be found in the literature.

most to the original function. In subpopulation-1 of S1, the power spectrum frequency at peak (*PSFP*), which is the frequency that contributes most to the firing pattern, usually matches the frequency of the tactile stimulus (Hernandez, Zainos, and Romo 2000; Salinas et al. 2000). In subpopulation-2 of S1, the firing pattern becomes less periodic, and is no longer matched to the frequency of the stimulus. However, the aperiodic firing pattern now correlates with stimulus frequency in terms of its rate, approximating a monotonic function of rate (Salinas et al. 2000).

In S2 and beyond, the rate correlation remains prominent, and the temporal, periodicity-based or phase-locked correlation, is no longer evident. An important difference emerges in S2. As in S1, there are subpopulations characterized by their differential responses to sensory stimuli, however, in S2 and in all of the more central areas of this circuit, the subpopulations are oppositely "tuned" (Salinas et al. 2000; Romo, DeLafuente, and Hernandez 2004). In S1, all neurons increase their firing with increases in stimulus frequency. In more central areas, approximately half increase firing rate as a monotonic increasing function of increasing stimulus frequency, whereas the other half decrease their rate as a monotonic decreasing function of increasing stimulus frequency. Thus, as stimulus frequency gets slower, the negatively tuned neurons increase their firing rate. Oppositely tuned subpopulations responsive to sensory stimuli are found in S2, PFC, VPC, and MPC (Romo, DeLafuente, and Hernandez 2004).

The above events occur during the presentation of the base and comparison stimuli. During the interstimulus interval (of 3-6 seconds, although this can be increased to 10-15 seconds without a significant difference in performance), no stimuli are presented. To successfully discriminate the first from the second tactile stimulus, and decide which has a greater frequency, the animal must maintain

something like a "mnemonic"[116] trace of the first stimulus. During this period, neurons in PFC correlate their firing rate with the frequency of the base stimulus, with approximately half showing a monotonic increasing relationship to frequency, and the other half showing a monotonic decreasing relationship (Romo et al. 1999). Correlated neural responses during the delay period are also found in S2, VPC, and MPC, also with oppositely tuned subpopulations (Hernandez, Zainos, and Romo 2002; Romo, Hernandez, and Zainos 2004; Salinas et al. 2000; Salinas et al. 1998) .

The comparison stimulus is then presented, whereby neural activity correlates as before in terms of phase-locking and periodicity in the periphery and early S1, and transformed into a rate correlation in S1 and then S2. Rate is also correlated with the stimulus in PFC, VPC, and MPC. Additionally, something like a "comparison and decision process" occurs, whereby the animal/its neurons "decide" which of the two frequencies is greater. The relationship of firing rate *R* to the base and comparison frequencies is given by the regression equation (Hernandez, Zainos, and Romo 2002; Romo, Hernandez, Zainos, Lemus et al. 2002; Romo, DeLafuente, and Hernandez 2004):

$$R = a_1 f_1 + a_2 f_2 + c,$$

where *c* is a constant, f_1 and f_2 are the frequencies of the base and comparison stimulus, respectively, and a_1 and a_2 are coefficients that determine the strength of the relationship between R and frequency. When either of the coefficients is zero, there is no detected correlation between rate

[116] I use scare quotes because whether the animal has memories, or its neural activities are memories, are tantamount to whether the representation and vehicle hypothesis, respectively, are true, and I haven't defended those claims yet.

and that coefficient's frequency. Importantly, when $a_1 = -a_2$, then firing rate is now correlated with neither f_1 nor f_2, but with the *difference*, $f_2 - f_1$.

During the comparison period, neurons in S1 only show correlation to f_2, throughout the stimulation period. In S2, some neurons begin the period correlated with f_2, then the population as a whole shifts towards correlation with the difference, $f_2 - f_1$ (i.e., $a_1 = -a_2$) (Romo, Hernandez, Zainos, Lemus et al. 2002). In VPC and MPC, there are several different populations. Some neurons begin the comparison period correlating with the base frequency, thus, perhaps they are something like "mnemonic traces", whereas others begin the period correlating with the comparison frequency as if they were sensory "representations". Towards the end of the comparison period, the majority of the responsive neurons in MPC and VPC correlate with the difference, $f_2 - f_1$ (Hernandez, Zainos, and Romo 2002; Romo, Hernandez, and Zainos 2004). Additionally, firing rates correlated with $f_2 - f_1$ are found in PFC (Romo, DeLafuente, and Hernandez 2004).

As with neural activity that correlates with the base or comparison frequency, the neural responses correlated with $f_2 - f_1$ (in S2, VPC, MPC, and PFC) show opposite slopes, where approximately half fire more strongly when $f_2 - f_1$ is positive, and the other half fire more strongly when $f_2 - f_1$ is negative.

Finally, M1 plays a crucial role in the animal's behavior during this task. While M1 shows no significant response above baseline activity during the base stimulus, delay period, or early in the comparison period, it does show neural activity correlated with $f_2 - f_1$, similar to the activity found in earlier areas, with subpopulations differentially responsive to the case where $f_2 > f_1$ and where $f_1 > f_2$ (Romo, DeLafuente, and Hernandez 2004).

In a different task, monkeys must categorize rather than discriminate the same type of tactile stimuli, simply saying whether a stimulus belongs to arbitrary categories of high or low learned during training (Salinas and Romo 1998). In this instance, firing rates had a sigmoidal shape: for a neuron that "preferred" higher speeds, its firing rate was essentially the same for stimulus speeds of 22-30 Hz. For a neuron that "preferred" lower speeds, its rate was essentially the same for stimulus speeds of 12-20 Hz (see Salinas and Romo 1998, figures 3 and 4). Thus, as found earlier, there are two subpopulations, each of which is selective for either high or low speeds. The sigmoidal shape of the firing rate as a function of tactile speed suggests that these neurons correlate with arbitrary, learned categories ("high" or "low"). Whether or not that analysis should be applied to the tactile discrimination task is uncertain. However, M1 does appear to play a role in the "decision" procedure for at least the categorization task, and it does have differential activity selective for the different "decisions" the animal may make (i.e., base greater than comparison or vice versa). Whether that differential activity participates in the "comparison and decision" procedure, or simply receives a copy of a "decision" already made, is unclear.

7.4 Defense of the Representation and Vehicle Hypotheses

I will defend the vehicle hypothesis, which states that phase-locked neural responses in the periphery, periodic responses in S1, and firing rate in S2, PFC, VPC, MPC and M1, *in this particular case only*, are physical vehicles of representation. The vehicle hypothesis only says *that* these states are representations, but not what their content is. We will consider how to identify their content shortly. My argument for the vehicle hypothesis needs the representation hypothesis, so let us begin with that.

7.4.1 The Representation Hypothesis

The *representation hypothesis* asserts that the macaque monkeys engaging in the vibrotactile discrimination task have representational states. Namely, the monkeys under consideration have sensory and mnemonic representations and motor plans, and they engage in a cognitive process of comparison and decision. *Sensory representations*, *working memory states*, and so forth, are theoretical constructs from cognitive psychology. However, these constructs share notions of surrogacy or pointing with the dictionary senses of 'representation', and thus deserve the title 'representation'. Additionally, since they are representational in cognitive psychology's sense, then they are representational in my core sense. To support the representation hypothesis, I will present three arguments, in what I take to be increasing order of strength.

Species argument: Macaque monkeys are mobile organisms, able to quickly react to varying external conditions in the environment, and to guide and direct their own behavior according to those varying conditions in ways that are conducive to survival and reproduction. This includes abilities to anticipate, to learn, to use tools, etc. The best explanation of these observations is that macaque monkeys have internal states that stand in for, mirror, or represent the world. More specifically, the best explanation is that macaques have sensory representations, short-term memory, and long-term memory, which together allow for anticipation, learning, tool use, and so forth. The particular monkeys engaging in the discrimination task are tokens of this species, therefore, these individual monkeys have representational states.

Individual best explanation argument: The individual monkeys engaging in this behavior reliably discriminate flutter frequencies at a very high accuracy level. More specifically, they reliably *press the medial button* when the comparison frequency is greater, at levels far above chance. For

example, a trained monkey, presented with a comparison that is 8 Hz higher than the base, will press the medial button 97% of the time, whereas an untrained monkey will either press the buttons randomly or not press them at all (Romo, DeLafuente, and Hernandez 2004). Certainly there is a neurochemical *basis* for the animal's observed behavior, and this is extremely important to a complete understanding of its behavior. However, positing working memory, sensory representations, motor plans, and a decision process explains and predicts the animal's behavior.

These posits explain behavior by incorporating the description of monkey behavior within a wider, reasonably well-established body of cognitive theory, which is itself grounded in some of our common folk concepts, such as memory, sensation, and so forth. It also allows for prediction: I can predict that a trained monkey performing this task will most likely press the medial button when presented with, say, a 10 Hz base stimulus and an 18 Hz comparison, on the grounds that the monkey has sensory and short-term memory representations of those stimuli, can compare them and decide which is greater, and will output motor behavior signaling its choice. This explanatory framework also allows for manipulation of manifest phenomena, since we can manipulate the animal's behavior. If I wanted to make it press the medial button, say, I could present it with frequencies in which the comparison is greater by 8 Hz.

The representation hypothesis may not allow for *differential* prediction, where it predicts something that a non-representational explanation would not. For example, a non-representational hypothesis that simply notes the correlation between behavior and frequency difference will allow for the predictions made above. In terms of prediction, we might consider the representational and non-representational hypotheses equally supported, at least thus far. However, the representational hypothesis adds explanatory unification that the non-representational hypothesis lacks. By describing

the animal's behavior in terms of categories from cognitive psychology, the explanation of behavior becomes integrated within a widely accepted conceptual framework.

Additionally, while the non-representational hypothesis may be equal to the representation hypothesis in terms of *predictive* power, the non-representational hypothesis doesn't actually explain *why* those correlations exist. The representation hypothesis does: The animal reliably behaves that way because it has learned that pressing the medial button when and only when presented with two frequencies in which the comparison is greater results in a squirt of juice, and the animal wants the juice. Since the representational explanation is the best explanation of the animal's behavior, we should conclude that the animal has representational states.

Humans and Occam's razor: Macaque monkeys and humans engaging in the vibrotactile discrimination task have indistinguishable psychophysical curves (Mountcastle, LaMotte, and Carli 1972; Talbot et al. 1968; LaMotte and Mountcastle 1975; Mountcastle, Steinmetz, and Romo 1990). That is, when we map response judgments against the various physical properties of the stimuli for both humans and macaques, we cannot reliably tell, from these curves alone, whether the subject was a human or a monkey. Were I to engage in the task, the best explanation of my behavior would be that I have a sensory representation, hold it in short-term memory, compare it, and so forth. I take it as given that I can remember things, and that I have short-term memory. Since (i) the stimuli presented are type identical, (ii) the behavioral outputs (press the medial or lateral button) are indistinguishable, and (iii) the psychophysical curves generated by those responses are indistinguishable, the best and simplest route is to explain both human and monkey behavior in the same way. Since the human explanation involves representational states, then the explanation of monkey behavior should as well. To grant the representation hypothesis of me, but to deny it of the monkey, would be arbitrary, unjustified, and violate the simplest explanation requirement. Therefore, these monkeys have representational states.

It may seem that I am pressing overly hard here; isn't it obvious that monkeys can remember? I do this because the representation hypothesis is a crucial element of the defense of the vehicle hypothesis, which is necessary to support claim 1. Thus, I'd rather risk overkill than not provide enough support for such a crucial claim. Second, there is a debate about what is known as *cognitive ethology*, which is only tangentially related to the representation hypothesis, but which I seek to avoid[117].

Ethology is the study of animal behavior, and typically emphasizes behavior in a natural environment. The central question of cognitive ethology is whether it is appropriate to attribute intentional states and processes to nonhuman animals in the explanation of animal behavior. This is only tangentially related to my representation hypothesis for several reasons. First, the experimental paradigm under consideration does not involve animal behavior in its natural environment, but only in a contrived laboratory situation, which is not the typical province of ethology. More substantially, the majority of the major figures in the cognitive ethology debate take it for granted that what is at issue are *intentional states*, which are something like the whole patterns of behavior that Dennett describes in his intentional stance[118], and which crucially involve the generation of intensional contexts in their description as well as carry the presupposition of minimal rationality. While an interesting question, this is not my question. I'm after something much more basic.

Nonetheless, I wouldn't want to rest my case on such an apparently flimsy defense. Therefore I've provided three arguments that I take to be in increasing order of strength. The species argument does not need ascriptions of rationality or generation of referential opacity; it only needs that we make the basic assumption that monkeys have states that point to, or stand in for, external events or things. The best explanation for the complexity and adaptiveness of monkey behavior involves adverting to

[117] See for example (Bekoff and Jamieson 1996; Cummins and Allen 1998; Allen and Bekoff 1997).
[118] Dennett is in fact one of the principal and founding players in this debate. See his (1983).

these states, although it may remain silent on whether the monkey is conscious, or has beliefs or desires, and so forth. The individual best explanation argument makes the case for sensory representations, working memory and motor plans even stronger, since the reliability with which certain manifest phenomena are observed is quite high, and positing representational states allows for explanation, prediction, and manipulation of those observations.

Finally, if any doubt remains, I take the humans and Occam's razor argument to be decisive. It is quite an interesting finding to discover that human and macaque psychophysical curves are indistinguishable for this task. A macaque monkey and a human will perform statistically the same. We have the same discrimination threshold of about 3 Hz. As the frequency difference widens and discrimination accuracy increases, it increases along the same curve for both humans and monkeys. Similar indentation amplitudes are required for successful performance on the task. Also relevant to this argument is that humans and macaques have similar neural structures that implement this task. We both have Meissner's corpuscles, the path into the central nervous system involves the same three neurons to get from the periphery to S1, and so forth.

It must be taken for granted that I have short-term memory and sensations, and this establishes the representation hypothesis for me. But since the stimuli and behavioral responses are indistinguishable, and since humans and monkeys share homologous neural structures that implement the task, it is arbitrary to deny the representation hypothesis of our Macaque cousins. This argument bypasses the cognitive ethology debate because there is no question of how best to interpret the animal's behavior. Since the animal's behavior is essentially identical to human behavior, we should interpret them both the same way, in representational terms.

The internal events that determine animal behavior, whatever they are, are best described in representational terms, just as we would describe the internal events that determine my behavior, were I to engage in this task. Thus, the monkeys have sensory representations, working memory, and motor plans. This establishes the representation hypothesis.

7.4.2 The Vehicle Hypothesis

The vehicle hypothesis asserts that phase-locked responses in the periphery, periodic responses in subpopulation-1 of S1, and the firing rate of individual neurons in subpopulation-2 of S1, S2, PFC, VPC and MPC are vehicles of sensory representation. It also asserts that the delay period responses in S2, PFC, VPC, and MPC are vehicles of mnemonic representations, and the firing rate of individual neurons in M1 are vehicles of motor representations, or motor plans.

I will focus initially on firing rate in S1, as establishing the vehicle hypothesis with respect to firing rate in S1 will support it for the other brain states.

The first step in supporting the vehicle hypothesis is the representation hypothesis. The representation hypothesis explains behavior, and in particular, the animal's button-pressing behavior in response to tactile stimuli. Since representations are causally efficacious we conclude that those states adverted to by the representation hypothesis are efficacious in the production of the behaviors involved in the discrimination task.

Second, as demonstrated by lesion studies (LaMotte and Mountcastle 1979; Zainos et al. 1997), firing rate activity in S1 is necessary for successful performance of the discrimination task. Monkeys will continue to perform the task, but without S1 intact, their discrimination accuracy is essentially at

chance. Additionally, microstimulation studies have demonstrated that firing rate activity in S1 is sufficient to drive the entire set of neural events involved in successful discrimination.

Romo and colleagues (Romo et al. 1998; Romo et al. 2000) used implanted electrodes not only to record, but also to inject pulses of current into S1. While the monkeys performed the task, electrical stimulation pulses, oscillating at the same frequency as the mechanical stimulation would have been, randomly replaced the mechanical stimulus, for all combinations of the base and comparison stimulus. They found that the monkeys were able to perform the discrimination task at accuracy levels indistinguishable from the natural, mechanical stimuli alone. On these grounds, we may conclude that firing rate activity in S1 is *causally relevant to*, and not merely correlated with, successful performance of the vibrotactile discrimination task.

Third, firing rate in S1 is reliably correlated with stimulus frequency. It bears emphasis that these stimuli are the ones about which the cognitive decision is made. The behavior of reliably pressing the medial button when the comparison frequency is higher is explained by the representation hypothesis, those representations cause this specific behavior, firing rate in S1 is causally necessary and sufficient for that same behavior, and firing rate in S1 is correlated with the very stimuli on which the behavior depends.

Fourth, there are correlations between firing rate in S1 and animal behavior, in two different ways. First, as mentioned above, without this activity the animal is unable to successfully perform the task. Second, even without lesions, when the animal makes a behavioral error[119] there is a correlation between the firing rate activity of its neurons and its behavioral error (Salinas et al. 2000).

[119] At this point in the dialectic the representation hypothesis is established. This justifies claiming that the animal made a *behavioral error*.

In S2 there was a significant difference in standardized rates in individual neurons when the monkey selected correctly and when it did not, in both types of tests (where base is greater than comparison and vice versa). The same rate discrepancy is found in S1[120]. Thus, if the animal made an error in behavioral discrimination, signaling that it found the comparison to be lower than the base when in fact the comparison was higher than the base, there is a significant likelihood that individual neurons in S1 and S2 were firing at a rate that is lower than they would have been firing, given that frequency, had the animal discriminated correctly. This correlation between neural activity and behavior holds, mutatis mutandis, for the case when the animal incorrectly judges the comparison to be higher than the base when it was in fact lower.

Fifth, the neurometric curves calculated with firing rate in S1 are very similar to the psychophysical curves of the animal (Hernandez, Zainos, and Romo 2000). A *neurometric curve* plots the probability that an ideal observer using (say) firing rate could correctly discriminate the two stimuli. That curve can be directly compared to a psychometric curve, which plots the animal's discriminatory behavior against the frequency difference. Neurometric curves for firing rate were computed using the following simple rule: if there are more spikes during f_2, then f_2 is higher. The neurometric curves calculated using firing rate for individual neurons in S1 were very similar to the animal's psychometric curves.

Sixth, the discrimination thresholds between the animal and the firing rate of individual neurons in S1 are indistinguishable (Hernandez, Zainos, and Romo 2000; Salinas et al. 2000). The *discrimination threshold* is the difference between base and comparison that is necessary for the animal (or neuron) to

[120] There's a qualification here: With measures of individual neurons in S1, the difference in standardized rate between hit and error trials was significant for the cases where the base is greater than the comparison, but in the case where the comparison is greater, the difference for individual neurons is within that expected by chance. See Appendix B for more on this.

correctly discriminate 75% of the time, where the neuron's ability to "discriminate" is calculated as above, using the ideal observer's rule.

To review, firing rate in S1 is causally relevant to successful vibrotactile discrimination behavior: It is both necessary and sufficient (jointly with other states), for driving the complete set of neural events involved in vibrotactile discrimination. Firing rate in S1 is correlated with the stimuli on which that behavior depends. Rate is also correlated with the behavior itself: If the animal presses the lateral button when in fact the comparison was higher, the firing rates of neurons in S1 will be lower than they would have been, had the animal pressed the medial button (i.e., had the animal discriminated correctly), and vice versa for the other type of error. The neurometric curves and neurometric discrimination thresholds for firing rate in S1 are respectively similar to and indistinguishable from the animal's psychometric curve and psychometric discrimination threshold. Finally, we have already accepted on independent grounds that the animal has sensory representations and those sensory representations are causally efficacious in driving vibrotactile discrimination behavior.

The simplest, most unified, and therefore best explanation of all of the above findings is that firing rate in S1 is a physical vehicle of representation. Firing rates of individual neurons in S1 just are (at least some of) those representations adverted to by the representation hypothesis. The vehicle hypothesis explains the above findings from the electrophysiological studies by unifying the two explanations. While the representation hypothesis explains behavior, the vehicle hypothesis explains both the discovered correlations as well as extends and refines the explanation of behavior provided by the representation hypothesis.

One might well ask[121]: What is added, what additional prediction is made, what is the extra explanatory payoff, of saying that these neural states represent, rather than simply that they are correlated with stimulus frequency at the fingertips? First, the empirical findings are not simply that firing rate in S1 is correlated with stimulus frequency at the fingertip. The empirical findings are much more robust and diverse. Nonetheless, we may still ask: What extra explanatory payoff do we get in assuming the vehicle hypothesis, rather than simply saying that these states correlate with the stimulus, with behavior in various ways, etc.?

The most important explanatory virtue provided by the vehicle hypothesis is explanatory unification[122]. We should accept the representation hypothesis on the independent grounds discussed in 7.4.1. But given that we've accepted that the monkey has representations and that they are causally efficacious in this particular behavior, and given the several different ways that these brain states are related to that behavior, by accepting that these brain states just are those representations, we get a unified explanation both of the animal's behavior as well as those correlations.

Second, without accepting the vehicle hypothesis, we have *no* explanation for some of the discovered correlations. We can explain the correlation between a neuron's firing rate and stimulus frequency at the fingertips in terms of anatomical connections and the electrophysiology of neural structures. However, that electrophysiological explanation begins to falter when we start adding up the various correlations to the animal's behavior. Why does the firing rate for a neuron correlate with the

[121] As Michael Levin has, in comments on this chapter.

[122] Feigl has this to say on unification: "The aim of scientific explanation throughout the ages has been *unification*, that is, the comprehending of a maximum of facts and regularities in terms of a minimum of theoretical concepts and assumptions" (Feigl 1970, 12). Philip Kitcher, from whom I draw the previous quote, says this: "a theory unifies our beliefs when it provides one (or more generally, a few) pattern(s) of argument which can be used in the derivation of a large number of sentences we accept" (Kitcher 1981, reprinted in Boyd, Gasper, and Trout (eds.) 1991, 333). Kitcher's theory gets very complicated as he cashes out his notion of argument patterns, but the basic idea is that unification involves explaining things that were previously thought to be disparate, in common terms, or with a single ontology.

monkey's button-pressing behavior, but only after the animal has been trained? Why is there a correlation between behavioral error and neural rates being different than they would have been had the animal not been in error? Why do the neurometric discrimination thresholds match up so well with the animal's psychometric discrimination thresholds, and similarly for the neurometric and psychometric curves? Surely there is a neurochemical *basis* for the neural and animal behavior, but, given the representation hypothesis, we get a more simplified and unified explanation of *both* the discovered correlations as well as the animal's behavior by accepting the vehicle hypothesis.

Further, the findings of causal efficacy should not be ignored. Since the representation hypothesis implies that the animal has unspecified states *X*, *Y*, and *Z* that are causally efficacious in producing successful discrimination behavior, and since the lesion studies and microstimulation studies demonstrate that firing rate in S1 is causally necessary and sufficient, jointly with other states, for producing that same behavior, it is reasonable to identify firing rate in S1 with *X*, *Y*, or *Z*. But since *X*, *Y*, and *Z* are, respectively, sensory representations, short-term memories, and motor plans, and since there is no reason to identify firing rate in S1 with short-term memory or motor plans, we should therefore conclude that firing rates in S1 are sensory representations.

A further explanatory payoff is that we now get a *better* explanation of animal behavior than we had from the representation hypothesis alone. When conjoined to the representation hypothesis, the vehicle hypothesis explains the animal's behavior at *both* a cognitive level and a neurological level. This is precisely what we should be after, since it provides a unification of biology with psychology. In providing that unification, we get greater predictive, manipulative, and explanatory utility than we had from the representation hypothesis alone.

For example, given certain parameters, such as a difference between base and comparison of 8 Hz, we could not predict with any reliability when the animal would make a mistake from the representation hypothesis alone. But, given the vehicle and representation hypotheses conjoined, we can predict when the animal will be in error from the firing rate of its neurons in S1. Additionally, we are afforded greater manipulative power, as demonstrated by the microstimulation studies. The representation hypothesis says nothing about particular brain states, and thus, could not say which neurons to stimulate, or how, in order to get the monkey to press the medial or lateral button, whereas the vehicle hypothesis does.

To clarify, we don't need the vehicle hypothesis to predict when the animal will *press the lateral button, given that the comparison was higher or lower*. That can be predicted without assuming that these states are representations. But given that we accept the representation hypothesis, we get the added predictive utility from the neuroscience as well as explanatory unification of cognitive psychology with neuroscience.

For all of these reasons, I conclude that the vehicle hypothesis, with respect to firing rate in S1 during the vibrotactile discrimination task in a trained monkey, is well supported. The vehicle hypothesis provides the extra explanatory payoff of explanatory unification. Without the vehicle hypothesis the *multiple* correlating relations between brain states, energy states, and behavior, are unexplained. The causal efficacy of these states, coupled with the causal efficacy of the representations adverted to by the representation hypothesis, is best explained by the vehicle hypothesis.

Once we have the vehicle hypothesis in hand for firing rate in S1, the vehicle hypotheses for phase-locked responses in peripheral afferents, for periodic responses in subpopulation-1 of S1, for firing rate in S2, VPC, MPC, PFC, and M1 become much more plausible. To accept the vehicle hypothesis

for firing rate in S1 but not in these other areas, when the same or similar considerations apply, is arbitrary and violates the simplest explanation standard. Thus, we should conclude that the vehicle hypothesis holds for each of those states as well.

While the representation hypothesis explains the animal's behavior and the vehicle hypothesis explains the discovered correlations in addition to extending and improving on the explanation of the animal's behavior, we do not as of yet have an explanation *of representation*. (That is, logically, in the dialectic here, we don't yet have that explanation.) While both of these hypotheses make use of the concept of representation, neither of them purports to give an explanation of what representation is, or of what determines representational content. To a certain extent then, both the representation and vehicle hypotheses, while explanatorily virtuous, are incomplete. They both depend on a concept that, while loosely understood, is problematic: It isn't clear how "pointing to", "being a surrogate for", or simply *aboutness*, are a part of the natural order. For that explanation, we turn to the structural preservation theory of representation.

I have defended the vehicle hypothesis without assuming SPT, or that representation is covariation, or any other substantial theory of representation. Rather, I have attempted to show that we can in fact identify the physical vehicles of representation in an active nervous system in at least one case, much the way we can identify water as such, without a substantial theory.

I would emphasize that the empirically discovered correlations are indicative, but not constitutive, of the presence of representational vehicles. Second, I only claim that my argument works for these brain states under consideration here, and I make no general claims on how to identify representations in the absence of a theory.

7.5 Applying SPT to the Brain: Claim 1

7.5.1 The Abduction

Claim 1: If we find representations in the brain, then they will have the properties attributed to them by SPT.

To support claim 1, we need a manner of independently identifying representations as such, and identifying their content. From the vehicle hypothesis we have identified several representations; however, we have not yet identified what they represent. Since these representations were identified partially on the basis of correlations with stimuli about which a cognitive decision was made, we can use those correlations to identify representational content. For example, if a phase-locked neuron in the periphery is firing at a rate of 25 bursts/sec, then we should pre-theoretically identify that vehicle as representing the vibratory rate of the stimulator as being 25 Hz.

I must admit that, while the claim that phase-locking is a vehicle of representation has been supported by my arguments above, I made no arguments supporting the further claim that we can identify the content of those vehicles using correlation. However, at this point in the dialectic, the vehicle hypothesis has been established: phase-locking is a vehicle of representation. Representational vehicles have contents. The correlation between a neuron's firing a spike or burst of spikes for each sinusoidal wave of the stimulus is part of what was appealed to in establishing the vehicle hypothesis. Additionally, phase-locking is *defined* by its firing a burst for each sinusoidal wave of the stimulus. From these considerations, and given the vehicle hypothesis, it seems implausible to claim that the representational content of a phase-locked response is something else, barring, of course, overriding reasons to the contrary. Additionally, and for the same reasons, I will determine the content of firing rate vehicles similarly.

I imagine the following objection: While we may grant the vehicle hypothesis, as you mention, the vehicle hypothesis does not establish the *content* of those vehicles, only that they *are* vehicles. Why should we accept what you call your "pre-theoretic" method of identifying content? You're just begging the question because you are going to identify contents in a way that will match your theory.

In reply, note that the vehicle hypothesis is not at issue. Thus, these states have some content, and all that is in question is what that content is. While it is open to an objector to deny that I have correctly identified the contents in this case, it is *not* open to an objector to do so without a better suggestion on what the content is. That is, *given* that burst rate is a vehicle of representation and thus has some content, what is its content, if not what I claim it is? This is the force of my implausibility argument: Given that it has some content, the content that I've identified is the only plausible one.

It is absolutely crucial to recognize that here I use correlation as an *indicator* of the presence of a representational vehicle and of its content, but I do not presuppose a covariation theory of representation, which says that representation *is* covariation. That is, correlation is *indicative*, but not *constitutive*, of representational content. In this case alone, I claim that correlation is a "superficial indicator" that we can use to identify representations and their contents, much as we use a thing's being a colorless liquid found in our lakes and streams as indicative but not constitutive of water.

Finally, I will clarify what SPT does and does not predict. Given SPT, we should expect to find that vehicles of representation are related to their contents in two ways: causal history and structural preservation. We should find that, if *R* represents, say, the vibratory rate of the stimulator, then *R* was caused by the stimulator. We should also find that, whatever frequency *R* represents the stimulator as having, that stimulator frequency and whatever parameter defines *R* (e.g., firing or burst rate) are members of independently specified relational systems that bear structural preservation to one

another. Additionally, when we confront the mapping component of the non-uniqueness problem, thereby defining the representation function *f* for SPT, it should turn out that *f* maps the right representation parameters to the right represented parameters. That is, *f* should match up with the pre-theoretic content identification from above, mapping firing rates or burst rates to stimulus frequencies in the same way. In other words, SPT predicts that if we find representations in the brain, we will find a *system* of representations, the members of which will be organized in such a way that that system structurally preserves a different system. The latter is a system of representeds, which, SPT predicts, will also be organized in such a way that the represented system and the representation system structurally preserve one another.

SPT does not make any predictions about which correlations will be found. It does not predict that neural activity in the periphery will be phase-locked, nor that responses in S1 will be periodic, nor that firing rate elsewhere will correlate with stimulus frequency. It does not predict that there will be oppositely tuned subpopulations. SPT does not predict anything about the *brain*, because it is not a theory about the brain. SPT is a theory about representation; it explains representation by saying what it is. More specifically, SPT explains representation by saying what it is to be a representation (mostly following Millikan here), and, what is my contribution, by saying what determines representational content. Strictly speaking, it is irrelevant to the logic of the argument that vehicles of representation were found in the brain. However, once vehicles of representation have been found, if SPT is true, then those vehicles must accord with the claims SPT makes about representation. If those vehicles do not accord with SPT's claims, then SPT is false.

7.5.2 Sensory Representations

I will discuss four different kinds of sensory representations: the peripheral burst code, the periodic/temporal code in subpopulation-1 of S1, and both the positively and negatively sloped rate codes in S2. I discuss the causal chain problem and error in the context of discussion of the rate codes. I discuss several different cases in detail for the following reasons. First, I find it an intrinsically interesting exercise to see if and how SPT applies to neural states, and I imagine others would as well. Second, and more important, given that this is an ampliative, abductive argument, the more examples that I find, the stronger the argument gets. On a related note, I want to show in what follows that these are not isolated cases. As it turns out, we will identify several different representations in several different neural areas (using the theory-independent vehicle hypothesis) that accord with SPT.

7.5.2.1 Peripheral Burst Code

The vehicle hypothesis establishes that phase-locked neural responses in the periphery are vehicles of representation, and we should further conclude that, when a neuron fires at, say, 25 bursts/sec, it represents the vibratory rate of the stimulator as being 25 Hz, for the reasons mentioned in 7.5.1. Now we look at the vehicles of representation, what they represent, and the relations between vehicle and content, and see what SPT would say. From the perspective of SPT, we need to first establish the metaphysical component of the thesis (i.e., is this thing a representation according to SPT?), then we'll move to the content component.

SPT, following Millikan, says that a state is a representation if it has the teleofunction of bearing some correspondence relation to some state of affairs, and indicatives are distinguished from imperatives by the teleofunctions of the consumer-states that use them. It is not tendentious to claim that sensory states have the teleofunction of covarying with energy states at the periphery of the

organism; hence, sensory states have the teleofunction of bearing some correspondence relation to states of the world. Further, those states are used by other neural mechanisms (i.e., the representation-consumers) to direct evolutionarily adaptive behavior of the organism as a whole. We have at least a preliminary argument in favor of the metaphysical component of SPT. I'll not delve more deeply into this issue since this is not my contribution and because I only endorse the broad outlines of Millikan's work here. Further, the content component would seem to be the more interesting and difficult component[123].

We'll define our relational systems as follows. Let $\mathfrak{A}$ = the stimulus relational system and $\mathfrak{B}$ = the physiological relational system. The domain of the stimulus relational system, A, consists of vibrotactile frequencies, and is ordered by $>_A$, the empirical higher-frequency-than relation. The domain of the physiological relational system, B, is *burst rate*, which is calculated as the number of bursts per stimulus presentation time. We then define a *burst* in terms of interspike intervals: a burst is "a group of spikes in which all intervals between consecutive spikes [is] less than τ msec" (Salinas et al. 2000). The shorter that τ gets, the closer burst rate will be to firing rate. For our purposes here, whatever τ maximizes the linear fit of the function from frequency to burst rate should be chosen to define 'burst'. The domain of B thus consists of burst rates, and is ordered by $>_B$, the empirical greater-burst-rate relation.

We'll need to know more about $\mathfrak{A}$ and $\mathfrak{B}$ in order to determine whether structure is preserved and if so, what kind of structural preservation obtains. Specifically, how many members do A and B

[123] There might seem to be an incongruity here. I spent a great deal of time in establishing the vehicle hypothesis, which is the pre-theoretic method of identifying vehicles of representation as such, and hence, is analogous to the metaphysical component of SPT. I've spent comparatively little energy in defending the manner in which I pre-theoretically identified the *content* of those vehicles, which is analogous to the content component of SPT. Now, when applying SPT, I seek to do the reverse. The incongruity is only apparent. The vehicle hypothesis is necessary in order to develop and defend the method of identifying contents. Once the vehicle hypothesis is established, the content-identification method (ampliatively) follows almost immediately. With respect to SPT, my contribution is principally in the content component, and that is what most philosophers focus on as well.

have? What kind of ordering do the empirical relations induce on the sets? Should we include frequencies outside the 5-50 Hz range in the domain of $\mathfrak{A}$?

To get a handle on these questions, consider the experimental procedures used as well as the theoretical assumptions that must be made if the measurement and statistical modeling is justified. In most of the experiments, only eight frequencies were used to determine the response properties of the various cells under investigation. From there, regression analyses were used to "fit" the responses of the cell to the properties of the stimulus. That is, using eight frequencies, a line is generated, which fills in the infinitely dense gaps between each of those frequencies. This constructs a continuous, monotonic function, which predicts what the response of the cell would be for any frequency within the modeled range. At least given the assumptions necessary to construct this model, we should then assume that the domains of both $\mathfrak{A}$ and $\mathfrak{B}$ are continua.

There are further assumptions underlying the measurement of firing or burst rate and frequency and the use of the various continuous statistical measures to make inferences about them. The assumption is that each rate/frequency is assigned a unique real number. Since the real numbers are continuous, if each rate or frequency is to be assigned a unique real number, the relational system composed of firing rates/frequencies must be isomorphic to $\mathfrak{N}^+ = \langle \mathbb{R}^+, \geq \rangle$. But this in turn implies that the empirical relations that order the empirical quantities are total orders, the domains of both relational systems have countable order dense subsets, and, again, have continuum many elements (see Appendix A).

Finally, it doesn't matter whether we include frequencies outside the 5-50 Hz range in A. Since I have included Swoyer's various relaxations in the definition of structural preservation (5.4.5), if structure is preserved when those frequencies are excluded, it will also be preserved with their

inclusion. For example, if they are excluded from A and isomorphism obtains from $\mathfrak{A}$ to $\mathfrak{B}$, then by including them but defining f only over the elements in the original set, there will exist an isomorphism* from $\mathfrak{A}$ to $\mathfrak{B}$. In both cases, structural preservation obtains from $\mathfrak{A}$ to $\mathfrak{B}$.

Thus, if we assume that the measurement and use of continuous statistical measures with respect to frequency and firing or burst rate is justified, then we can assume that $\mathfrak{A}$ and $\mathfrak{B}$ both have uncountable domains with countable order dense subsets, and are total orders. Given these assumptions, it follows that $\mathfrak{A}$ is isomorphic to $\mathfrak{B}$ (5.4.6, U1). Now we need to solve the mapping component of the non-uniqueness problem. As discussed in 5.4.4, while from the perspective of measurement theory it is acceptable to prove a Uniqueness Theorem, from our perspective, we need to rule out every isomorphism-determining function except for one, which is the *representation function*. All that we have shown thus far is that there exists a structure-preserving mapping from A to B. We do not yet have an argument for which is the one that determines f-predicative content.

The mapping component of the non-uniqueness problem is solved with teleology. Of the numerous structure-preserving mapping functions connecting the two independently specified relational systems, the representation function is the one which is the device's *teleofunction* to bear to some state of affairs. I've argued, following Millikan, that a state is a representation if it has the teleofunction of bearing some particular correspondence relation, so that it's doing so is adaptive for the organism of which that state is a part. So we can ask: Do we have any reason to believe that the states of the primary, secondary, and tertiary afferents in the rapidly adapting circuit under consideration have the teleofunction of bearing some correspondence relation to anything? As a start in answering that question, we ask a simpler question: Do those states reliably covary with anything, and if so, can we define that covariance in terms of a mapping function from the domain of $\mathfrak{A}$ to the domain of $\mathfrak{B}$? For these latter questions, we have answers.

The electrical states of the peripheral afferents under consideration, defined in terms of burst rate, reliably covary with vibrotactile frequency in the flutter range of 5-50 Hz, each at their particular receptive fields. That reliable covariance is the simple one of phase-locking to the stimulus, hence, is adequately described by the function $r_1: A \rightarrow B$, where $r_1(x) = x$. This function maps frequencies to frequencies, where x Hz stimulator frequency maps to x bursts/sec.

My *teleofunction hypothesis* here is that those cells have the teleofunction of covarying, according to r_1, with vibrotactile stimulations in the flutter range occurring at the receptive field of the respective cells. Another way of stating that hypothesis is this: Burst rate covaries with vibrotactile frequency because, in the course of evolutionary history, there was selection for peripheral nerves that emitted a burst at a rate equal to frequency of a sine wave of pressure on the fingertip[124]. On what grounds should we believe the teleofunction hypothesis?

First, consider the specificity of ambient energy needed to reliably generate a train of action potentials. Due to the microanatomy of Meissner's corpuscles, only vibrating mechanical energy in the 5-50 Hz range, at the superficially located level (around 500μm beneath the surface), will generate trains of action potentials. Faster or deeper vibrations simply won't activate the Meissner's circuit, but will instead activate Pacinian corpuscles, and slower indentations in the form of constant pressure will activate the slowly adapting mechanoreceptors and their associated afferents (Gardner, Martin, and Jessell 2000; Gardner and Kandel 2000). And these are each forms of tactile, mechanical energy. Electromagnetic, chemical, thermal, or acoustic mechanical energies won't activate this circuit at all.

[124] I borrow this second way of formulating the teleofunction hypothesis from Michael Levin. Notice that teleofunction is explicated in terms of evolutionary history: Whatever tokens of a type of state *did*, which is causally responsible for the existence or preservation of that type of state, is the teleofunction of those states. Thus these two formulations are equivalent.

The specificity of kinds and levels of energy required to activate this circuit is governed by basic physical laws and the microanatomy of the relevant cells[125]. Thus, r_1 is not only a structure-preserving function connecting $\mathfrak{A}$ and $\mathfrak{B}$. Additionally, r_1 describes a regular covariation that is grounded in physical laws. Thus, there is a nomically grounded covariation between the frequency of superficial mechanical indentations in glabrous skin and burst rate in the rapidly adapting peripheral afferents, which is described by r_1. This is my first premise in support of the teleofunction hypothesis.

Second, the tactile sensitivity of the glabrous areas of primate skin makes possible various evolutionarily adaptive behaviors, such as grasping objects and tactile recognition, which in turn aid us in getting food into our mouths. We primates do all sorts of things with our hands, which contribute to behavior that is conducive to survival and procreation. While we should be wary of just-so stories, in the cases under consideration the presumption should be in favor of the claim that the nomically grounded covariations under consideration are or were evolutionarily adaptive. Thus, we should conclude that the teleofunction of the primary, secondary, and tertiary afferents associated with the rapidly adapting circuit is to covary with mechanical deformations at their respective receptive fields, according to r_1. Since, according to SPT, teleofunction defines the representation function and solves the mapping component of the non-uniqueness problem, we therefore conclude that according to SPT, r_1 is the *representation function*.

Given specifications of relational systems as well as a representation function, we now have solutions to the systemic and mapping components of the non-uniqueness problem. This provides us with determinate f-predicative content for the various states of the peripheral afferents under

[125] We should avoid getting embroiled in discussions about the nature of explanation in general, or how it applies to neuroscience. However, I should mention that I'm not assuming a covering-law model of explanation: The mechanistic model of explanation (see for example Craver 2007) fits here as well. I argue that the activity of Meissner's corpuscles is *governed* by lawlike regularities in the universe. This does not imply that explanation involves deducing their behavior from a statement of those laws.

consideration. If the afferent is firing a burst of action potentials at a rate of, say, 25 bursts/sec, then that state is a representation, which f-predicates the property of having a frequency of 25 Hz, to whatever it f-refers to.

F-reference is determined by causal history. Although we haven't delved into the details, we can easily see that these states will fall into the indicative category of representations, and hence, they f-refer to what caused them, rather than what they caused. Assuming that the afferent's bursting activity was caused by the mechanical stimulator, that activity f-refers to the stimulator. Putting the components together, 25 burst/sec activity of the rapidly adapting afferents, assuming it was caused by the experimental stimulator, has the representational content that *the stimulator is vibrating at 25 Hz*. If the stimulator is indeed vibrating at 25 Hz, then the activity of this neuron *is true*, and if not, then the activity of this neuron *is false*[126]. More colloquially, SPT claims that if the stimulator is vibrating at 25 Hz then the burst rates of these neurons correctly represent the vibratory rate of the stimulator as being 25 Hz. This is in agreement with the pre-theoretic identification of content based on the vehicle hypothesis, which is what we should expect if SPT is true.

7.5.2.2 Temporal Code in S1

The vehicle hypothesis established that periodic firing in subpopulation-1 of S1 is a vehicle of representation. The pre-theoretic identification of the contents of these vehicles should run in a similar fashion to burst rate: The neural firing represents the vibratory rate of the stimulator as being at whatever frequency contributes most to the temporal pattern of that neural firing. The same grounds that justified content identification for phase-locked peripheral responses justify this method of content

[126] It might seem a bit jarring to say that the activity of a neuron *is true*. Nonetheless, from the vehicle hypothesis, which confirms what SPT says about this, this bit of neural activity has representational content. Sentences, for example, *have* representational content, and sentences *are true* or *are false*. Similarly, this neural activity *has* content, and so, *is true* or *is false*.

identification here. Namely, the vehicle hypothesis is established, and so it is not in question whether periodic responses in S1 are representations. They are vehicles of representation so they must have content. Given that they have content, and that part of the reason why the vehicle hypothesis has been established is the correlation between the temporally defined activity of the neuron and stimulator frequency, it follows that the only plausible claim here is that the neural firing represents the vibratory rate of the stimulator as being at whatever frequency contributes most to the temporal pattern of that neural firing. Now we'll see if SPT agrees with this pre-theoretic claim.

Let $\mathfrak{A}$ = the stimulus relational system and $\mathfrak{B}$ = the physiological relational system. I'll continue to make the idealizing assumptions about $\mathfrak{A}$; thus, its domain has a countable order dense subset yet has continuum many elements, $>_A$ induces total ordering on A, and the question of whether to include frequencies outside the 5-50 Hz range is irrelevant. If $\mathfrak{A}$, exclusive of items outside that range is isomorphic to $\mathfrak{B}$, then including further items in A, yet defining r only over the initial set, gives us an isomorphism* from $\mathfrak{A}$ to $\mathfrak{B}$.

Subpopulation-1 is best described in terms of its correlation between periodicity and frequency. To define $\mathfrak{B}$, we'll define the members of B in terms of PSFP, or power spectrum frequency at peak (Salinas et al. 2000). Briefly, recall that PSFP is calculated with a Fourier decomposition of the time course of neural activity, then the frequency bin with the peak power is found, and its median taken. This is the frequency that contributes most to the oscillatory activity of the particular neuron under consideration. Each member of B is a frequency, and so the natural ordering relation to choose is the greater-frequency-than relation, $>_B$.

As with the calculation of burst rate discussed previously, notice the initial inductive reasoning. Eight frequencies are used experimentally, and the PSFP for each neuron is calculated for that frequency. The gaps are filled by a regression, creating a continuous monotonic function that predicts,

for any given stimulus frequency and neuron, what the PSFP will be for that neuron and that frequency. Second, this is not a measurement of "more or less" periodicity. It is a measurement of which frequency component of the overall activity of the neuron contributes most to its oscillatory activity.

The same idealizing considerations apply for PSFP as they do for firing and burst rate. If the measurement of PSFP is justified, if the regression analysis which fits PSFP to stimulus frequency in a continuous line is justified, and if we assign unique real numbers to each frequency, but no two frequencies to the same number, then it follows that $\mathfrak{B}$ is isomorphic to $\mathfrak{N}^{+}$. From this it follows that $\mathfrak{B}$ is a total order and B has a countable order dense subset (Appendix A), and hence, $\mathfrak{A}$ is isomorphic to $\mathfrak{B}$ (5.4.6, U1).

Given the definitions of $\mathfrak{A}$ and $\mathfrak{B}$ and the finding of structural preservation between them, we have a solution to the systemic component of the non-uniqueness problem. The mapping component is solved as it was for the peripheral burst code, and for the same reasons. Regression analyses and information calculations are ultimately attempts to find reliable correlations between events, and there are indeed reliable correlations between PSFP and stimulus frequency. It is reasonable to assume that neurons in subpopulation-1 have the teleofunction of covarying with ambient energy states at the periphery, and more specifically, with superficial flutter-vibrations at the respective receptive fields for each neuron. As above, the function that best describes the correlation is the simple $r_2(x) = x$, where if the stimulus frequency is x Hz, then the PSFP will be x Hz. Note that $r_1$is distinct from r_2: the first is a function from frequencies to burst rates, while the second is a function from frequencies to PSFP. Since neurons in subpopulation-1 of S1 have the teleofunction of covarying with peripheral energy states according to r_2, it follows that according to SPT r_2 is the representation function.

Finally, given that the stimulator caused the neural activity in S1, we can assign f-referential content, via causal history, as the stimulator. Putting f-predication and f-reference together, the

representational content of the periodic, temporal organization of neural spike trains in subpopulation-1 of S1, assuming its PSFP is, say, 25 Hz, is *the stimulator is vibrating at 25 Hz*. Or, we may say that neural firing here represents the vibratory rate of the stimulator as being 25 Hz. As above, the independent vehicle hypothesis and its related pre-theoretic method of content-identification determined the same thing, and this is what is to expected, if SPT is true.

7.5.2.3 Positively and Negatively Sloped Rate Codes in S2, PFC, VPC, and MPC

Firing rates in subpopulation-2 of S1, S2, PFC, VPC, and MPC, during the base and comparison periods of the vibrotactile discrimination task, are vehicles of sensory representation. This was established by the vehicle hypothesis. We pre-theoretically identified the content of the burst-code and periodicity-code vehicles in terms of their discovered correlations. I propose to do the same here, and the reason is the same: In the absence of a strong reason otherwise, and given the vehicle hypothesis, this is the only plausible suggestion. As above, the discovered correlations are indicative but not constitutive of representational content.

Consider subpopulation-2 of S1, where we find aperiodic, stimulus-dependent firing rate responses. In this subpopulation of S1, firing rate, but not periodicity, changes as a function of stimulus frequency. Let's define our relational systems with the usual nomenclature.

$\mathfrak{A}$ is the totally ordered stimulus relational system, whose domain, A, has a countable order dense subset. $\mathfrak{B}$ is the physiological relational system, whose domain, B, consists of firing rates ordered by the empirical, total ordering, greater-firing-rate relation. Given the same idealization assumptions for measurement discussed above, we get the result that B has a countable order dense subset and is totally ordered and hence, $\mathfrak{A}$ is isomorphic to $\mathfrak{B}$, thus solving the systemic component of the non-uniqueness problem. What about the mapping component?

Here, as above, it is reasonable to take the reliable covariation between physiological states and energy states at the periphery of the organism, coupled with the survival-conducive behavior made possible by sensory systems in general, as evidence that the neurons in question have the teleofunction of corresponding to those energy states. The mapping function that describes the correspondence between firing rate and stimulus frequency, like that between burst rate or PSFP and frequency, is discovered experimentally. As described in (Salinas et al. 2000, 5506), the typical relationship between rate and frequency in S1 is described by:

$$r(s) = 22 + 0.7s + \sigma\epsilon,$$

where *s* is stimulus frequency, $r(s)$ is rate described as a function of frequency, ϵ is noise with zero mean and unit variance, and σ is the standard deviation of the mean firing rate. For our purposes here, we should delete the final noise term, since noise is, by definition, not a signal[127]. Thus, the function $r_3: A \rightarrow B$ that describes the correspondence relation, which is the neuron's teleofunction to bear to peripheral energy states, is this:

$$r_3(s) = 22 + 0.7s.$$

For these neurons, baseline firing is at 22 spikes/sec, and this increases linearly with stimulus frequency, with a slope of 0.7. Since these neurons bear the teleofunction of corresponding to peripheral energy according to r_3, i.e., they exist and have the properties they do because they covaried with energy according to r_3 in the environment of evolutionary origin, it follows that, according to SPT, r_3 is the

[127] Noise is an important conceptual issue that must be dealt with. We'll discuss this in chapter 8.

representation function, and thus determines f-predicative content. Causal history determines the f-referential content as the experimental stimulator.

For these neurons, firing at a rate of, say, 50 spikes/sec does *not* f-predicate having a frequency of 50 Hz. Rather, a rate of 50 spikes/sec f-predicates having a frequency of 40 Hz. Given that the stimulator caused this neural firing, we can say that, according to SPT, the representational content of a neuron in supopulation-2, firing at 50 spikes/sec, is *the stimulator is vibrating at 40 Hz*. In agreement with this, the vehicle hypothesis and its associated method of identifying content in terms of the discovered correlations say the same, thus providing further empirical support for SPT.

In general, we are going to find different representation functions in different populations, even if the relational systems involved are the same or similar. For example, while neurons in subpopulations-1 and 2 both have the teleofunction of covarying with stimulus frequency, subpopulation-1 is supposed to be more closely phase-locked, so that its temporal structure covaries with frequency. On the other hand, subpopulation-2 begins to abstract the temporal structure of the stimulus frequency out of the representational code. Using an aperiodic version of the task, the neurons in S1 that covaried according to their rate with the periodic frequency (i.e. subpopulation-2), also covaried, presumably according to r_3 (this specific information was not reported in the data), with the aperiodic stimulus frequency (Salinas et al. 2000; Hernandez, Zainos, and Romo 2000). Thus, we should conclude that subpopulation-2 has the teleofunction of covarying with frequency, regardless of whether it is periodic or aperiodic, according to r_3.

The specific linear equations describing the discovered correlations in S2 and more central areas have not, to my knowledge, been published, although the differences in the slopes of different subpopulations have. What deserves special mention in these areas, as distinct from S1, is that they each have specialized subpopulations with opposite slopes. Conceptually, this makes no difference and

is easily accommodated within the structural preservation framework. Practically, the difference is as follows. While the positively sloping population might have its correlation function look something like this,

$$r_3(s) = 22 + 0.7s,$$

the negatively sloped population might have its correlation function look like this:

$$r_4(s) = 65 - 0.5s.$$

I have stipulated r_4, but let's assume for the sake of the argument that r_4 describes the negatively sloped subpopulation of S2. The vehicle hypothesis determines that rate in S2 is a vehicle of representation, and its associated method which uses correlations as indicative but not constitutive of representational content identifies the content of these neurons in accordance with the correlation described by r_4.

SPT would describe the situation as follows. From the reliable covariation coupled with the utility of sensory systems in general, we conclude that those neurons have the teleofunction of covarying, according to r_4, with flutter frequencies at their respective receptive fields. From there, we conclude that r_4 is the representation function for that population, and thus have solutions to both the systemic and mapping components of the non-uniqueness problem, and thus an analysis of f-

predication. F-reference is determined as above, via causal history, as the experimental stimulator. As a result, firing at 50 spikes/sec in the positively sloped population f-predicates the property of vibrating at 40 Hz, whereas the same firing rate in the negatively sloped population f-predicates the property of vibrating at 30 Hz. In both cases, the f-referent remains the same, as the experimental stimulator tip. The difference is that neurons in different populations have the teleofunction of corresponding, via different correspondence relations (i.e. different representation functions), to flutter frequencies at the fingertip.

While the independent method clearly identifies the content of each of these neurons as the vibratory rate of the stimulator, since SPT appeals to causal history to determine f-reference as the stimulator tip, SPT must deal with the causal chain problem.

The causal chain problem arises in every case, but it seems more acute in the more central areas than in the periphery. Is it arbitrary to claim that firing rate activity in subpopulation-2 of S1 f-refers to the stimulator, rather than, say, activity in subpopulation-1 or in the peripheral afferents? Recall from 6.4 that the causal chain problem is really another version of the systemic component of the non-uniqueness problem. The problem involves determining which of the potentially infinite number of relational systems is the one of interest, in an objective, non-relative and non-arbitrary fashion. We begin with causal etiology, which rules out things like vibrating stimulator tips on the other side of the world. Then we consider structural preservation: there is no reason to assume that there will be structurally preserved relational systems at every link in the causal chain from content to representation. For the causal chain problem as it arises for the peripheral afferents, this consideration is enough to rule in the frequency of the stimulator tip, and rule out things like ion channel opening and closing, which do not obviously constitute a relational system that structurally preserves the physiological relational system of interest. Recall that we can always stipulate a relational system that

has whatever properties we like, but that is irrelevant. However, at least for the neural activity in S1, causal etiology and structural preservation alone are not enough to non-arbitrarily weed out, for example, neural activity upstream of S1, such as in the spinal cord.

As discussed in 6.4, teleology plays a critical role here as well. That which confers survival advantage, at least in this case, is not the covariation of neural activity in S1 with other neural activity. Rather, by covarying with energy states at the periphery of the organism, in well-defined ways, distinct neural mechanisms can use that activity in S1 to perform transformations and computations which ultimately result in behavior of the organism that is appropriate to the environment. More specifically, those later neural mechanisms can use that activity in S1 to generate behavior that is appropriate to the energy changes at the fingertip. As a result, we can non-arbitrarily claim that states of the stimulator tip constitute the non-physiological relational system, which the representation function connects to the physiological relational system. Further, we can non-arbitrarily claim that the neural representations f-refer to the stimulator tip, not to intermediate points in the causal chain. Similar considerations apply when considering sensory representations in S2, PFC, VPC, and MPC.

Finally, while the monkeys are highly successful in this task, indistinguishably so from humans, they do occasionally make errors. When this occurs, there is a correlation between standardized measures of firing rate in S1 and S2, with behavioral error. For example, if the monkey presses the medial button, signaling that the comparison was higher when in fact it was lower than the base, it is likely that the firing rates of its neurons in S1 and S2 are less than they would have been, had the animal made an accurate discrimination.

What would the vehicle hypothesis and its associated method of content identification say about this? First, we already accept the representation and vehicle hypotheses as well as the correlation-as-an-indicator method of identifying representational content, so these are not at issue.

Neither is the claim that the animal has made a mistake, nor that, when the animal successfully discriminates, that the neurons' representational content can be identified in accordance with, say, r_3. All that is at issue is what, if anything, the independent method of identifying representational content would say here, and it seems clear that we should say that this is case of neural *error*. As with the identification of representational content in general, in the absence of any overriding reason to the contrary, it is implausible to claim anything else. The implausibility argument is particularly strong here because we already accept that when the animal correctly discriminates, that its neural firing *correctly* represents the vibratory rate of the stimulator. Thus, in the absence of a theory of representation, we should independently identify this as a case of neural error. If SPT is true, then we should expect SPT to do so as well.

To see what SPT would say about this case, assume that r_3 is the representation function for neuron *n*, and that the stimulator is vibrating at 40 Hz. Given that frequency, *n* has the teleofunction of firing at a rate of 50 Hz. However, *n*'s firing rate, assume, is 60 Hz. This neural activity then f-predicates 54.3 Hz vibration, and f-refers to the stimulator tip. Thus, the firing rate of *n* has the representational content, *the stimulator tip is firing at 54.3 Hz*. However, the stimulator tip is not vibrating at 54.3 Hz, but at 40 Hz. Thus, *n*'s activity is false or in error. Like the pre-theoretic method, SPT identifies this as a case of neural error.

If only one or two neurons represent falsely by f-predicating the wrong property of the stimulator tip, the animal's behavior as a whole will likely be unaffected. But as the number of neurons in error begins to mount, it becomes increasingly likely that the animal will behaviorally signal in error. And, as found by Romo and colleagues, this is exactly what happens. If the animal has made a mistake, it is significantly likely that its neurons in S1 and S2 will be firing at a rate significantly higher (or lower) than they should have been, given that frequency. Thus, independent identifications of several different

kinds of sensory or indicative, neural representations, including neural error, agree with SPT's analysis of those same findings.

7.5.3 Motor Representations

Primary motor cortex, or M1, is one of the key brain areas responsible for directing behavioral output. Let us see how the present analysis applies to M1.

According to the vehicle hypothesis, firing rate in M1 is a vehicle of representation. M1 is principally a motor area, and is the primary area responsible for directing the animal's movement. Neural activity in M1 is essentially silent during the sensory and mnemonic components of the task in the base, delay, and comparison periods (Romo, DeLafuente, and Hernandez 2004). Its activity begins late during the comparison period (Romo, Hernandez, Zainos, Brody et al. 2002; Romo, Hernandez, and Zainos 2004).

Given the vehicle hypothesis, the timing of neural activity in M1 during the task, and the fact that M1 has been well established as a motor area, we should pre-theoretically identify the contents of its states as being behaviors, not, for example, peripheral stimuli. When the animal correctly discriminates and when $f_2 > f_1$ (i.e., the comparison is greater than the base), the animal presses the medial button, and when $f_1 > f_2$ the animal presses the lateral button. Firing rate in M1 can be characterized in terms of its correlation with either $f_2 > f_1$ or $f_1 > f_2$, and thus, we should use these correlations between rate and the inequalities, *and* the inequalities and behaviors, to identify the representational contents of firing rate as something like a motor command to press the medial or lateral button. As with each other type of neural representation, given the vehicle hypothesis and the discovered correlations, and in the absence of overriding reasons to the contrary, it would be implausible to identify the contents of the vehicles in M1 as being other than motor commands to press

the medial or lateral button. In chapter 8 I will discuss some reasons to the contrary, where I provide an alternate interpretation of the pre-theoretic identification of representational content as representing the difference, $f_2 - f_1$. I argue there that these are not overriding reasons, and we should accept my initial characterization. In the meantime, let us assume this initial, pre-theoretic identification of content. If SPT is true, then we would expect its analysis of these states to determine both that they are representations (this is the metaphysical component of SPT), and that their contents involve something like motor commands to press the medial or lateral button (this is the content component of SPT).

To begin, note that the teleofunction hypothesis, which will simultaneously help to determine which is the representation function as well as answer the metaphysical question (is this a representation and if so what kind?), is not going to be as straightforward as it was for the sensory representations. The motor output of pressing the medial versus lateral button in response to a comparison of two vibrating stimuli is learned, not evolved. Nonetheless, the animals do achieve high accuracy levels, and a reasonable teleological argument can be made on these grounds. The monkeys have learned that pressing the medial button when and only when the comparison stimulus is higher results in the acquisition of juice, and mutatis mutandis for the lateral button. Further, after learning, certain neural activities have come to be regularly correlated with the muscular motions associated with medial and lateral button-pressing. We may then conclude that the consumers of the neural activity under question (i.e. the neural mechanisms downstream of M1, in the basal ganglia, cerebellum, spinal cord, and motor neurons at the periphery) have the teleofunction of producing the state of affairs corresponding to the motor plan. Or in other words, if the motor plan says "my right arm is pushing the medial button", then the consumers of that motor plan have the teleofunction to make that true. Hence, we may at least provisionally conclude that, according to SPT, the neural activity to be discussed below constitutes imperative representations (i.e. motor plans).

For simplicity we will consider the activity in M1 which correlates with the difference, $f_2 - f_1$, via distinct subpopulations preferring $f_2 > f_1$ or the opposite.

Let us define our relational systems as follows. We begin with the positively sloped subpopulation, which "prefers" $f_2 > f_1$. The Romo group has not published the actual equations describing the relationship of firing rate to $f_2 - f_1$, but for concreteness I stipulate one, as follows. Notice that $a_1 = -a_2$, and that the constant is the point at which the function crosses the *y*-axis. Thus, if $f_2 = f_1$, the neuron will fire at the constant rate.

$$g_1(f_1, f_2) = -2f_1 + 2f_2 + 44$$

Let A = the set of firing rates, ordered by the greater-firing-rate relation, as above. $\mathfrak{A} = \langle A, >_A \rangle$ is a total order, and its domain has an order dense subset. It is therefore isomorphic to $\mathfrak{N}^+$, which allows us to create a function from rates to the nonnegative reals and then use its inverse to create range equivalence classes (5.4.6). Given this, we'll partition A as follows. If $a \in A$ is between 0 and 44 spikes/sec, then $a \in A^-$. If $a \in A$ is greater than 44 spikes/sec, then $a \in A^+$. I've chosen the +/- notation to partition A because of g_1; 44 spikes/sec is the point where g_1 crosses the *y*-axis. Below that, $f_2 - f_1$ is negative (f_1 is greater) and above that $f_2 - f_1$ is positive and f_2 is greater. Let $A^{RE} = \{A^-, A^+\}$. We'll define $>_{AR} = \{\langle A^+, A^- \rangle\}$, and finally, $\mathfrak{A}^{RE} = \langle A^{RE}, >_{AR} \rangle$.

B has only two elements; B = {is pushing the lateral button, is pushing the medial button}. (I'll abbreviate these two elements as *L* and *M*.) Now let's define a relation to order B, as $R = \{\langle M, L \rangle\}$. $\mathfrak{B} = \langle A, R \rangle$.

Getting to the representation function will take several steps. First, we define $r_5: A^{RE} \rightarrow B$, very simply, as $r_5(A^-) = L;\ r_5(A^+) = M$. The function r_5 defines an isomorphism from $\mathfrak{A}^{RE}$ to $\mathfrak{B}$. However, we need to connect firing rates, not equivalence classes, to the members of B. We won't get a homomorphism from $\mathfrak{A}$ to $\mathfrak{B}$, since $>_A$ is a total order, not a weak order. Thus, it won't be the case that $a >_A b$ iff $f(a)Rf(b)$. However, this is where another of Swoyer's relaxations will come in handy for defining structural preservation (see 5.4.5 and Appendix A).

Define $\mathfrak{A}^{\Delta/\Psi} = \langle A, =, >_A \rangle$, and let $\mathfrak{B}^{\Delta/\Psi} = \langle B, =, R \rangle$. Now we'll define our representation function $r_6: A \rightarrow B$ as follows. If $a \in A^-$, then $r_6(a) = L$. If $a \in A^+$, then $r_6(a) = M$. From here, we note that r_6 defines a Δ/Ψ -morphism from $\mathfrak{A}^{\Delta/\Psi}$ to $\mathfrak{B}^{\Delta/\Psi}$. Let the identity relation be an element of Δ, and let $>_A \in \Psi$, and further, there are no other elements of Δ or Ψ. Note that $r_6(a)Rr_6(b)$ implies that $a >_A b$, but the implication does not go the other way. Hence, r_6 counter-preserves but does not preserve $>_A$. On other hand, $a = b$ implies that $r_6(a) = r_6(b)$, hence, r_6 preserves identity. Since these are the only respective elements of Δ and Ψ, r_6 preserves all of the relations in Δ and counter-preserves all of the relations in Ψ, and hence defines a Δ/Ψ -morphism from $\mathfrak{A}^{\Delta/\Psi}$ to $\mathfrak{B}^{\Delta/\Psi}$, and this gives us structural preservation between $\mathfrak{A}^{\Delta/\Psi}$ and $\mathfrak{B}^{\Delta/\Psi}$.

Without the jargon, we've mapped every firing rate from 0 to 44 spikes/sec to something like, "push the lateral button", and every rate from 44 spikes/sec and up to something like, "push the medial button". With respect to f-predication, we conclude the following. Since r_6 (i) defines structural preservation between $\mathfrak{A}^{\Delta/\Psi}$ and $\mathfrak{B}^{\Delta/\Psi}$, (ii) is the correspondence relation between neural states and muscle states that it is the teleofunction of representation-consumers to produce, and (iii) maps 55 spikes/sec to *is pressing the medial button*, it follows that, according to SPT, neurons in the positively sloping subpopulation in M1, firing at 55 spikes/sec, f-predicate the property of pressing the medial button, to whatever they f-refer to.

F-reference is determined by causal history. As discussed in 5.5, imperative representations f-refer to what they caused, not to what caused them. The neural activity from M1 currently under consideration causes changes in the contraction levels of the various muscle groups of the animal's right arm. It also causes changes in neural activity downstream, such as in the spinal cord, but we can use an argument from teleofunction, parallel to the sensory version, to avert this causal chain objection: It is in virtue of the fact that M1 controls muscle contractions that proves to be survival-conducive, hence, while there are other links in the causal chain from representation to content, it is not arbitrary to claim that firing rates in M1 f-refer to muscles in the arm, not neural activity in areas between M1 and those muscles. Assuming that this 55 spike/sec neural activity does in fact cause changes in the arm's position, SPT would attribute the following content: *my right arm is pressing the medial button*. The negatively sloping subpopulation in M1 can be analyzed according to SPT in a similar fashion.

In contrast to previous examples, the representations are in the domain, not the range, of the representation function. I argued in 5.4.8 that the only empirical consideration relevant to determining whether the representations are in the domain or range had to do with the size of the domains of the two relational systems. I considered a case where we might plausibly argue that there are more elements in the stimulus relational system than in the physiological system, and on that basis should consider the representations to be in the domain of the representation function. This way it is possible to define it as a function. In the motor case, the opposite occurs: There are more firing rates than muscle position categories (here there are only two). On these grounds then, we should conclude that the representations are in the domain of the function, not the range.

Error is described as it was for sensory representations, both by the pre-theoretic identification and by SPT. I'll consider the positively sloping population, whose activity accords with g_1. Suppose that $f_2 = 50$ Hz, and $f_1 = 48$ Hz. Then $g_1(f_1, f_2) = 48$ spikes/sec, which is very close to the "zero" point of

44 spikes/sec. At this point, downstream noise is more likely to have a bigger effect on the motor neurons than it would if firing rate were, say, 84 spikes/sec, as it would be with a 20 Hz difference in frequency. According to the appropriate representation function, r_6, 48 spikes/sec maps to medial button-pressing. Imagine however that the arm presses the lateral button. In this case, both SPT and the pre-theoretic identification would claim that we have error. It would be more colloquial to say that the motor plan was not carried out, or that the command was not followed or not satisfied, than to say that the neural activity "is false". However, from the viewpoint of SPT, it is exactly the same as representing falsely. Thus, when the pre-theoretic method identifies a case of a motor or imperative representation not being satisfied, SPT accords with that account, providing further empirical support for the theory.

7.5.4 The Role of the Empirical Literature

In this subsection I clarify my arguments above, discuss some objections, and discuss the role of the empirical literature.

First, SPT does not predict that the correlation and hence representation function between neurons and stimulus frequency will be described by r_1, r_2, or r_3. If some other relationship had been discovered, say that baseline firing is at 30 instead of 22 spikes/sec, SPT would have been consistent with that, due to the non-uniqueness of isomorphisms. That does not show that SPT is a non-empirical theory or that the findings described above do not provide evidence for the theory. SPT is not a theory about the brain and so makes no predictions about the brain. In particular, it does not predict how neural states will correlate with energy states, if at all.

Rather, SPT predicts that, if we were to find representations with content *C*, those vehicles would be related to *C* by way of causal history and structural preservation. Additionally, if SPT is true,

then the representation function defined by SPT will match up vehicles with contents in the same way that the independent identification matches vehicles with their contents. SPT predicts that vehicles of representation will be an *organized system*, and it predicts that what those vehicles represent will also be an organized system, and further that the two systems will structurally preserve one another.

In 7.5 I claim that that is exactly what has happened, yet it need not have. There is no a priori guarantee that the independent identification of vehicles and their contents would be related by causal history, for example. It could have turned out that our independent characterization identified the content of firing rate in S1 as a stimulator tip on the moon, or an apple in Japan, or the number three. There is no a priori guarantee that the independently specified relational systems would have been related via structural preservation, since structural preservation is not guaranteed to exist. It could have turned out that the systemic component of the non-uniqueness problem for SPT determines a pair of relational systems unrelated to the pre-theoretic identification method. It could have turned out that the independent method identified the representational content of firing rate as the color of the stimulator tip, or as the metal out of which it is constructed, or as 4.9% of its vibrating frequency, except when the frequency is 25 Hz. Any of these and more could have occurred, but they did not. There is no a priori guarantee that the vehicles of representation that we found would be related to their contents in the way that we have discovered them to be. This is a significant finding, and it does indeed constitute empirical confirmation of SPT.

One may object with any number of complaints, and I would like to air them to be sure the dialectic is clear. One may object that these are not vehicles of representation, or that we should not identify their contents as such, or that the vehicle hypothesis is insufficiently supported. If this objection is sound, then I certainly concede that I have not provided empirical evidence of SPT, because doing so depends on a prior and independent method of identifying representations and their contents.

However, this would simply be an objection to the vehicle hypothesis or its method of identifying contents. It would not be the more general objection that no theory of representation can be supported by evidence.

However, to deny the vehicle hypothesis, one must provide a better explanation of the discovered correlations *as well as* the animal's behavior; the vehicle hypothesis conjoins itself to the representation hypothesis in extending and improving on the explanation of the animal's behavior. The vehicle hypothesis provides explanatory unification with the representation hypothesis. To provide a better explanation, my objector would then need to deny the representation hypothesis in order to avoid the explanatory dis-unification charge. But that means denying the humans and Occam's razor argument, which I take to be extremely implausible, since it entails either denying that I have working memory, or making an arbitrary distinction between macaques and me, even though our behavior is indistinguishable. As I said, that is implausible in the extreme.

A second objection would be that I have begged the question by using covariation as an "indicator", and then importing the covariation into my theory of representation, in this way guaranteeing that whatever I would find would satisfy my theory. This objection misunderstands both SPT and the methods used for identifying representations in the brain. I have noted that we must distinguish between using covariation as an indicator of the presence of a representation and of its content, and identifying representation or representational content with covariation. Representation *is not covariation*, and it cannot be covariation, for the reasons discussed previously. Most substantially, and decisively, I argued in 5.2.2 that no representation has content, either veridical or mistaken, except insofar as it both points to something as well as says something about what it points to. Without *both* f-reference and f-predication, or something like them, there is no truth or error and hence, no content.

Covariation does not have the resources to explain both f-reference and f-predication. Thus, representation is not covariation.

I have not assumed a covariance theory in identifying representations. Rather, I used the discovered correlations as a fallible indicator of representational content. The assumption that covariation is constitutive of content is a distinct, and false, claim. Further, SPT is not a covariance theory of representation. I have repeatedly noted that SPT repudiates covariation as constitutive of either being a representation, or of having representational content. It is consistent with (but not implied by) SPT that covariation can play an evidential role for certain fallible statements about teleofunction. However the grounds on which we believe that device *D* has teleofunction *T* are irrelevant to SPT. All that matters is that it *has* that teleofunction.

A third objection is that, whatever content the pre-theoretic method determined firing rate to have, SPT is malleable enough that it could have been made to fit that content. Especially given the non-uniqueness of structural preservation, this might seem plausible. This objection, like the objection above, misunderstands SPT. First, causal history determines f-reference on SPT, and causal history is not malleable. Should the pre-theoretic method have identified something that is not causally linked to the representation as that representation's content, SPT would not be consistent with that. Second, even given the f-referent as causally linked, SPT must solve the mapping component of the non-uniqueness problem: It must say which of the numerous mapping functions is the representation function, and it does this on the basis of teleology. If the pre-theoretic method identified the content of the firing rate as something like 4.9% of the frequency except when it is 25 Hz, SPT would not be consistent with this. While almost any arbitrary mapping function can be constructed that connects the two relational systems, SPT *chooses only one*. If any other mapping function is chosen by the pre-theoretic

identification, then that is not consistent with SPT. What follows is a fourth objection, quoted from comments by Michael Levin on this chapter:

> The question of whether SPT is empirically supported comes down to the question of whether there is some independent criterion for the existence (and content) of representational states which are then predicted by SPT. The alternative view, of course, is that SPT is a good analysis of our criterion for identifying representational states and ascribing them content … What you say is entirely consistent with [the independent method for identifying content] tacitly employing SPT. For instance, we see the monkeys responding in systematically adaptive ways to varying pressure on their fingertips, assume there is some internal state correlated with the pressure, and take each such state to "mean" that a certain pressure is being applied to the fingertips. We tacitly use SPT to postulate representations.

This objection is another version of the circularity worry discussed in the second and third objections above, and the reply is the same, although I will emphasize certain points. I began by defending the representation hypothesis, ultimately on the basis of the indistinguishability of human and macaque behavior, and on the undeniable starting point that I have short-term memory. This established the representation hypothesis without any underlying theory of representation.

My defense of the vehicle hypothesis depended on several factors, only one of which is the correlation between internal states and pressure on the fingertip. In addition, I appealed to the causal efficacy of those states for producing behavior in this particular task, as evidenced by the lesion and microstimulation studies. I also appealed to the correlation between behavior and those internal states,

as well as the neurometric and psychometric curves and discrimination thresholds. Each of the latter two elements is a way of relating animal behavior or neural activity to properties of the stimulus, other than a simple correlation between varying pressure and neural activity. And, what is certainly an important element, I appealed to the reliable covariation between neural activity and pressure on the fingertip. Given all of these various factors, and given the representation hypothesis, the simplest and most unified explanatory approach is to explain all of them in the same way, and thus conclude that these states are representations. This establishes the vehicle hypothesis through a combination of various factors; however, not one of them involves an appeal to teleology, nor, particularly, isomorphism or structural preservation. Additionally, while I appeal to the reliable covariation between neural events and stimulus frequency, I do not appeal to causal *history* between them. I do appeal to the causal efficacy of these states in producing successful discrimination *behavior*, but, again, I do not appeal to the causal history between neural events and pressure on the fingertip.

Finally, given both the representation and vehicle hypotheses, I used the discovered covariations to pre-theoretically determine representational content; that is, I used those correlations as indicative but not constitutive of content. However, covariation, in any form, is irrelevant to SPT, as I mentioned above. As a result, in using covariation, I *couldn't have been* tacitly employing SPT, because covariation plays no role in the actual theory[128]. That covariation can and should be used to help fallibly identify the teleofunctions of various biological devices is completely irrelevant to SPT. Thus, this version of the circularity worry, like those mentioned above, is avoided.

A second aspect of Levin's concern is that SPT is really a clarification of the commonsense concept of representation ("SPT is a good analysis of our criterion for identifying representational states

[128] Rather, I should say that covariation plays no role in the theory *proper*. In other words, covariation does not determine f-reference, f-predication, or establish the metaphysical component. But it does play an evidentiary role, for example, in helping to discover various teleofunctions.

and ascribing them content"). By analyzing the everyday concept, I arrived at SPT, thus elucidating our criterion for identifying representational states and ascribing them content. If this is the case, even the claim that I've repeatedly stated to be undeniable, that *I* have short-term memory, would be made on the basis of assuming SPT. Since the covariation-as-indicator method of content identification depends on the defense of the vehicle hypothesis, and the vehicle hypothesis depends on the defense of the representation hypothesis, which ultimately depends on the claim about my own representational states, this would serve to collapse the entire structure. If SPT is but an analysis of our ordinary concept, and if SPT is the basis upon which the commonsense use of 'representation' licenses the claim that I have representations (via short-term memory and sensations), then this would be a troubling objection. But it isn't.

There is no basis for claiming that SPT is an analysis of the "ordinary" concept of representation. I made a similar point earlier in the context of a discussion of Ramsey (2007), in 1.3.1. The only authority on the ordinary use of any term is the dictionary, and the various dictionary senses of 'represent' have notions of surrogacy and pointing in common, but nothing else. Certainly, the commonsense, ordinary language concept of representation has nothing to say about systems of representations structurally related to each other, connected via causal history. Neither does the ordinary concept of representation, if there is such a unitary thing, say anything about teleofunction and the role it plays in determining which of the very many different systems and mappings are involved in representation. To claim that SPT has really been a part of our ordinary notion of representation all along seems fairly incredible.

There might be a worry about a fallacy of equivocation here. Perhaps the worry is that I cannot move from the undeniable claim that I have memory, to the representation hypothesis as it applies to me, or, to my use of 'representation' in this context, since the former claim is based on ordinary

language and the latter on a technical use. However, I carefully circumscribed my use of 'representation' in chapter 1, so that it accords, to a reasonably sufficient degree, with the ordinary language sense of surrogacy and pointing (although this does not imply that the ordinary language sense assumes the complicated machinery of SPT). I also made it explicit that if a state is representational in cognitive psychology's sense (e.g., as short-term memory is), then it is representational in my sense. Thus, I appeal to the cognitive psychologist's use of 'short-term memory', which is in reasonable accordance with the ordinary sense of surrogacy and pointing, to make the claim that I have short-term memory. From there, I move to the claim that I have representations, using 'representation' in the sense that I've outlined as being the common core of various theoretical senses. This justifies the representation hypothesis with respect to me, and it does so without assuming SPT. The worry that SPT is but an ordinary-language analysis, which would collapse the entire structure of my arguments for the independent, pre-theoretic method of content identification, is unfounded.

I will briefly discuss three more clarifications. For the simple domain under consideration, it might seem that a covariance account can do the trick, and the more complicated theoretical machinery need not be introduced until we start dealing with more complex cases of representation. However, covariance cannot do the trick even for the relatively simple cases under consideration here, for the numerous reasons mentioned earlier, not least because it cannot handle the necessary distinction between what a representation is about (i.e. f-reference) and what a representation "says" regarding whatever it is about (i.e. f-predication). Additionally, note that I have not used the empirical literature as an argument *against* any theory of representation, but only *in support of* SPT.

What if another theory is compatible with the findings? Can the empirical literature arbitrate? Should there be another theory of representation that is equally compatible with the pre-theoretic description of content, then either conceptual considerations will arbitrate, or the two theories will be

equally supported by both conceptual and empirical considerations. This does not remove any of the support for SPT gained through the empirical literature.

Finally, if someone does not accept SPT, should the empirical work and claim 1 sway them? It is important to recognize that all of my arguments in support of SPT are ampliative. Thus, each additional argument serves to provide further support for my theory. The chief support is to be found in earlier chapters, where I argued that SPT satisfies certain adequacy conditions that other theories do not, and most importantly, allows for a naturalistic explanation of representational content and error that other theories do not. The role of claim 1 is to provide additional support, over and above what is already in the conceptual work, but whether claim 1 succeeds is dependent on whether the vehicle hypothesis succeeds. So it's hard to say whether the additional empirical work, still based on ampliative arguments, should be the final straw that tips the scales in my favor. I do contend however that the vehicle hypothesis does succeed, and that claim 1 provides additional support on behalf of SPT.

The relation of theory to empirical discovery in the case under consideration is two-fold. With respect to claim 1, if the vehicle hypothesis is successful, then we can compare representational vehicles and their relations to what they represent with what SPT claims representation and representational content to be. This is what I have done in this chapter, arguing that there is a match-up between SPT's conception of representation and the independent, pre-theoretic method of representation-identification, thus providing empirical support on behalf of SPT. With respect to claim 2, appealing to the empirical literature can only serve to illustrate, clarify and refine SPT, but not confirm it.

The dual-approach methodology begins with a conception of representation, provisionally identifies states as representations in light of that conception, and then uses an analysis of those states in an attempt to clarify and refine the theory. The success of the dual-approach methodology and the

justification for using it depends on the prior adequacy of the conception of representation that we start with. I have approached claim 1 first, because claim 1 provides additional support for SPT, over and above the conceptual support of previous chapters.

In the following and final chapter, we will use the dual-approach strategy to clarify and refine SPT. I argue that doing so allows for several important refinements and clarification of my theory. Thus, examining how representation is *implemented* allows for a refinement of our conception of the *nature* of representation.

Chapter 8: The Implementation of Representation II - Clarifying and Refining Structural Preservation Theory

8.0 Introduction: Applying SPT to the Brain from the Perspective of Claim 2

Claim 2: Structural preservation theory is the best theory of representation. Therefore, if SPT determines that *X* is a representation, then we should conclude that *X* is a representation, *on the grounds that SPT implies this*.

From the perspective of claim 2, we do not confirm SPT, but we use an analysis of the empirical literature to illustrate, clarify, and refine it. My goal in this chapter is to do just that. I will apply SPT to the findings from the electrophysiological studies reviewed in chapter 7 and Appendix B, yet this time from the perspective of claim 2. In so doing I will clarify and refine several important conceptual issues with respect to the nature of representation. The structure of this chapter is as follows. I begin with a discussion of working memory. In 8.2 I provide an analysis of the cognitive/neural computational process involved in deciding which of two vibrotactile frequencies is greater. In 8.3 I address the induction of artificial percepts generated by intracortical microstimulation, and in 8.4, the issue of neural noise and how it affects a theory of representation. Then in 8.5 I clarify the roles of each component of the theory, in light of the refinements and illustrations provided by the empirical literature. I also discuss the causal efficacy of representations in this section. In 8.6 I discuss how this theory makes contact with the more traditional project in this field, why my theory is a competitor to Fodor, Dretske, Millikan, et al., and how to "scale up" the theory in order to apply it to less basic kinds of representations. Finally, in 8.7 I conclude the dissertation.

8.1 Working Memory

In 7.5 we discussed sensory representations and motor plans, but ignored everything in between. However, working memory plays a critical role in the animal's cognitive decision process. Sensory representations of the base frequency are converted to short-term memories, which are then compared to sensory representations of the comparison frequency. The outcome of this decision process is a motor plan.

Working memory presents a small challenge to the theory presented thus far, because it seems that we need a tense-modifier. That is, with respect to memory, we no longer want to conclude that neural activity represents that *a* is *F*, but rather, that *a* was *F*. As it stands, SPT does not have the resources for that. However, with just a bit of tweaking, we can see the broad outlines for how that story should go.

First, all non-imperative representations, including sensory representations, are representations of the past. For example, latency data in ms of sensory representations of the base stimulus from S1, S2, VPC, and MPC, respectively, are 41±08, 58±06, 61±10, and 135±22 (Romo, Hernandez, and Zainos 2004, 170, table 1). Thus, simply saying that memory is a "representation of the past" does not distinguish mnemonic from sensory representations. To account for this, we might relativize our time-scales to behaviorally or neurally relevant time periods, then assert that representations are mnemonic if their instantiation occurs *after* the event that they represent, on the understanding that *after* is relative to neurally or behaviorally relevant time-scales. In that case, everything else stays the same: the

representation f-predicates *F* of whatever *a* it f-refers to. However, colloquially, we would *describe in English* the representation's content as saying something like, *a was F*, rather than that *a is F*.

Importantly, this is a slight modification of the metaphysical component of SPT, not the content component. We have been working with Millikan's (1984, 100) teleological conditions on what it is to *be* a representation, on the understanding that further work will be needed. Thus far we have been thinking only in terms of indicatives and imperatives, but a theoretical framework for the differences between sensory, mnemonic, and motor representations seems to be an important addition. I've added the qualifier, 'seems', because it may also turn out that we do *not* need to make a sharp distinction between sensory and mnemonic representations. Perhaps we should simply say that all indicatives represent past events, and then we distinguish whether how far in the past the event occurred is behaviorally relevant or not. This is an issue for another day.

One difference for the working memory component of the discrimination task is that, in addition to positive and negative sloping populations like those found with sensory representations, these neural representations are also sub-categorized according to when they occur. There are early, persistent, and late neurons, which fire according to their respective representation functions throughout the delay period (persistent), early in the delay period, or late in the delay period (Romo et al. 1999). While this is an interesting experimental finding that helps to detail the inner computational workings of the brain as it solves the discrimination problem, it does not have any implications for explaining representational content.

8.2 Computation and Representation: A Comparison and Decision Procedure

8.2.1 Computation and Representation

The cognitive/neural computational procedure by which the animal compares two representations of stimuli and decides which is greater presents an interesting opportunity to clarify the relationship between computation and representation, and the role of SPT with respect to each. Additionally, in 7.5.3 I noted that we would return to the pre-theoretic identification of representational content for the motor representations in M1. Specifically, perhaps intuitions run in a different manner, and suggest that we should identify the content of these motor plans with something like the decision that, say, $f_2 > f_1$, or perhaps as a representation of the difference, $f_2 - f_1$.

To get a clear understanding of the theoretical underpinnings of the cognitive/neural processes of comparison and decision, we'll take a step away from the neural details, and just consider computation. In general, computation is thought of as the manipulation of syntactically structured symbols according to quasi-linguistic, syntactically defined rules. We can drop the linguistic trappings, and simply consider computation to be the process of taking a meaningful or content-bearing input, manipulating it according to some procedure, and generating a meaningful or content-bearing output. Mathematical functions are roughly analogous to computations. The function $f(x) = 2x$, for example, takes numbers as input, and generates numbers as output. However, nothing mediates between the input and the output. What physically instantiated computations do, and mathematical functions do not, is make use of some mediating *procedure*. With $f(x) = 2x$, there is nothing between (say) 2 and 4; f simply maps 2 to 4, and that is all there is to the function. By contrast, a machine that instantiated that function, or that performed that computation, must necessarily *do something* in order to move from the inputted symbol '2' to the outputted symbol '4'. That is the computational procedure that mediates between meaningful input and meaningful output.

In the vibrotactile discrimination task, the brain performs computations. Since computation involves the manipulation of content-bearing items, we should consider the procedure as moving from sensory *representations* to motor *representations*. If the sensory representations represent the comparison as lower, then the computational procedure should output a motor command to press the lateral button. Considered as a function (in the mathematical sense), the inputs include sensory and mnemonic representations of the stimuli, and the outputs are motor plans to press either the medial or lateral button. Considered as a computation, the inputs and outputs are the same. However, this is a physically instantiated computation, and so must necessarily make use of a *procedure*. The brain, qua computing machine, must *do* something in order to get from sensory representations to motor plans. The best way to consider the relevant sequence of events is this: physical stimulus → sensory and mnemonic representation of the stimulus (input to the computational procedure) → manipulation of sensory and mnemonic representations (the computational procedure itself) → motor plans, or, representation of the behavior (output of the computational procedure) → behavior. To understand the implemented computational process, we have to keep clear on the distinction between the *inputs and outputs* of the procedure, which are representations, and the *procedure itself*.

We'll begin with a brief review, recalling that there is a full rendition of each of these events (with a greater or lesser emphasis on different elements of them) in each of S2, PFC, VPC, and MPC (Hernandez, Zainos, and Romo 2002; Romo, Hernandez, Zainos, Brody et al. 2002; Romo, Hernandez, and Zainos 2004). During the comparison period, some neurons fire at a rate that covaries with the base frequency f_1, instantiating a mnemonic representation, while others fire at a rate that covaries with the comparison frequency f_2, instantiating a sensory representation. As the comparison period proceeds, neurons begin to fire at a rate that corresponds to some combination of f_1 and f_2 (that is, both a_1 and a_2 in the multiple regression analysis are significantly different from zero). Towards the end of the

comparison period, neural activity now correlates with neither f_1 nor f_2, but with the difference, $f_2 - f_1$.

In this case, $a_1 = -a_2$, and it is not the frequencies themselves that determine firing rate, it is the difference between them that determines rate. For example, a base and comparison of 40 and 20 Hz, respectively, will elicit the same response as 30 and 10 Hz, respectively. As we've seen several times before, there are specialized subpopulations. One subpopulation "prefers" the case where $f_2 > f_1$, whereas the other "prefers" $f_1 > f_2$. For the former population, as $f_2 - f_1$ gets (positively) larger, firing rate increases. For the latter, the opposite occurs: as $f_2 - f_1$ gets progressively more negative, firing rate gets progressively faster[129].

8.2.2 A Possible Disconfirmation of SPT

In 7.5.3 I argued that the pre-theoretic identification of content for the neural activity in M1 that correlates with $f_2 - f_1$ is a motor plan that represents an "intended" arm movement. Then I argued that SPT agrees with that initial characterization, thus providing evidence for the theory. Here, we will consider a different suggestion: When $a_1 = -a_2$, that neural activity represents the difference of $f_2 - f_1$, or the inequality $f_2 > f_1$. Or maybe we should say something like: that neural activity constitutes the neuron's/animal's decision that $f_2 > f_1$. I argue that this analysis should be rejected, and in so doing, I'll defend the manner in which I originally (pre-theoretically) characterized representational content for these states. Additionally, the subsequent discussion allows for a clarification of computation, representation, and the role that SPT plays with respect to each.

[129] It is unclear what the mechanisms are by which this transformation occurs, but one hypothesis mentioned in the literature is the subtraction hypothesis (Gold and Shadlen 2001). According to the subtraction hypothesis, the activity of neurons tuned in opposition to each other with respect to the same stimulus parameter can be subtracted, in order to reach a sensory "hypothesis". Notice however that this is a functional/computational level description; even if it is true it does not describe the mechanisms by which neural populations realize this subtraction process.

In chapter 7 I used covariation as an identifier of representational content; thus, in the absence of any reason not to do so here, we ought to identify the content of firing rate as that which it most obviously correlates with. When $a_1 = -a_2$ in the regression analysis, firing rate correlates with the difference, $f_2 - f_1$. On this initial analysis, the representational content of those neural vehicles is the difference, $f_2 - f_1$, or the inequality, $f_2 > f_1$.

If this is the correct description of the initial data, then we have a possible disconfirmation of SPT. When I applied SPT to this scenario in 7.5.3, I argued that according to SPT, the firing rate in M1 that correlates with the frequency difference has arm movements as its representational content, not an inequality or a difference between two frequencies. Given a possible recalcitrant observation, and given the Duhem-Quine thesis, we have at least three options: (i) re-describe or revise our initial characterization of the data (i.e., that firing rate in M1 represents an inequality or a relation among frequencies), (ii) check if there are unsupported auxiliary hypotheses operating in the background in need of revision, or (iii) abandon SPT as having been empirically refuted.

Unsurprisingly, I don't think we should abandon SPT just yet. Since SPT is otherwise well-supported, I propose instead to explain why we have the intuition that firing rate in M1 represents a difference, but why that initial intuitive judgment is mistaken. Thus, we have a case in which the most obvious correlation is *not* an indicator of content. Additionally, I will review the positive reasons I gave on behalf of my initial non-theoretic characterization.

Because this is such a critical point, I will say it again in a different way. However, note that this is not different from my characterization of the dialectic surrounding claim 1 in chapter 7. The basic assertion that I need for claim 1, and to support my claim that there is a possible disconfirmation here, is as follows. It is possible to identify a representation with content *C* in the absence of a theory of

representation. Or, it is possible to point to a thing and say, "that's an *F*", even though we do not yet have a deep understanding of what it is to be an *F* (or in the case here, we lack an understanding of what it is to be a representation *and* what it is to have representational content). I support this claim with my analogy to water, although the examples can be multiplied endlessly. Unless it is possible to begin with a pre-theoretic characterization of a thing as belonging in some particular category, it is impossible to ever make inductive generalizations, or to probe deeper into the nature of what things in those category *are*, or in short, to do science. But it is possible to do science, hence, in many cases, it is possible to simply point and say, "that's water", "that's a fish", and so forth, even in the absence of a theory of chemistry or evolution.

Representation *seems* like a different case because it seems like whether something is a fish, or is water, is something that we can just *see*, whereas something's being a representation is not observable, but must be inferred on the basis of a theory. But this is not a good distinction to make. We do not "just see" *that* a thing is an *F*. Perhaps we just see *F*s, but the characterizing judgment *that* a thing is *F*, say, is water, or is an apple, or is a representation, must be based in either a tacit or explicit manner of categorizing the world[130]. However, this does not render the observation that *that is an apple* a non-empirical datum. Similarly, it does not render the observation that *firing rate represents frequency* a non-empirical datum either.

For a different example, it seems like whether burst rate or periodicity correlates with peripheral energy states is clearly an empirical datum, whereas that firing rate is a representation with content *C* is not. However, these correlations are not directly observed, but are inferred on the basis of

[130] I accept the thesis that all perception is theory-laden, thus eliding the theory/observation distinction. Nonetheless this is controversial so I don't want to rest my case on it here. Here I make the slightly weaker claim that the characterization of any empirical datum which says, in effect, "that's an *F*" must be based in a tacit or explicit manner of categorizing the world. That "manner of categorizing" should be considered weaker than a *theory* in the traditional sense. However it still makes my point that even paradigm observation statements have some categorization judgment backing them.

numerous statistical assumptions, ultimately grounded in possibly ineffable assumptions about inductive reasoning. To get from what the scientists actually observe, which is a monkey pressing buttons, to the claim that firing rate in S2 correlates with stimulus frequency, we must pass through numerous theoretical assumptions and inferences. However, that rate correlates with frequency is in fact an empirical datum.

Since the "observed" correlations are empirical data, even though they are not directly observed but require inferences and a background theoretical framework, and since the observation that *a is an apple* is an empirical datum, even though that categorization judgment is based in a tacit manner of categorizing, why should we outright reject the assumption that *firing rate represents frequency* is, similarly, an empirically observable datum?

Surely, we may disagree about whether firing rate is a representation or about what it represents. In fact, given that we really don't know what representation is, we should expect initial disagreement here. But all that we are disagreeing about is the initial *characterization* of the datum. We are not disagreeing that it is or is not a datum. However, what we can do is provisionally identify firing rates as representations on the grounds that SPT implies this, and then shuffle back and forth between initial characterizations of the data and the theory. That is what I do in this chapter. In the previous chapter, I used independent arguments, namely, those supporting the representation and vehicle hypotheses, to identify representations without assuming a theory, in such a way that we could test what those independent characterizations said against what SPT has to say.

In this section we are considering a possible empirical refutation of SPT because, if the characterization of the data under consideration is correct, then that characterization does not bear out what SPT predicts. Namely, SPT has no way of making sense of a representation that refers to *both* f_1

and f_2. This holds whether we characterize firing rate as predicating a relation between the base and comparison frequency (e.g., that $f_2 > f_1$) or predicating that their difference is *x*. I cannot make SPT come out saying that "the difference between f_1 and f_2 is 5 Hz", because the subject term refers to both the base and comparison frequencies. Even if we grant that a difference can have causal powers, and I'm not sure that we should, nonetheless, the difference here is a difference between f_1 and f_2, and so we would need a way to get the representation, according to SPT, to refer to both frequencies. It would seem that SPT is stumped. We would need to either significantly revise SPT or abandon it, if this is the correct initial characterization. However it is not the correct characterization of the data. I will explain away the intuition that it is the correct characterization, and argue for my distinct initial characterization. Note once again that both of the initial characterizations, the one I support and the one I reject, are independent of a theory of what representation is or what determines representational content, and hence independent of SPT. The question here is: How best to initially characterize the data?

8.2.3 The Computational Rule

Without the linguistic accessories, computation is the process of generating a content-laden output, given a content-laden input, according to a rule. The best description of the vibrotactile discrimination task is as follows. Physical stimuli get transduced into sensory and mnemonic representations. A computational procedure ensues, whereby those representations are manipulated and altered, and a motor representation is produced as output. The motor plan then gets re-converted, from a representation of motor activity, into actual motor activity.

The intuition that the neural activity in question implements a cognitive decision, but is *not* a motor representation, is a result of the failure to keep clear the distinction between the inputs/outputs

of the computational process, which are representations, and the intervening computational procedure itself. The comparison of f_2 and f_1, which occurs when both a_1 and a_2 in the regression analysis are nonzero, is the computational procedure, mediating between representational inputs and outputs. The "decision" that $f_2 > f_1$ is, in part, an element of that procedure and so not a representation at all.

The rule which governs the computational procedure is something like this. "If the inputs represent f_2 as greater than f_1, then output a motor plan to press the medial button. If the inputs represent f_1 as greater than f_2, then output a motor plan to press the lateral button." The experimental literature has given us a glimpse into how this rule gets implemented in the brain. In S2, for example, no representation of f_1 persists throughout the delay period, yet during the comparison period, there are both sensory representations of f_2 and mnemonic representations of f_1. As the comparison period proceeds, neural activity then shifts, to where it is a function of both f_1 and f_2, and then finally, to a function of $f_2 - f_1$, split into positively and negatively sloping subpopulations (Romo, Hernandez, Zainos, Lemus et al. 2002). This description of neural activity is a description of the computational procedure itself. The activities that occur (temporally) in between the sensory representations and the motor plans in M1 are *not representations*.

One might ask: Doesn't this activity represent the computational rule itself? No. The computational rule is simply our description of the neural/cognitive activity that in fact occurs during this task. In virtue of producing the motor representation to press the medial button as output, given an input in which, say, f_1 is represented as being 20 Hz and f_2 is represented as being 40 Hz, the neural activity is *following* the rule. Another way to say this is that it *embodies* or *realizes* the computational rule in virtue of what it does: Given an input of *x* and *y*, the neural circuit will reliably output *z* (where *x*, *y*, and *z* are representations). In virtue of this, it follows, embodies, or realizes a computational rule. It does not follow from this, however, that the computational rule itself finds an explicit representation

anywhere in the brain[131]. Perhaps the cognitive "decision", as well as the other elements of the computation (such as the comparison) can be identified with a functionally defined state, in something like the traditional functionalist theories of mind (Putnam 1960; Lewis 1966; Armstrong 1968). I won't provide any discussion of this proposal here, as it will take us too far afield.

If the intuition under consideration is based solely on confusing computations with representations, the above reply is sufficient to clarify the issue. However, we may also consider the cognitive "decision" to be the output of the computational process. As such, it is a representation, and so however it gets initially characterized, SPT should accord with that characterization.

To the extent that the cognitive "decision" is in fact an *output* of the computational procedure and hence a representation, that decision just is the motor command to press one of the two buttons. If the decision that $f_2 > f_1$ is distinct from the motor plan to press the medial button, then we should be able to find them both: We should be able to experimentally tease them apart. One way of doing this would be to compare passive and active versions of the task. In the passive version the otherwise free hand is restrained, no discrimination is made, and no reward is offered, yet the stimuli remain the same. If the activity in question, which I claim to be a motor plan, is not what I claim but instead is a representation of the decision that $f_2 > f_1$, then we would not expect a difference between the active version, which involves a motor output, and the passive version which does not.

To my knowledge, this version of the discrimination experiment has not been performed. However, in the related categorization task (Salinas and Romo 1998), the authors found that, of the neurons in M1 that correlated with the learned stimulus categories of *high* and *low*, none of them did so during a passive version of that task. If this result is replicated with the discrimination version, then we

[131] It also does not imply that it doesn't. Maybe there is another brain area where we can find that rule explicitly represented. If so, that doesn't change any of the present discussion.

should provisionally conclude either that there is no distinction between the neural implementation of the motor plans and the cognitive decision, or at least that the neural activity in question *is* a motor plan (leaving open whether it is also a cognitive decision).

A confound in the above setup is the possibility that motor plans and discrimination decisions always co-occur, even if they are distinct. Hence, altering the experimental setup in such a way that the animal does not engage in the discrimination *behavior* would imply that the animal also does not engage in the *cognitive* task of discrimination. If this is the case, then we would find neither motor plans nor the cognitive decision that $f_2 > f_1$ during the passive condition. We will have to keep this confound in mind as a caveat.

Nevertheless, the intuition that we should initially characterize the neural activity under consideration as representing an inequality or a difference rather than as a motor plan is accounted for in terms of the computational comparison and decision *procedure*, as opposed to the inputs and outputs of that procedure. Given that consideration as well as the tentative empirical conclusion above (qualified by the confound caveat), it follows that we should at least tentatively reject the characterization of the data in those terms. Further, given the vehicle hypothesis, latency data, and the fact that M1 is well-established as a motor area from other experimental paradigms, the best initial characterization of the data is the one I provided in 7.5.3: Neural activity in M1 that is modeled by a regression analysis in which $a_1 = -a_2$, which, during successful behavior, covaries with the behavior of pressing either the medial or lateral buttons, is the implementation of a motor command to press one of those buttons.

This discussion clarifies the role that SPT plays with respect to explaining or describing the neural implementation of the computational comparison and decision procedure. SPT explains what

representation is and what determines representational content, and thus applies to the inputs and outputs, but it does not apply to the intervening computational procedure. Thus, we have a case where it is important to distinguish neural activity that is representational, from neural activity that is computational but not representational.

8.3 Artificial Percepts and the Failure to Represent

One interesting variant on this experiment used implanted electrodes to inject current directly into the columns of the rapidly adapting circuit, in S1 (Romo et al. 1998; Romo et al. 2000). The investigators found that the monkeys were able to perform the discrimination task, based on cortical microstimulation, at levels indistinguishable from the natural stimuli alone (see Appendix B for discussion). This case provides a clear example of the distinction between the two kinds of error, which I've called *failure to represent* and *representing falsely* (introduced in 4.4.3.3).

I've argued in 7.5.2.2 and 7.5.2.3 that neurons in subpopulation-1 of S1 have the teleofunction of covarying, according to r_2, with vibrating mechanical energy in the 5-50 Hz range, at superficial levels of around 500 μm under the skin, at each neuron's respective receptive field. Neurons in subpopulation-2 have the teleofunction of covarying with superficial mechanical energy, at their respective receptive fields, according to r_3. Whether the injected current stimulates subpopulation-1 or 2, in both cases, we get the same result. While these neurons have the teleofunction of covarying with energy states at the periphery, in the microstimulation cases, they fail to perform that function. Neural activity that occurs as a result of cortical microstimulation does not covary with energy states at the periphery.

Here we have a case of the failure to represent. According to SPT, this particular neural activity is a representation, and it is supposed to have representational content, but it has none. Further sensory representations in more central areas of the cortical circuit that are generated by cortical microstimulation also fail to represent, and thus, have no representational content. As a comparison, consider a case of representing falsely, which we found in S1 and S2, and discussed in 7.5.2.3.

A neuron in S2 has the teleofunction of covarying according to r_3 with mechanical energy at the periphery. If it is firing at 50 spikes/sec and is caused by the stimulator, then it has the content, *the stimulator is vibrating at 40 Hz*. However, it may still covary with energy states at the periphery while representing falsely. Assume that a neuron in S2 is caused by the stimulator vibrating at 40 Hz, and that neuron fires at a rate of 60 spikes/sec. In this case, the neuron has the content, *the stimulator is vibrating at 54.3 Hz*, and so its representational content is false, or, it represents falsely. By contrast, if that neuron isn't covarying with peripheral energy states at all, then it doesn't matter what its firing rate is. If it is firing, but not in covariance with peripheral energy, then it is weakly normatively defective; or, it is failing to perform its teleofunction.

This case allows for conceptual clarification of two points. First, I have earlier argued that the neuron has the teleofunction of covarying, *according to* r_3, with energy states at the periphery. The difference between representing falsely and failing to represent turns on whether the neuron performs its teleofunction. However, if the stimulator is vibrating at 40 Hz and the neuron is firing at 60 Hz, then it is failing to covary *according to* r_3, because it should be firing at 50 spikes/sec, yet the correct thing to say here is that it represents falsely, not that it fails to represent. We can use this case as a conceptual tool to go back to SPT and clarify it.

We must distinguish the teleofunctions that determine *that* a neuron has content, from those which determine *what the content is*. In the case under consideration, covarying with energy states at the periphery determines *that* the neural activity has representational content. By not covarying with peripheral energy states, the microstimulation-caused activity fails to perform its teleofunction and thus lacks content. On the other hand, *given* that the neural activity is covarying with peripheral energy states, r_3 determines representational content.

Thus, neurons in S2 have the teleofunction of covarying with energy at the periphery. If they are firing but not in covariance with peripheral energy, then regardless of their rate, they have no content because they fail to perform the teleofunction that determines that a representation has content at all. If these neurons do have content, then their rate matters. The f-predicative content of a particular neuron is whatever frequency r_3 maps to that neuron's current firing rate. If r_3 maps the actual frequency to the current firing rate (e.g., 40 Hz to 50 spikes/sec), then that activity is true. If r_3 maps some other frequency to the current firing rate, as in the case where the neuron fires at 60 spikes/sec and thus f-predicates a frequency of 54.3 Hz, then that neural activity represents falsely.

Second, we'll have to be clearer about 'covariation'. Even when the neurons are caused to fire by microstimulation, in a sense, they still *counterfactually* covary with peripheral energy states. But that's not what I'm getting at here. Rather, this is a kind of causal history, although it is a hybrid of causal history and counterfactual covariation; I'll call this *currently covarying*, as I did above.

I'll begin with an example. Suppose injected current caused neuron *n* to fire. If the injected current were to cease, then *n*'s firing would return to baseline. If the injected current were to increase its rate, then *n*'s firing rate would increase. In this case, *n*'s activity is currently covarying with current pulses injected into the cortex, but not with energy states at the periphery. Alternatively, suppose the

vibrating stimulator caused *n* to fire. If the stimulator were to increase its frequency, then *n*'s firing rate would increase. If the stimulator were to cease, then *n*'s rate would return to baseline. Here, the activity of *n* currently covaries with energy states at the periphery, but not with injected current pulses. The notion of currently covarying is essentially counterfactual causation tied to a particular causal history. Let's see if I can make that more precise.

Let $F = \{F_1, F_2, \dots\}$ be a set containing properties that vary along a range. Millikan would call these properties a "specifiable range of determinates under a determinable under which [F] falls" (Millikan 1984, 20). The Romo group would call this a *parameter* or a *scalar* (Romo et al. 1999). For example, F is the group of vibrating frequencies, whose members include 20 Hz, 21 Hz, etc. Each F_i is the (or a) causally efficacious property of *a*. Let $G = \{G_1, G_2, \dots\}$ be a set (or a parameter) also containing properties that vary along a range, where the members of G are members of the domain of the physiological relational system. The members of G are the causally relevant properties of *b*, with respect to F (for example, firing rates). The hybrid notion of currently covarying is this: (i) that *a* is F_i caused *b* to be G_i. Further, (ii) had *a* been different with respect to F, then *b* would have been different with respect to G[132], and (iii) should *a* change with respect to F, then *b* would change with respect to G.

These two clarifications apply to the present case as follows. Firing rates in S1 and S2 have the teleofunction of covarying with frequencies. In the case where the rate is 60 spikes/sec and has been caused by peripheral mechanical energy at a frequency of 40 Hz, then firing rate is *currently covarying* with frequency: the firing rate was caused by frequency, and if the frequency were to change, the firing rate would also change. However, r_3 determines f-predication, and a rate of 60 spikes/sec f-predicates a 54.3 Hz vibration. Since the stimulator is vibrating at 40 Hz, it follows that that neural activity is false.

[132] I'm borrowing this from Millikan's notion of reproduction. Compare: "Roughly, the law in situ implies that *had A been different* with respect to its determinate character *p* within a specifiable range of variation, as a result, *B would have differed accordingly*" (Millikan 1984, 20).

By contrast, when neurons in S2 are firing at any rate as a causal result of cortical microstimulation, they are not currently covarying with energy states in the periphery, and thus they are failing to perform their teleofunction and do not have representational content.

8.4 Noise in Neural Systems

The linear equation describing the relationship of firing rate in subpopulation-2 of S1 with stimulus frequency is

$$r(s) = 22 + 0.7s + \sigma\epsilon,$$

where *s* is stimulus frequency, $r(s)$ is rate described as a function of frequency, ϵ is noise with zero mean and unit variance, and σ is the standard deviation of the mean firing rate. I argued that, since noise is by definition not a signal, we should ignore the final noise term. By dropping the last term, we get r_3, the representation function. However, there is no such as a thing as a noiseless signal in the brain. Many of the mechanisms that open ion channels, as well as vesicle release and ion diffusion, are stochastic processes. Thus, there will always be "random" electrical activity, which is not a result of stimulus representation or neural processing and computation. We call this activity *noise*. But this would seem to have an unwanted implication: Neural activity such as firing rate is always false.

Representation function r_3 determines that 50 spikes/sec f-predicates the property of vibrating at 40 Hz. But even if everything is working properly, and the stimulus is at 40 Hz, the neuron will fire at

50 spikes/sec ± *noise*. Hence, it will not fire at 50 spikes/sec. Whatever rate it is actually firing at, that rate will f-predicate a property close to, but distinct from, 40 Hz, and thus the neural activity will be false. This is an unwanted implication and the theory should be adjusted to deal with it. In what follows I will outline a possible solution to this problem, leaving a more careful analysis for another day.

First, we distinguish three different kinds of noise. There is *c-noise*, *i-noise,* and *o-noise*. The prefixed letters stand for 'covariance', 'informational', and 'objective', respectively. The first and third kinds of noise are objective, while the second is relative to a cognitive agent's decisions on relevance. In general, noise is an alteration of the state of the receiver that is not due to an alteration at the source. Let's have some examples.

Assume that there is a one-to-one phase-locking of spikes of a primary afferent and frequency. For every amplitude peak of the sinusoidal frequency, the neuron fires a spike. Assume that, due to the pseudo-random processes mentioned above, the neuron fires a spike that was not caused by an amplitude peak of the stimulus, and does not correlate with that amplitude peak. That spike constitutes c-noise. It is objective and not relative to anyone deciding which possibilities are relevant for some particular purpose.

Information theory, on the other hand, is a useful theoretical tool that is ultimately dependent on certain fundamental probability assignments, which cannot be made in the absence of assumptions of relevance (see 2.4.2). In virtue of making those relevance assumptions, we necessarily render mutual information, entropy, and the other information-theoretic quantities, including noise and equivocation, non-objective. This is not problematic when we use information theory as a tool to aid us in discovering underlying nomic regularities, but for the purposes of reducing representation to physical entities, quantities, or characteristics, it is. I-noise is a conglomerate of four different quantities: average noise

and the noise associated with a single event, as well as average equivocation and the equivocation associated with a single event. Each of these quantities is relative to a judgment on relevance in order to assign probabilities.

What we are ultimately interested in is o-noise. That there is such a thing as objective noise is dependent on the fact that there is such a thing as objective, determinate representational content. Noise is always relative to a signal. If something's being a signal is relative to a cognitive agent's so deciding, or, the use for which a cognitive agent puts something, then we are not yet dealing with *objective* noise, because then noise itself is relative to a cognitive agent's decision or use. This is Dretske's noise and the noise of information theory. By contrast, if representation (and thus, something's *being a signal*) is not relative to any cognitive agent's intentional/representational states, then neither is noise. Thus, to understand noise, first we need to understand representation and representational content.

SPT explains representational content, so, to understand o-noise, we define it relative to representational content as defined by SPT. O-noise, then, is *an alteration of the content-bearing properties of a vehicle of representation, in the absence of an alteration in that vehicle's content*. Firing rate, for example, is a content-bearing property of representation in the brain. An alteration in firing rate in the absence of an alteration in content, is noise. But how, consistent with SPT, can we differentiate alterations in firing rate that change the f-predicate, from those that do not?

First, note that the representation function determines f-predication. A mapping function is a representation function if it is the teleofunction of the representations to correspond, according to that mapping function, with some parameter (e.g., with vibrations at the periphery). We infer that representational vehicles such as firing rates have that teleofunction partially on the basis of covariance.

But (here's the important part): sometimes alterations in the content-determining property of the vehicle (i.e., in the firing rate) occur as a result of c-noise. These alterations are "outside" of the content-determining milieu. As a result, we may non-arbitrarily ignore them as o-noise.

How do we know which alterations are content-determining, and which are not? Or, which are a result of c-noise? Information theory and other statistical measures of covariance are indispensable here, as fallible signs of an underlying nomically grounded regularity. For the same reason, information theory, and hence i-noise, are useful indicators of the prevalence of c-noise. Alternatively, the various other statistical assumptions, such as those used in regression analyses, or the assumption that noise has a normal distribution, are each fallible epistemic guides to the prevalence of c-noise. Given those fallible epistemic guides and those assumptions about the prevalence of c-noise, we can differentiate content-determining from non-content-determining alterations in firing rate by considering whether those alterations are within the range of i-noise.

In other words, what we need are the resources to distinguish a signal that is noisy-but-true, from a signal that is noisy-and-false. That difference is how far away the signal is from what it is supposed to be, given (i) its representation function, (ii) the value of the parameter it represents, and (iii) the noise range for that type of neuron[133]. If the range of noise for the neuron in question is 2 spikes/sec, the neuron fires at 51 spikes/sec, and the stimulator is vibrating at 40 Hz, then this is a true-but-noisy signal, because it is within the range of noise. The theoretical justification for this is as follows. That single extra spike/sec does not count as part of the signal, because it is a result of c-noise. Whereas our evidence that it is a result of c-noise is ultimately based on considerations regarding i-noise, since firing rate is a representation (or a signal), that extra spike/sec is actually o-noise. It is not part of the

[133] Here, the noise range is the c-noise range, to which we have epistemic access by virtue of i-noise.

content-determining alteration in firing rate. If, instead, the neuron is firing at a rate that is outside the range of noise, then it is a false-and-noisy signal. We thus have the resources to account for o-noise.

Importantly, there is a difference between (i) what determines representational content, and (ii) how we come to know what the representational content of a given vehicle is. As I mentioned above, both c-noise and o-noise are objective. They occur whether or not we make any epistemic judgments about them. However, for us to ascertain the content of some given representational vehicle, we need to know which aspects of firing rate are signal and which are noise. To get that, we need extra information, such as the frequency. But this is an epistemic problem and a practical problem, not a theoretical problem about the semantics of representational content. It also presents a practical problem for downstream neurons that use the signal for further processing and computation, but through massively parallel encoding as well as the use of oppositely tuned populations, neural systems have ways of getting around this[134]. The theoretical question of what determines representational content, however, is explained by SPT, and the demand for an incorporation of noise into that theory can be satisfied.

8.5 Objections and Clarifications

8.5.1 Clarifying the Roles of Each Component

Structural preservation, causal history, teleology, and causal covariation each have roles to play in the theory, but it is important to understand where each fits. Without structural preservation, there is no f-predication. Even though causal history determines *that a* caused representation *R*, it does not

[134] (Romo et al. 2003), for example, argue that oppositely tuned populations have the effect of reducing noise because the noise from oppositely tuned populations cancel each other out.

say which of the (perhaps infinite) aspects or properties of *a* that *R* f-predicates, nor does it allow for misrepresentation. The role of structural preservation is to match up empirical relational systems to each other, in such a way that values of some parameter map to values of some other parameter.

Notice that we have made use of several different kinds of structural preservation. Namely, we've used isomorphism, isomorphism*, and Δ/Ψ-morphism. The reason that isomorphism has been so prominent is that we have allowed the assumptions that our domains have the power of the continuum, their relations induce total ordering, and the domains have order-dense subsets. However, this does not imply that the other forms of structural preservation are not relevant elsewhere. For example, in Appendix C I present a different method of independently specifying relational systems based on the empirically discovered neurometric discrimination thresholds, using the range equivalence method. There we end up with finite relational systems. Ultimately I argue that that is not the best method of independently specifying relational systems in this case, however, I present it in an appendix to show that SPT has the resources to handle several different methods of specifying relational systems. It is a pliable theory.

Causal history maps vehicles to f-referents, without determining which property the representation f-predicates of the f-referent. Thus, there is no such thing as "false" f-reference, nor "mis-f-reference". Representational vehicles either f-refer or they do not. For example, when firing rate in S1 is caused by intracortical microstimulation, those vehicles fail to f-refer as a result of their failure to satisfy the teleofunction that determines that they have content, even though they still are representations.

Teleology plays several roles. First and foremost, biological devices, including representations, are what they are in virtue of teleofunction. Teleology thus answers the metaphysical question.

Teleology also plays a role in helping to select relational systems and mappings between them, by solving the mapping component of the non-uniqueness problem. This helps to define the empirical relational system at issue as well as select which among the infinite number of isomorphism-determining functions is the representation function. In virtue of so doing, teleology also helps to solve the causal chain problem. How do we know that these peripheral afferents have this teleofunction? This is where covariation plays its role.

It is because these afferents do so covary, coupled with the adaptiveness of sensory systems in general, that we conclude that these afferents have the teleofunction of doing what they do. Notice that covariation does not, in any case, determine either f-predication, f-reference, or answer the metaphysical question. Merely covarying is insufficient for any of these. However, finding that biological devices or systems do covary with energy at the periphery (for example) is indispensable for determining that that is what those biological systems are *supposed* to be doing. Compare, for example, the heart. The fact that it does circulate oxygenated blood is not enough to conclude that it is supposed to circulate oxygenated blood. However, that is certainly a crucial consideration when we come to argue that, indeed, the heart is *supposed* to do that.

Second, I used covariation as an indicator of representations and representational content, to initially characterize some of the data without presupposing a theory of representation. I emphasize that I only argued that this method works in the specific cases I mentioned, and I make no claims that it generalizes[135].

[135] As a general diagnosis of previous literature, we can see where much of the literature has gone wrong, and why. Covariation is an epistemic *indicator* of representations and their contents (at least in this one case), and covariation is an epistemic *indicator* of the presence of teleofunctions that partially determine that a thing is a representation, and help to determine content. But while covariation is *indicative* of these things, it is not *constitutive* of them. Failing to recognize this distinction may have resulted in many of the covariance or information theories that dominate this literature.

Finally, we discovered in 8.4 that we must distinguish counterfactual covariation, which helps to determine what a thing's teleofunction is, from currently covarying, which helps to determine whether a thing is currently performing its function. Currently covarying is a hybrid of causal history and counterfactual covariation; if a thing is currently covarying with whatever it is supposed to be, it is performing its teleofunction. For example, neurons in S1 might be currently covarying with energy states at the periphery, even though their activity is false. By contrast, neurons in S1 might be currently covarying with intracortical microstimulation, in which case they are not currently covarying with peripheral energy states, are not performing their teleofunction, and hence, they are in error by failing to represent. However, they still counterfactually covary with peripheral energy states, and this is part (but not all) of the reason why we attribute to them the teleofunction of covarying with peripheral energy states.

8.5.2 Objections: Teleology and Covariance

Fodor has argued that teleological accounts of representation cannot work because teleology cannot pull apart nomically necessarily coextensive predicates (Fodor 1990). Since actual history, not counterfactuals, determines a thing's teleofunction, and since natural selection is blind to these coextensive predicates, it follows that teleological accounts are also blind to these distinctions, and that, Fodor claims, is fatal to any teleological theory of representation. This is, for example, the fly-or-BB problem, where Fodor argues that neither covariation nor teleology can decide among the predicates 'fly' or 'fly-or-BB'.

On my theory, teleology has a role to play, because it helps in defining relational systems, it helps to solve the causal chain problem for f-reference, and it helps to solve the mapping component of the non-uniqueness problem. Does some version of Fodor's worry about teleology affect my theory?

We can, for example, set up a scenario that parallels the fly-or-BB problem. Why say that firing rate in S1 has the teleofunction of covarying with vibrating mechanical energy at the periphery? Why not instead say that it has the teleofunction of covarying with the disjunctive property, peripheral vibrations-or-intracortical stimulation? With respect to covariation, firing rate more closely covaries with the disjunction than with peripheral vibration alone. Since teleofunction is determined by actual evolutionary history, and since direct intracortical microstimulation has not been around long enough to have any evolutionary effect, it might seem to follow that teleology alone does not differentiate between peripheral energy and peripheral energy-or-intracortical stimulation. Thus, I cannot non-arbitrarily define my relational systems the way that I have, nor can I solve the causal chain problem using teleology as I have (so the objection might go).

The second group of objections involves my use of covariation. Since I've gotten so much mileage out of my arguments detailing the non-objectivity of all different kinds of information, don't these arguments apply to my use of covariation? What, exactly, do I mean by 'covariation'? The concept of covariation, as we've learned over the years, is a difficult one. If we mean exceptionless natural necessity, that would be conceptually coherent but never instantiated. There are no circumstances under which *X* and a representation of *X* always co-occur, especially since neurons run via stochastic processes. We've also learned above that I need to make a distinction between counterfactual covariation and currently covarying, which is a hybrid of causal history and counterfactual covariation. Even assuming that covariation is perhaps unquantifiable yet still legitimate in the role that I've assigned to it, am I nonetheless faced with some version of a disjunction problem? Further, since I clearly eschew exceptionless natural necessity (that is, having a conditional probability of 1), must I answer the question of how much covariation is enough to count for my purposes?

All of these objections are related, as are their replies. Let's start with covariance. With respect to non-objectivity, what makes the various kinds of information non-objective is their reliance on "relevant" background conditions, a distinction between signal and noise that is relative to our interests, or relativity to domains or reference classes, which are also relative to the interests of a cognitive agent. However, that the various kinds of information are not objective does not imply that covariation itself is relative to a cognitive agent's interests. Rather, given the same background conditions, if we reliably find the same number of bits of mutual information, this can be used as evidence, as an epistemic guide, that an objective correlation does in fact exist. The problem that I discussed previously is that, for the purposes of a philosophical, naturalistic reduction of mind, we cannot say that representation = information (or any variant of this), because to do so is to reduce representation to, essentially, itself, and hence is not to reduce it. What I do, by contrast, is reduce representation to causal history and structural preservation. The role of covariation in the theory is as a *fallible epistemic guide*, revisable by further evidence, that can be used as a reason to provisionally conclude that some biological device has such and such a teleofunction. So the non-objectivity problems of information are irrelevant to the way that I use covariation in the theory.

By 'covariation', I am not referring to exceptionless natural necessity, but rather, a regular and reliable co-occurrence of types of events. But if not covariance with a conditional probability of 1, then how much? Is there a non-arbitrary cutoff point? For this, I simply co-opt Millikan's response (Millikan 2004, 44). However much covariance was enough to be efficacious in the evolutionary history of the organism, is how much covariance must occur in order to count for my purposes of assigning teleofunction to a biological device.

The disjunction problem, similarly, is not a problem for my theory. The role of covariation is as a fallible guide to determining teleofunction. It is true that firing rate in S1 more closely covaries with

peripheral-energy-or-microstimulation. However, it is because of S1's covariance with peripheral energy that the brain evolved the way it did. The organisms that happened to have neurons that reliably covaried with subtle, fine-grained changes in peripheral mechanical energy at the fingertips were able to engage in various behaviors involving fine motor control of the fingertips, ultimately resulting in an evolutionary advantage over their competitors. When I say 'reliably', what I mean is that these neurons covaried more closely than the neurons of the animal's competitors, but not so closely that by so doing some other advantage (perhaps in cortical energy use?) was lost. The search for maximal covariance is an artifact of causal-informational theories of content that appeal to exceptionless natural necessity.

This clarification of the role of covariation helps to address Fodor's worry about teleology. First, the concern about teleology is that it doesn't break disjunctions generated by maximal covariance theories. When we are looking for nomically necessary covariation, as Dretske and Fodor were, we get a disjunction problem, and the suggestion is that teleology will break it. But I don't get a disjunction problem in determining content, because I use causal history for f-reference and structural preservation for f-predication. The need for teleology to break apart disjunctions arises with information theories because *the theory itself* generates these disjunctions. My theory, by contrast, does not generate these awkward disjunctions, and so there is no need to appeal to teleology to break them.

Nonetheless, there are well known difficulties with determining a thing's teleofunction. Are ascriptions of function necessarily indeterminate? If so, will this generate a problem with my frequent appeals to teleology? It need not. We acknowledge from the start that any item (and thus any biological device) will maximally covary with some disjunction of things or properties. But neither I nor Millikan have ever claimed that to determine a thing's teleofunction, we should determine what it maximally covaries with.

Rather, to determine a thing's teleofunction, we determine what a thing does, and what it did, such that its doing so resulted in evolutionary advantage. That is, because its tokens did such and such in the past, a type of thing is what it is today. However, any ascription of teleofunction may be false. This does not imply that there is a necessary indeterminacy; to think that it does (as Dennett 1987 does), is to confuse epistemology with metaphysics. We may not always be able to *know* what a thing's teleofunction is, partially because a thing's teleofunction is determined by what it did in the past and we do not always have access to the past. But even if we cannot know in principle, this does not imply that teleofunction is itself indeterminate. Thus, Fodor's worries about teleology and disjunctions can be assuaged with respect to SPT.

8.5.3 Causal Efficacy

The problem of mental causation is typically thought of in this way: Are representations causally efficacious *in virtue of their content*? That is, it is not enough to simply claim that vehicles of representation are causally efficacious. For example, Dretske describes a soprano hitting a high note that shatters glass (Dretske 1988, 79). The vehicle, that is, the sound wave, is causally efficacious, but not in virtue of its content[136]. It doesn't matter *what* she was saying, and in fact it doesn't matter if she was saying anything at all; that sound wave would have the effect it did regardless. That is not the kind of efficacy that is at issue here.

As it applies to my project, I've argued that representations are efficacious on every conception of 'representation', and thus we must ask: Can representations as defined by SPT be efficacious in

[136] Typically, it is said that *events* cause other events, not that things (e.g., representational vehicles) cause events. Thus the question of mental causation is the question of whether mental events cause things in virtue of the psychological properties that participate in, or instantiate, the mental event. I follow the custom in this literature of speaking loosely in terms of things (e.g. beliefs and desires) causing events. Nothing turns on it.

virtue of their content? I argue that, to the extent that this is a legitimate question, they can. However, in a sense this question is ill-formulated in the first place.

The question of whether a thing is efficacious in virtue of *X*, or qua *X*, should be translated into counterfactuals. That is, if firing rate is efficacious in virtue of having content *C*, then, had firing rate had a different content, that rate would have had a different causal effect. For example, assume that my words are efficacious in virtue of their content, and I say, "Please bring me a glass of water". Assuming my companion desires to do as I ask, a glass of water will be brought to me. But if I said something with a different content, namely, "please bring me a glass of milk", there would be a different outcome. By contrast, had the soprano asked for a glass of milk or water while singing, the glass would have shattered regardless of what she had said. In that case, content is irrelevant to the causal effects of the vehicle.

In the case of firing rate, since we assume that it is a vehicle with content *C*, the question of causal efficacy boils down to the counterfactual: Had rate had a different content, would the causal outcome have been different?

This question is answered in the affirmative. For concreteness, let's have an example. Firing rate in subpopulation-2 of S1 represents stimulus frequency, with f-predication determined by r_3. Maintaining our familiar example, if a group of neurons are firing at a rate of 50 spikes/sec, then they each f-predicate the property of vibrating at 40 Hz to their f-referent. Assume that the represented frequency is the comparison, and that the base frequency was 30 Hz. In virtue of their content, that is, because these neurons "say" that the comparison is 40 Hz, the animal will ultimately, with a great deal of reliability, press the medial and not the lateral button. If the neurons in question fire at a different

rate, where they f-predicate, say, 20 Hz, the animal will behave differently. Similar stories can be told about the motor plans in M1, the mnemonic representations throughout the circuit, and so on.

It is in virtue of their firing at *that rate* that the animal does what it does. Had the rate been different, the content would have been different, and had the content been different, the rate would have been different: They are one and the same thing. Further, had the rate been different, the causal effect would have been different, or what is just the same thing, had the content been different, the causal effect would have been different. Thus, changing the rate is the same as changing the content, which results in a change in behavior.

To the extent that this is a well-formulated question, it should be translated into counterfactuals. By doing so, we see that representations as defined by SPT are efficacious *in virtue of* their content. We see this because, had the content been different, the causal outcome would have been different. However, it may appear that I have trivialized the question by resorting to token identity, thus avoiding what is really at issue, which is whether, qua representation, a state is efficacious. Or rather, is a state efficacious in virtue of being a content-bearing state, or in virtue of something else? The soprano's high note is efficacious in breaking glass in virtue of its amplitude and frequency, not its content. In the case of representations as defined by SPT, are representational vehicles efficacious in virtue of being content-bearing states or having content *C*, or something else? But the important question here is: What else is there to "in virtue of", if not the counterfactual translation that I've provided?

Here is an analogy. Suppose we are playing catch with a baseball, and I throw the ball at you when you are not looking; it hits you in the head and creates a bruise. The baseball caused the bump, but in virtue of what? Is the ball efficacious in virtue of being white, or in virtue of being mine? No,

clearly the answer is that the ball is efficacious in virtue of being hard. That is, *qua* hard object thrown with sufficient force, the ball did what it did. We can tell this by translating the question into counterfactuals. Had the ball been different with respect to its hardness, it would have had a different effect. If it were harder, it would have created a larger bruise or perhaps done more damage. If it were soft, it would not have created a bruise at all. We should also examine the counterfactuals for the other properties. If the ball had been red, would it have had a different effect? If the ball had been yours would it have had a different effect? Obviously, the effect would have been the same in these counterfactual scenarios.

Thus, had the ball been different with respect to its hardness, it would have had a different effect, whereas if it had been different with respect to its color or ownership, it would have had the same effect. As a result, we conclude that it is in virtue of being a hard object, or, qua hard object, that the ball had its causal efficacy. Other than these counterfactuals, I ask again: What else is there to "in virtue of"? If nothing, then representations as defined by SPT are efficacious in virtue of their content.

A clarification is in order, made possible by the discussion of artificial percepts and the failure to represent. In the artificial stimulation scenario, firing rate is a representation, but according to SPT, has no content. However, note that firing rate has the *same* effect as it does in the natural stimuli scenario. In the latter case, firing rate does have content. It would seem that this draws apart the causally efficacious property of firing rate, and shows that it is *not* its having content that is efficacious, because the counterfactuals do not come out the way they should: Had firing rate lacked content, the causal outcome would be *the same*. But instead we expect this: Had firing rate lacked content, the causal outcome would be different. Thus, it would seem that rate is not efficacious in virtue of being a content-bearing state.

Even if we conclude that rate is not efficacious in virtue of being a content-bearing state, we may still conclude that rate is efficacious in virtue of *having content C*. In other words, the difference between having and not having content may not be a causally relevant difference, at least in this case. However, *given* that a representational vehicle has content, the specific content it has *is* a causally relevant factor. This is shown by the counterfactuals above: If the firing rate had a different content, it would have had a different causal effect.

While I don't contend that I have solved the problem of mental causation, I have at least shown that it is consistent with SPT for its vehicles of representation to have causal efficacy *in virtue of what they say*. Given that they say something, *what* they say makes a causally relevant difference.

8.6 Scaling Up the Theory

Philosophers in this area are not typically interested in the minutiae of neural details underlying a monkey's button-pressing in response to a stimulus. By using the experimental paradigm I have as a centerpiece in illustrating my thesis, and by pressing so hard on the distinction between intentionality and representation "as I've characterized it", it may seem that what I have to say is not relevant to the standard set of philosophical concerns here. Thus I have the added burden of showing that my theory is in fact relevant to those traditional concerns, and is a competitor to Fodor, Millikan, Dretske, and others. In this penultimate section, I address these concerns.

Because I am concerned with neural implementation, and have used a relatively simple animal behavior as the paradigm example of how my theory works, it might seem that my theory is not actually a competitor to Dretske et al. This is not the case: my theory is in direct competition with each of their

theories. Each of the above philosophers seek to explain intentionality, which involves aboutness, the capacity for error, causal efficacy, a relation one of whose relata need not exist, and the generation of intensionality. I, however, have restricted my explanandum to the first three components of intentionality. I want to show how a state in the world can "point to" or be about another, and can be mistaken. But that is exactly what these authors want to do as well. The difference is that they have added burdens that would be better addressed only after the core components of aboutness and the capacity for error have been explained. That is, I have taken the same problem that Dretske et al. have been working on, and broken it down into its constituent parts. I've adopted an incremental approach to solving the overall problem, by only attempting to solve the component problem that lies at its core: What is aboutness? Therefore, my theory is in direct competition with each of their theories.

Second, do not read too much into my use of monkey brains as illustrative of structural preservation theory. If anything, I've moved closer to humans than Millikan and Dretske have, with their bees (Millikan 1984), hoverflies (Millikan 1993), and magnetotaxic bacteria (Dretske 1986). Further, given the indistinguishable psychophysical curves between humans and monkeys, as well as the homologous neural structures involved in this task, we may see the set of experiments I've discussed as using a legitimate animal model of human neural activity. Thus, the description of how SPT applies to the monkey brain could very well be a description of how SPT applies to human brains, how humans implement sensory representations, working memory, and motor plans, and how SPT explains what representation is, as instantiated by human brains.

Another way that my theory contributes to addressing the traditional set of philosophical concerns is in clarifying the preliminary characterization of the appropriate explanandum. By analyzing the various ways that 'representation' is used in different theories, and by analyzing what they have in common with each other and with the dictionary senses of 'representation', I have provided a *better*,

more careful initial characterization of the explanandum that we are all working on. The fact that the concept of representation appears again and again, in different forms, in several different explanatory approaches, is significant and should not be ignored. Further, that there is a common, recurring core to all of these concepts, adds additional plausibility to the claim that I have provided a better initial characterization of the explanandum.

In particular, the requirement that a theory of basic, original representation explain the generation of intensionality seems misguided. Referential opacity is a logical phenomenon associated with sentences. More specifically, it is associated with representations of representations. For example, the sentence "Alfred believes that snow is white" is a representation of Alfred's belief, and thus, represents a representational state. This brings in added complexity because there are two contents involved. Intensionality is a property of something like higher-order representations, and its explanation should be found in the philosophy of language (as, for example, Millikan suggests in chapter 7 of her 2004), not in a theory of original representation, which is more fundamental, and must be presupposed by a theory of language.

SPT doesn't attempt to explain every facet of intentionality, but only the core elements common to the concepts of intentionality and representation as used in several theories. It explains what it is to be a representation and to have representational content, thus providing an explanation of aboutness and the capacity for error in reductive, materialistic terms. If successful, that is an extremely important contribution to the traditional project. Additionally, we can then build on the advances offered by SPT, to handle more of the components of intentionality[137]. Concerns about intensionality and fine-grained

[137] That is, should we decide that the intentional states associated with folk psychology do in fact exist in a manner consistent with the folk theory. I am skeptical of the veracity of folk psychology, but that is a different topic, and my skepticism takes nothing away from the need for an explanation of representation and representational content.

content can only be successfully addressed once an explanation of that common core is established, and that is what I contend that I have provided.

Given these considerations, it is clear that my theory addresses the traditional philosophical concerns in this area. Further, structural preservation theory, without modification, is a direct competitor to Fodor's, Dretske's, Millikan's, and the theories of everyone else who writes on intentionality and representation. As a final exercise, it will be useful to see how SPT can be extended or built on, to generate possible explanations of less basic representations. I wish to make clear however that my theory is, prima facie at least, *compatible with* each of the following extensions. SPT does not imply any of them.

First, SPT can be extended straightforwardly. Anything can be a member of a relational system, not just parametric energy states at the periphery of the organism. Thus, in addition to mechanical, electromagnetic, thermal, and other forms of energy, relational systems may include things like predator, food source, conspecific, shelter, and even perhaps truth, justice, and virtue. On the representation side, relational systems need not have domains constituted only by single neuron firing rates. Population measures, including vectors of arbitrary sizes, are easily accommodated within the theoretical framework developed here. Further, the full resources of structural preservation theory include much more than just isomorphism between relational systems ordered by a single relation. There are many different kinds of structural preservation, and arbitrarily large and complex ways of defining relational systems. With this adaptable framework in hand, it is surely possible to simply extend structural preservation theory to explain less basic representations.

Second, as I described in 6.3, SPT is consistent with the language of thought hypothesis. That hypothesis only claims that thoughts are syntactically structured, with parts that are both meaningful

and transportable. It says nothing about how those parts get their meaning. This is consistent with SPT in the following way. Basic representations have their content determined in the manner explained by SPT. However, at least some non-basic representations are the sorts of things adverted to by language of thought theorists. That is, they are the meaningful constituents of syntactically structured thoughts, which play roles analogous to predicates and names in natural languages. Those non-basic representations get their semantics by bearing a suitable relation to the basic representations. I don't know what that suitable relation might amount to. Perhaps it is something like being used in the appropriate way by a cognitive agent. For our purposes here, it doesn't matter, and it is worth pointing out that no one else knows what the suitable relation between basic and non-basic representations is either. SPT can be marshaled as an explanation of the most fundamental kind of representation. Then, less basic kinds of representation get their semantics by virtue of their relation to basic representations, in a manner consistent with the language of thought hypothesis.

Third, my theory is consistent with a holist, functional role semantics. One basic problem for these sorts of theories is that it is difficult to make sense out of communication or psychological explanations that advert to shared mental contents. This may be remedied by appealing to content similarity, rather than content identity. However, the standard reply here is that content similarity depends on content identity, because we still need to define the dimensions along which two concepts or thoughts have similar content. To do that, we need to be able to say that the dimensions are semantically identical. However, if we again make the distinction between basic and non-basic representations, we can use SPT to explain the semantics of basic representations. Then we explain the semantics of non-basic representations in something like functional role, and deal with the content identity problem by appealing to content similarity. By marrying itself to SPT, functional role semantics can avail itself of the determinate content explained by SPT in order to define the dimensions along

which less basic, meaningful states (i.e. thoughts, concepts, propositional attitudes) are similar. As an example, recall the comparison and decision computational procedure discussed in 8.2. Perhaps certain states of that computation can be defined functionally, and assigned their content in that way. Those would then be non-basic representations, and their semantics would ultimately depend on the representational content of the basic sensory, mnemonic, and motor representations, explained by SPT.

Fourth, Paul Churchland has a holistic theory of semantics which he calls *state space semantics* (Churchland 1986, 1989). State space semantics is similar to the f-predicative component of structural preservation theory. It maps vector-defined neural states (this is a population measure) to a "semantic space". In virtue of an isomorphism between the two, and a mapping function from one to the other, Churchland argues that content can be determined. The basic problem with this view is that, like holism, it does not have the resources to "tie down the edges" of the semantic space. In other words, while he does attempt to define similarity in an objective way, as discussed above, his theory does not account for how the dimensions of semantic space are themselves non-arbitrarily determined. However, we can accept the basic/non-basic distinction, and use SPT to explain the content of basic representations, and thus "tie down" semantic space. Then Churchland's population measures and vector spaces can be used to explain the content of less basic representations, while anchored to basic representations explained by my theory.

Fifth, SPT fits hand-in-glove with Millikan's theory of language. My chief complaint against her theory is that it is incomplete, because the mapping rules component of the theory doesn't work. The mapping rules component is, simply, what Millikan calls *Fregean sense*. However, we can replace Millikan's local information with structural preservation theory in order to explain Fregean sense. Everything else – dictionary sense, reference, intensions, proper functions, etc. – can all be left as is. If

the rest of Millikan's theoretical framework works, it can be supplemented with SPT, to provide a more wide-ranging theory of representation and meaning.

Finally, Prinz's proxytype theory of concepts (Prinz 2002) has several components, one of which is what he calls *intentional content*, which is distinct from *cognitive content*. In the language I've been using, both intentional content and cognitive content would be included in a theory of intentionality, while intentional content is simply what I've been calling representation. It is essentially aboutness with the possibility of error. Prinz uses a combination of information and etiology to explain intentional content, but does not provide separate theories for f-referential and f-predicative content as I do. However, we can replace the informational/etiological account of intentional content with SPT, while leaving everything else the same. If the rest of the theory works, then we have a more wide-ranging explanation of both representation and concepts.

There are many different ways that structural preservation theory can be extended or coupled with other theories of less basic kinds of representation. By providing an explanation of the most fundamental elements of the traditional philosophical concerns, namely, of aboutness and the capacity for error, structural preservation theory builds a foundation, and opens several doors for future investigation into some of the less basic, but equally important, philosophical concerns.

8.7 Conclusion

In this dissertation, I have attempted to shed light on a fundamental question about human nature: How do mental states arise from physical states and processes? While consciousness and intentionality are distinct, it seems to many that intentionality, representation, or aboutness is more

fundamental: Many theories attempt to explain consciousness in representational terms. It therefore awaits a theory of representation.

Yet, we don't have that theory. We don't know how matter and energy configures itself into states that are about, of, or directed at other states. To make progress on that project, I began with the recognition that there are several viable theories about the mind, behavior, or the human organism, among them our folk psychology, that make use of representation as a theoretical posit. While there are differences among these explanatory posits, there is also a common core. My target in this dissertation has been to explain that common core.

The end result is the structural preservation theory of original representation. While I contend that it enjoys great theoretical virtues, I have no doubt that it is very far from the final word, if there is such a thing. However, I also contend that it provides a conceptual and theoretical framework that may aid us, even if only a little, in the age old project of making sense of ourselves. It does not purport to explain consciousness nor even every aspect traditionally associated with intentionality, but it does purport to make progress in the explanation of what is probably more fundamental than either: the nature and implementation of representation in biological systems.

Appendix A: Measurement Theory and Empirical Relational Systems

A.0 Introduction

Measurement theory is concerned with two fundamental questions: Is the assignment of numbers to some phenomena justified? If so, to what degree is that assignment unique? These questions are answered by (i) proving a Representation Theorem, which demonstrates that at least a homomorphism connects an empirical relational system with a numerical relational system, and (ii) by proving a Uniqueness Theorem, which specifies the relation among different homomorphism-determining functions connecting the empirical and numerical systems. Properly speaking then, measurement theory *does not apply* to the case of two empirical relational systems: Measurement theory only applies to the case of one empirical system to be measured, and one numerical system that does the measuring.

Nonetheless, many of the concepts used by measurement theorists, such as isomorphism and homomorphism, are appealed to by philosophers seeking to understand representation. While it is easy to see the broad outlines of how these concepts can apply to the case of two empirical relational systems, the details are not so obvious. Specifically, proving that isomorphism or homomorphism connects two relational systems is a difficult and non-trivial task. In measurement theory that task is aided by certain properties of numerical systems. For example, the order density of the rationals allows the construction of a structure-preserving function from a countable infinite system to a numerical system. For another example, the induction of total ordering (for example, by $\geq$) is required to construct a bijection rather than a surjection between any two systems, and bijection is necessary for

isomorphism. However, it is not guaranteed that empirical relational systems will have these properties, so a different strategy is called for.

The goal of this appendix is as follows. I want to extend the results provided by measurement theorists, to the hitherto underexplored domain of two empirical relational systems. While much of what follows may be helpful as a review of Suppes and Zinnes (1963) and Krantz et al. (1971), the interspersed discussion of how to connect two empirical relational systems is my own contribution. I hope that it proves useful.

The proof that two systems are isomorphic is in every case an existence proof: We must show that a function exists that has certain properties. In what follows I review methods for constructing functions from empirical to numerical systems in cases where the empirical domain is finite, infinite but countable, and uncountable, in the context of measurement theory. Using that as a guide, I then show how to construct a parallel function from one empirical system to another. In some cases this is very simple, while in others it is not. This exercise will prove useful, because in so doing we will be able to explicitly articulate some of the empirical conditions on relational systems that suffice for using these methods. Thus, when we investigate whether empirical relational systems bear structural preservation to each other, all we need to do is determine certain of their properties.

A.1 Finite Domain

The physiological and non-physiological empirical relations of relevance to a theory of representation usually establish an ordering among a group of states. The firing rate of a neuron is one

obvious example, with greater-firing-rate-than establishing an ordering over the set of rates. It is useful to begin with some important classifications of orderings[138].

Let $\mathfrak{A} = \langle A, \succcurlyeq \rangle$ with $\succcurlyeq$ a binary relation on A. (The symbol '$\succcurlyeq$' denotes a relation generally, without distinguishing whether it is empirical or numerical. The numerical analogue of $\succcurlyeq$ is $\geq$, and the empirical analogue of $\succcurlyeq$ is $\geq_E$.) $\mathfrak{A}$ is a *weak order* (or, $\mathfrak{A}$ is *weakly ordered*) iff, for all $a, b, c \in A$, the following axioms are satisfied:

1. Connectedness: Either $a \succcurlyeq b$ or $b \succcurlyeq a$.
2. Transitivity: If $a \succcurlyeq b$ and $b \succcurlyeq c$, then $a \succcurlyeq c$.

All weak orders are also *reflexive*, since axiom 1 implies $a \succcurlyeq a$ for all a. If we add the further requirement of *anti-symmetry* then we get a *total order*:

3. Anti-symmetry: If $a \succcurlyeq b$ and $b \succcurlyeq a$ then $a = b$.

The distinction between a weak and a total order is significant. With a weak order, it is possible that $a \succcurlyeq b$ and $b \succcurlyeq a$ yet $a \neq b$. This happens often with empirical orderings. For example, two rods a and b may be neither greater than nor lesser than the other, yet the rods themselves are distinct elements of A. For another example, two distinct stimuli may be so close that they are physiologically

[138] My exposition of the following elementary concepts is drawn from (Krantz et al. 1971).

indistinguishable, in which case it may make sense to use an empirical ordering that takes into account their indistinguishability. There is a certain sense in which a and b, though distinct, are equivalent. The idea of equivalence is made more precise as follows.

If $\succcurlyeq$ is a binary relation on A, define $\sim$ and $\succ$ as follows:

$$a \sim b \text{ iff } a \succcurlyeq b \text{ and } b \succcurlyeq a,$$

$$a \succ b \text{ iff } a \succcurlyeq b \text{ and not } (b \succcurlyeq a).$$

If $\succcurlyeq$ induces a weak ordering on $\mathfrak{A}$ then $\sim$ is an *equivalence relation*. Thus, $\sim$ is reflexive, symmetric, and transitive. Equivalence relations define equivalence classes:

$$\text{Let } \boldsymbol{a} = \{b | b \in A,\ b \sim a\},$$

then $\boldsymbol{a}$ is the equivalence class determined by a. It is the set of all of the members of A that are equivalent to a. Equivalence classes are important because they partition a set into disjoint subsets, where every member of A is in exactly one equivalence class. The set of equivalence classes in A is called $A/\sim$, and $\succcurlyeq$, the weak order on A, induces a different relation, $\succcurlyeq_{Eq}$ on $A/\sim$, which totally orders $A/\sim$, defined as follows:

$$\boldsymbol{a} \succcurlyeq_{Eq} \boldsymbol{b} \text{ iff } a \succcurlyeq b.$$

This is important for the following reason. Because of the partition created by the equivalence classes, if we can show that $\mathfrak{A}' = \langle A/\sim, \succcurlyeq_{Eq} \rangle$ is isomorphic to a numerical relational system, we have thereby shown that $\mathfrak{A} = \langle A, \succcurlyeq \rangle$ is homomorphic to that same numerical relational system. Here's why: if we can define a function $f: A \to N$, $N \subseteq \mathbb{R}$, such that

$$a \succcurlyeq b \text{ iff } f(a) \geq f(b),$$

and if we assume that $a \sim b$, then both $f(a) \geq f(b)$ and $f(b) \geq f(a)$. However, $\geq$, the numerical relation, induces a total ordering, which implies that $f(a) = f(b)$. Thus, every member of $\boldsymbol{a}$ maps to the same element in N. If an isomorphism exists from the set of equivalence classes to a numerical relational system, then the equivalence classes map one-to-one onto a numerical set, in a relation-preserving way, and the members of those equivalence classes map onto, but not one-to-one, to that same numerical set, in a relation-preserving way. The proof of the first Representation Theorem consists in constructing a function $\boldsymbol{f}$ from $A/\sim$ to N (rather than a function $f: A \to N$). We want this:

$$\boldsymbol{a} \succcurlyeq_{Eq} \boldsymbol{b} \text{ iff } \boldsymbol{f}(\boldsymbol{a}) \geq \boldsymbol{f}(\boldsymbol{b}),$$

and then we obtain f by setting $f(a) = \boldsymbol{f}(\boldsymbol{a})$.

Now that we have defined weak and total orders and equivalence classes, we can drop the nonspecific relation $\succcurlyeq$, because the difference between an empirical relation and the numerical relation that respects it, is significant. To distinguish numerical from empirical relations, I use the subscript 'E' to signify that a relation is empirical. Later, we distinguish different empirical relations from each other. In that case, I use 'E' to signify that it's empirical, as well as 'A' or 'B' to connect it to a particular set. Finally, I continue to use the subscript 'Eq' to signify that this is a relation among equivalence classes of an empirical set. What we want then, rather than the above, is this:

$$\boldsymbol{a} \geq_{Eq} \boldsymbol{b} \text{ iff } \boldsymbol{f}(\boldsymbol{a}) \geq \boldsymbol{f}(\boldsymbol{b}),$$

where $\boldsymbol{a} \geq_{Eq} \boldsymbol{b}$ iff $a \geq_E b$. Thus, $\geq_{Eq}$ is the relation on equivalence classes defined over elements of the empirical set A, and $\geq_E$ is the empirical relation that weakly orders A.

Following (Krantz et al. 1971, 14-17), we'll construct that function in the following way. For each $\boldsymbol{a} \in A/\sim$, let $\boldsymbol{f}(\boldsymbol{a})$ be the number of distinct equivalence classes $\boldsymbol{b}$ such that $\boldsymbol{a} \geq_{Eq} \boldsymbol{b}$. This generates a counting process that assigns 1 to the lowest equivalence class, 2 to the next one up, and so forth. Now we need to show that, for every $\boldsymbol{a}, \boldsymbol{b} \in A/\sim$,

$$\boldsymbol{a} \geq_{Eq} \boldsymbol{b} \text{ iff } \boldsymbol{f}(\boldsymbol{a}) \geq \boldsymbol{f}(\boldsymbol{b}).$$

We'll start by proving the conditional from left to right. Assume that $\boldsymbol{a} \geq_{Eq} \boldsymbol{b}$. From axiom 2 (transitivity), if $\boldsymbol{a} \geq_{Eq} \boldsymbol{b}$ then for every $\boldsymbol{c}$, if $\boldsymbol{b} \geq_{Eq} \boldsymbol{c}$ then $\boldsymbol{a} \geq_{Eq} \boldsymbol{c}$, so if $\boldsymbol{c}$ is counted for $\boldsymbol{f}(\boldsymbol{b})$, then $\boldsymbol{c}$ is counted for $\boldsymbol{f}(\boldsymbol{a})$, hence, $\boldsymbol{f}(\boldsymbol{a}) \geq \boldsymbol{f}(\boldsymbol{b})$. Now we'll prove the conditional from right to left by proving its converse.

Assume *not* ($\boldsymbol{a} \geq_{Eq} \boldsymbol{b}$). From axiom 1 (connectedness), it follows that ($\boldsymbol{b} \geq_{Eq} \boldsymbol{a}$). Then there is at least one $\boldsymbol{c}$ (namely, $\boldsymbol{b}$) counted in $\boldsymbol{f}(\boldsymbol{b})$ but not counted in $\boldsymbol{f}(\boldsymbol{a})$, so, $\boldsymbol{f}(\boldsymbol{b}) > \boldsymbol{f}(\boldsymbol{a})$, hence, *not* ($\boldsymbol{f}(\boldsymbol{a}) \geq \boldsymbol{f}(\boldsymbol{b})$). This completes the proof that $\boldsymbol{a} \geq_{Eq} \boldsymbol{b}$ iff $\boldsymbol{f}(\boldsymbol{a}) \geq \boldsymbol{f}(\boldsymbol{b})$. However, it also needs to be shown that $\boldsymbol{f}$ is bijective for it to define an isomorphism (as opposed to a homomorphism or something else). Bijectivity follows from the above result. While Krantz et al. claim that it follows "immediately", I find it helpful to explicitly say how weak orders define homomorphisms and total orders define isomorphisms.

To show that a function is bijective we show that it is surjective or onto and injective or one-one. For surjectivity, define N as the image of $A/\sim$ under $\boldsymbol{f}$ (this is the set of all elements of $\mathbb{R}$ to which $\boldsymbol{f}$ maps an element of $A/\sim$). For one-one, we need to show that $\boldsymbol{f}(\boldsymbol{a}) = \boldsymbol{f}(\boldsymbol{b}) \rightarrow \boldsymbol{a} = \boldsymbol{b}$. Assume $\boldsymbol{f}(\boldsymbol{a}) = \boldsymbol{f}(\boldsymbol{b})$. Then $\boldsymbol{f}(\boldsymbol{a}) \geq \boldsymbol{f}(\boldsymbol{b})$ and $\boldsymbol{f}(\boldsymbol{b}) \geq \boldsymbol{f}(\boldsymbol{a})$. Since $\boldsymbol{a} \geq_{Eq} \boldsymbol{b}$ iff $\boldsymbol{f}(\boldsymbol{a}) \geq \boldsymbol{f}(\boldsymbol{b})$ then $\boldsymbol{a} \geq_{Eq} \boldsymbol{b}$ and $\boldsymbol{b} \geq_{Eq} \boldsymbol{a}$. From axiom 3 (antisymmetry), it follows that $\boldsymbol{a} = \boldsymbol{b}$. So $\boldsymbol{f}$ is bijective and defines an isomorphism. Notice that f is not bijective and does not define an isomorphism, but instead defines a homomorphism. From the definition of f in terms of $\boldsymbol{f}$, we know that $a \geq_E b$ iff $f(a) \geq f(b)$. We also know that f is surjective (by definition of N). However, our above conclusion that $\boldsymbol{a} = \boldsymbol{b}$ does not imply that $a = b$, because ($a \geq_E b$ and $b \geq_E a$) does not imply that $a = b$, but only that $a \sim b$. This is how a weak order can define a homomorphism but not an isomorphism, while a total order, because of antisymmetry, defines an isomorphism.

While weak orders do not define isomorphisms, all weak orders are associated with a total order through equivalence classes. Partition the original set by defining equivalence classes over it, define the set of equivalence classes, and define a new relation over the equivalence classes in terms of the original weak-order-inducing relation. This new relation will be a total order.

Note the relations between isomorphism, homomorphism, and embeddings. What we've shown above is that an isomorphism exists between $\mathfrak{A}' = \langle A/\sim, \geq_{Eq} \rangle$ and $\mathfrak{N}' = \langle N, \geq \rangle$, with $N =$ the image of $A/\sim$ under $\boldsymbol{f}$, and thus a homomorphism from $\mathfrak{A} = \langle A, \geq_E \rangle$ to $\mathfrak{N}' = \langle N, \geq \rangle$. Further, $\boldsymbol{f}$ defines an isomorphic embedding from $\mathfrak{A}'$ to $\mathfrak{N} = \langle \mathbb{R}, \geq \rangle$ and a homomorphic embedding from $\mathfrak{A}$ to $\mathfrak{N}$. While these considerations may seem pedantic, they have philosophical significance when we apply this to the problem of original representation.

I've discussed how to construct a numerical function and then prove that an empirical relational system is structurally preserved in a numerical relational system. But for a theory of representation we need to take this a step further and demonstrate that one empirical relational system is structurally preserved in another empirical relational system. For the finite case here under consideration this can be achieved in a simple way. First, use the Krantz et al. method from above to map $\mathfrak{A}'$ to the natural numbers. The empirical ordering in $\mathfrak{A}'$ will be preserved by the numerical $\geq$ ordering of the natural numbers. Then use that mapping to index the elements of $A/\sim$ as $\boldsymbol{a_1}, \boldsymbol{a_2}, \ldots \boldsymbol{a_n}$. The index ordering is taken from the $\geq$ order in the natural numbers, and preserves the $\geq_{AEq}$ ordering of $A/\sim$ as well as the $\geq_{AE}$ ordering of A (notice the subscripts used to differentiate the empirical relations associated with A from other empirical relations). Next, repeat this process for $\mathfrak{B}' = \langle B/\sim, \geq_{BEq} \rangle$ the equivalence class relational system associated with a second empirical relational system $\mathfrak{B}$, and index the elements of $B/\sim$ as $\boldsymbol{b_1}, \boldsymbol{b_2}, \ldots \boldsymbol{b_m}$. Again, since we've indexed them according to the Krantz et al. function described above, which is structurally preserving, the ordering induced by the indices is structurally preserving.

Finally, map $\boldsymbol{a_1} \rightarrow \boldsymbol{b_1}, \boldsymbol{a_2} \rightarrow \boldsymbol{b_2}$, etc. In this case, we must make the assumption that $n = m$; that is, there are as many elements of $A/\sim$ as there are of $B/\sim$. This mapping (let's call it $\boldsymbol{g}$), with the assumptions made, guarantees that for all $\boldsymbol{a}, \boldsymbol{b} \in A/\sim$, $\boldsymbol{a} \geq_{AEq} \boldsymbol{b}$ iff $\boldsymbol{g}(\boldsymbol{a}) \geq_{BEq} \boldsymbol{g}(\boldsymbol{b})$, and since both $\geq_{AEq}$ and $\geq_{BEq}$ induce total orders, they are antisymmetric and thus define an isomorphism from $\mathfrak{A}'$ to $\mathfrak{B}'$.

Here we have structural preservation between two equivalence class relational systems, which are derived from empirical relational systems. The numerical method above begins by constructing equivalence classes, and then proves isomorphism between $\mathfrak{A}'$ and $\mathfrak{B}'$, which establishes homomorphisms from $\mathfrak{A}'$ to $\mathfrak{N}' = \langle N, \geq \rangle$ and from $\mathfrak{B}'$ to $\mathfrak{N}'$. However, we can't use the indexing method described above to map the members of A to the members of B. While we can establish a one-one mapping of $\boldsymbol{a_1}, \boldsymbol{a_2}, \ldots, \boldsymbol{a_n}$ to $\boldsymbol{b_1}, \boldsymbol{b_2}, \ldots, \boldsymbol{b_m}$, this does not imply that the members of $\boldsymbol{a_1}$ will map one-to-one, in a structure-preserving way, onto the members of $\boldsymbol{b_1}$, because in equivalence classes there is no ordering of their members. The ordering of $A/\sim$ is what makes the construction of a function based on that order possible. Another option would be to simply map a_1 to the first member b_1 of $\boldsymbol{b_1}$. But without an order, there is no "first" member of $\boldsymbol{b_1}$ (or of $\boldsymbol{a_1}$). We also can't map a_1 to every member of $\boldsymbol{b_1}$ because that would not be a function.

What makes isomorphism between $\mathfrak{A}'$ and $\mathfrak{N}'$ possible is that both of these systems are totally ordered. The homomorphism from $\mathfrak{A}$ to $\mathfrak{N}'$ is derivative on that. What causes the problem when we're dealing with $\mathfrak{A}$ and $\mathfrak{B}$ is that neither of them are total orders. However, we can make an empirical assumption, and suppose that $\geq_{BE}$ is antisymmetric (for concreteness, this might be the assumption that the empirical greater-firing-rate relation is antisymmetric). That is, $(b_1 \geq_{BE} b_2 \; and \; b_2 \geq_{BE} b_1)$ implies that $b_1 = b_2$. With this assumption we construct a function as follows.

First, create equivalence classes on A (where A consists of stimulus frequencies, for example), then index them according to the method described above. Second, given our assumption of antisymmetry, B is totally ordered by $\geq_{BE}$, so we don't need equivalence classes here. Instead, index the members of B in the same way, and then map the equivalence classes $\boldsymbol{a_1}, \ldots, \boldsymbol{a_n}$ derived from A to the members $b_1, \ldots, b_m$ of B. Then we have an isomorphism from equivalence classes of frequencies to firing rates, and thus a homomorphism from frequencies to firing rates. Every member of B (each firing rate) gets several members of A mapped to it (that is, every member of $\boldsymbol{a_1}$, all of which are equivalent). Perhaps we should say that content is indeterminate but bounded, or that b_1 determinately represents every member of $\boldsymbol{a_1}$, or that b_1 represents the equivalence class itself. Or perhaps none of these. That is a different discussion.

A second option is to make the assumption of antisymmetry with respect to the stimulus relational system as well, and then prove isomorphism (not homomorphism) directly from the stimulus relational system to the physiological relational system. A third option is to eschew these empirical assumptions (perhaps they seem ad hoc), hence, both empirical relational systems of interest are only weakly ordered. Then construct equivalence classes and prove isomorphism among the equivalence class relational systems as above. Perhaps then we should claim that, if an element of B falls into $\boldsymbol{b_1}$ then that element represents the equivalence class that maps to $\boldsymbol{b_1}$, or at least one element of that equivalence class, or every element of the equivalence class. Fourth, the assumption that $m = n$, that is, that there are just as many elements of $A/\sim$ as there are of $B/\sim$ (and the correlative assumptions that must be made if we assume that one or both of $\geq_{BE}$ and $\geq_{AE}$ are antisymmetric), may not be justified. In this case, Swoyer's relaxations come in handy (see 5.4.5). If it turns out that there are more elements of $A/\sim$ than there are of $B/\sim$, then there will be some elements of $A/\sim$ that do not get

mapped to anything, and we have a morphism*. If there are more elements of $B/\sim$ then we have some kind of an embedding.

Here is a further complication. The measurement of $\mathfrak{A}$ with $\mathfrak{N}$ in the above example constitutes the use of an ordinal scale, which is essentially arbitrary except for order. That is, the assignment of numbers to the members of A (or $A/\sim$) is unique up to a monotone transformation[139]. The indexing process that makes possible the order-preserving mapping of $A/\sim$ to $B/\sim$ is not unique. Does this cause a problem for the mapping of empirical relational systems? Is there an empirical analogue of a monotone transformation? The answer to the first question is no. Imagine that, rather than mapping $A/\sim$ to $\{1,2,\dots,n\}$ and using that mapping to index the members of $A/\sim$, we instead use a monotone transformation of $\boldsymbol{f}$. Then we would map $A/\sim$ to, say, $\{\alpha 1, \alpha 2, \dots, \alpha n\}$. The order would still be preserved (this is what makes the original assignment non-unique, or, unique only up to a monotone transformation). So long as internal relations are preserved, we can still use the elements of the new set $\{\alpha 1, \alpha 2, \dots, \alpha n\}$ to index the elements of $A/\sim$, and then use the indexing of $A/\sim$ and $B/\sim$ to map their members. This does not imply, however, that there is only one structure-preserving mapping of $A/\sim$ to $B/\sim$, or $A/\sim$ to B, etc. In general there will be more than one. There may still be several "equally good" mappings that preserve structure between empirical relational systems, perhaps through an empirical analogue of a monotone transformation. The philosophical ramifications of this for a theory of representation are discussed in chapter 5, but especially see 5.4.7.1.

[139] A function $f: A \to B$ is monotone increasing iff for all $x, y \in A$,if $x < y$ then $f(x) < f(y)$. It is monotone decreasing iff for all $x, y \in A$, if $x < y$ then $f(x) > f(y)$. A function is monotone is if it monotone increasing or monotone decreasing.

A.2 Infinite Countable Domain

We constructed various order-preserving functions above based on the assumptions that the domain is totally ordered and finite. Hence, there exists a lowest member in the ordering, and we defined the function in terms of that element. However, perhaps we should individuate empirical domains in such a way that they are not finite. Some neurons for example do not fire action potentials but only graded potentials. In this case, these neurons do not "fire" at all, so individuating their states in terms of firing rate is senseless, and a better, independently motivated way of typing neural states here is in terms of voltage. But how should we individuate voltage states? Continuously? Are there countably many voltage states? I'm not going to try to answer these questions here, but in anticipation that in some cases it will make sense to individuate states of the world (physiological or otherwise) as varying along a dimension with an infinitely fine grain, we will discuss what to do with empirical relational systems whose domains are infinite but countable, and then in A.3 we will deal with the uncountable case.

The proof of a Representation Theorem is in every case an existence proof, which is achieved by constructing a function and then demonstrating that that function has the requisite properties. For the denumerable case, I draw on (Suppes and Zinnes 1963, 23-28). My purpose in describing these mapping functions, as above, is twofold. First, I want to use the ideas implicit in the numerical function to construct a function from empirical to empirical sets. Second, I want to show what properties of the empirical relational systems would suffice for the existence of structural preservation between them.

The general strategy is the same as for the finite case. We begin with a weakly ordered relational system, then derive a totally ordered relational system from it by defining equivalence classes and an antisymmetric relation over them, where that relation is derived from the original, weak-order-

inducing relation. Then prove isomorphism from the total order to a numerical relational system, and thus homomorphism from the original weak order to the same numerical relational system.

Let $\mathfrak{A} = \langle A, \geq_{AE} \rangle$ be an empirical relational system with $\geq_{AE}$ defining a weak order. Let $\mathfrak{A}' = \langle A/\sim, \geq_{AEq} \rangle$ be the relational system derived from $\mathfrak{A}$, with $A/\sim$ the set of equivalence classes derived from A, and $\geq_{AEq}$ the total order-inducing relation derived from $\geq_{AE}$. Let $\mathfrak{N} = \langle \mathbb{R}, \geq \rangle$ and $\mathfrak{N}' = \langle N, \geq \rangle$, with $N =$ the image of $A/\sim$ under f_1 (to be defined shortly). We're going to construct a function f_1 that defines an isomorphism between $\mathfrak{A}'$ and $\mathfrak{N}'$, a homomorphism from $\mathfrak{A}$ to $\mathfrak{N}'$, an isomorphic embedding of $\mathfrak{A}'$ in $\mathfrak{N}$, and a homomorphic embedding of $\mathfrak{A}$ in $\mathfrak{N}$. Finally, let $\mathfrak{B} = \langle B, \geq_{BE} \rangle$ be a weakly ordered empirical relational system and $\mathfrak{B}' = \langle B/\sim, \geq_{BEq} \rangle$ be the totally ordered relational system associated with $\mathfrak{B}$ by construction of equivalence classes. To maintain contact with intuition, we may interpret B as the set of firing rates and A as the set of vibrotactile frequencies. However it's important to keep in mind that this discussion is intended to be fully general, so that we can simply apply what we learn here to various other empirical relational systems of interest to a theory of representation.

The assumption that A is denumerable is crucial, because it implies that the members of A (and hence also the members of $A/\sim$) can be placed in one-one correspondence with the natural numbers. Thus we begin by enumerating the elements of $A/\sim$ as $\boldsymbol{a_1}, \boldsymbol{a_2}, \ldots, \boldsymbol{a_n}, \ldots$. In addition, the rational numbers are also denumerable, as $r_1, r_2, \ldots, r_n, \ldots$. The enumeration $\boldsymbol{a_1}, \boldsymbol{a_2}, \ldots, \boldsymbol{a_n}, \ldots$ is *not* the ordering induced by $\geq_{AEq}$, but is an enumeration of the elements that we know to exist from the assumption of denumerability. We'll now define $f_1: A/\sim \rightarrow N$ by induction. First set

$$f_1(\boldsymbol{a_1}) = 0,$$

and then consider $\boldsymbol{a_n}$, of which there are three cases.

Case 1. $\boldsymbol{a_n} \geq_{AEq} \boldsymbol{a_i}$, for $i = 1,2, \ldots, n-1$. Then set

$$f_1(\boldsymbol{a_n}) = n.$$

Case 2. $\boldsymbol{a_i} \geq_{AEq} \boldsymbol{a_n}$, for $i = 1,2, \ldots, n-1$. Then set

$$f_1(\boldsymbol{a_n}) = -n.$$

Case 3. If neither cases 1 nor 2 apply, then there exist integers $1 \leq i, j \leq n$ such that $\boldsymbol{a_i} \geq_{AEq} \boldsymbol{a_n} \geq_{AEq} \boldsymbol{a_j}$. Then we'll use our enumeration of the rationals to define $f_1(\boldsymbol{a_n})$. First, define

$$\boldsymbol{a_n^*} = max\{\boldsymbol{a_j} | \boldsymbol{a_n} \geq_{AEq} \boldsymbol{a_j} \,\&\, j < n\},$$

$$\boldsymbol{b_n^*} = min\{\boldsymbol{a_i} | \boldsymbol{a_i} \geq_{AEq} \boldsymbol{a_n} \,\&\, i < n\}.$$

Then set

$$f_1(\boldsymbol{a_n}) = r_{1st},$$

where 'r_{1st}' denotes the first rational number in our enumeration of the rationals such that $f_1(\boldsymbol{a_n^*}) < r_{1st} < f_1(\boldsymbol{b_n^*})$. By induction f_1 is now defined for all $\boldsymbol{a_n}$. Further[140],

$$\boldsymbol{a} \geq_{AEq} \boldsymbol{b} \text{ iff } f_1(\boldsymbol{a}) \geq f_1(\boldsymbol{b}), \text{ for all } \boldsymbol{a}, \boldsymbol{b} \in A/\sim.$$

The proof of the above biconditional proves the various morphisms noted above (i.e., f_1 defines an isomorphism between $\mathfrak{A}'$ and $\mathfrak{R}'$, its associated function defines a homomorphism from $\mathfrak{A}$ to $\mathfrak{R}'$, etc.). It will be helpful to say that again but without the formalisms, to get a better feel for how the function is constructed and what complications it introduces when we use it for our purposes.

The enumeration of the equivalence classes is essentially random with respect to the empirical ordering[141]. Element $\boldsymbol{a_1}$ is not the "first" or lowest member of $A/\sim$ in the empirical $\geq_{AEq}$ ordering, it is simply whatever equivalence class happened to be chosen to correspond to 1. To construct a function from equivalence classes to a subset of the reals, we begin by setting $f_1(\boldsymbol{a_1}) = 0$, and then we consider

[140] For the proof see (Suppes and Zinnes 1963, 23-28). For a slightly different way to construct the function for a countable total order see (Krantz et al. 1971, 38-40).

[141] I should not call $\geq_{AEq}$ an empirical ordering because it is an ordering over equivalence classes of empirical things, not the empirical things themselves. It is however derived from the empirical ordering $\geq_{AE}$ (recall that $\boldsymbol{a} \geq_{AEq} \boldsymbol{b}$ iff $a \geq_{AE} b$), nothing really turns on it, and it makes exposition easier, so I do it anyway.

the other elements of $A/\sim$, which must fall into one of three cases. In the first case, if all of the $\boldsymbol{a_i}$ that are below $\boldsymbol{a_n}$ in the numerical ordering, are also below $\boldsymbol{a_n}$ in the empirical ordering, then we set the value of $\boldsymbol{a_n}$ to n. If all of the $\boldsymbol{a_i}$ that are below $\boldsymbol{a_n}$ in the numerical ordering are above $\boldsymbol{a_n}$ in the empirical ordering, then we set the value of $\boldsymbol{a_n}$ to $-n$. That is the second case. Finally, it could be the case that, among all of the $\boldsymbol{a_i}$ that are below $\boldsymbol{a_n}$ in the numerical ordering, at least one is above and at least one is below $\boldsymbol{a_n}$ in the empirical ordering. What we want then is for $f_1(\boldsymbol{a_n})$ to fall between the values of the two $\boldsymbol{a_i}$ that are closest to $\boldsymbol{a_n}$. So we define $\boldsymbol{a_n^*}$ and $\boldsymbol{b_n^*}$ as the two $\boldsymbol{a_i}$ that are closest to $\boldsymbol{a_n}$, with $\boldsymbol{a_n^*}$ above $\boldsymbol{a_n}$ and $\boldsymbol{b_n^*}$ below $\boldsymbol{a_n}$, in the empirical ordering (the "max" and "min" are with respect to the empirical, not the numerical ordering). Further, since the rational numbers are denumerable, we enumerate them, and set $f_1(\boldsymbol{a_n})$ to the first rational number (in the rational number enumeration) that falls between $f_1(\boldsymbol{a_n^*})$ and $f_1(\boldsymbol{b_n^*})$ (where between is here with respect to the $\geq$ ordering). This effectively fits $f_1(\boldsymbol{a_n})$ into the ordering that we're establishing among the real numbers in such a way that it preserves the empirical ordering of $A/\sim$. It is guaranteed that such a rational number exists because every $\boldsymbol{a_i}$ gets sent to either an integer (n or $-n$), or to a rational number that is not an integer, and between any two rational numbers there exists another.

Intuitively, we can see that this function preserves order. Setting $f_1(\boldsymbol{a_1}) = 0$ is like setting the value of $\boldsymbol{a_1}$ right in the middle. Then for $\boldsymbol{a_2}$, if it is above $\boldsymbol{a_1}$ in the empirical ordering, we set it to 2; then $\boldsymbol{a_2} \geq_{AEq} \boldsymbol{a_1}$ and $f_1(\boldsymbol{a_2}) \geq f_1(\boldsymbol{a_1})$. If $\boldsymbol{a_3}$ is between $\boldsymbol{a_2}$ and $\boldsymbol{a_1}$ in the empirical order, then we set $f_1(\boldsymbol{a_3})$ to a value that is between 2 and 0, and so forth.

Ultimately we want a structure-preserving function from A to B. In the finite case we were able to index the equivalence classes of empirical elements via Krantz's function, where the index-ordering preserved the empirical ordering, and then use the indices to define a function from equivalence classes of empirical objects to equivalence classes of empirical objects. From there, depending on what

empirical assumptions seem legitimate regarding the empirical relations, we can complete the transition to some type of structural preservation among the empirical relational systems. In this case, that process doesn't work.

We know by assumption that both A and B are denumerable, and let's assume, for ease of exposition, that the empirical relations on them induce total orders. We also know from Suppes and Zinnes' work described above, that both $\mathfrak{A}$ and $\mathfrak{B}$ are isomorphically embedded in $\mathfrak{N}$[142]. Prima facie, a reasonable strategy would be to use the values of f_1 and f_2 (where f_2 is a function from B to N analogous to f_1) to index the elements of A and B in the same way we did for the finite case, then map the elements of A to B using the two index-orderings. That worked in the finite case when we mapped the elements of A to integers, but it doesn't work in the infinite case.

The function from A to B would be defined in terms of the indices associated with the members of A and B, so we can just speak in terms of the indices for a moment. Given Suppes and Zinnes' function, we have a set C and a set D (the members of these are the indices on A and B), both of which are proper subsets of the rationals, totally ordered by $\geq$, and both including 0 as a member. Since they are both infinite yet countable, we know that they are equinumerous and can be mapped one-one to the integers. However, what we cannot infer is that there is an order-preserving bijection from C to D, or, that there exists a function $f_3: C \rightarrow D$ such that, for every $a, b \in C$, $a \geq b$ iff $f_3(a) \geq f_3(b)$. We may begin by setting $0 \in C$ to $0 \in D$, but from there on up (or down) there is no guarantee that there is a "next" member, because we are dealing with an infinite subset of the rationals, which have the following property. For any two elements a, c with $c \geq a$, there exists a third element b such that

[142] Actually, we know from Suppes' and Zinnes' work that $\mathfrak{A}'$ and $\mathfrak{B}'$ are isomorphically embedded in $\mathfrak{N}$, but with the added assumption that the empirical relations induce total orders, we thus can infer that $\mathfrak{A}$ and $\mathfrak{B}$ are isomorphically embedded in $\mathfrak{N}$. I only make that assumption so we can discuss this in terms of A and B rather than the equivalence classes derived from them. Nothing substantial turns on the assumption.

$c \geq b \geq a$. Since we don't know anything else about the sets C and D other than that they are countable infinite proper subsets of the rationals, we have no way of guaranteeing that there will be a next member, and thus no way of preserving the $\geq$ ordering from one set to the other[143]. We therefore can't use the indexing method described above to preserve the empirical order of A in B.

There is a way to get around this, and the same property of the rationals that causes the problem also provides a solution. It will be beneficial to first approach the uncountable case, then we'll see how to construct a structure-preserving function for two empirical relational systems for both countable and uncountable sets.

A.3 Uncountable Domain

The rationals are a countable subset of the reals, which have the important property of being "thoroughly interspersed" among the reals: between any two real numbers there exists a rational number. This interspersion is what makes possible the preservation of structure of an uncountable relational system in a numerical relational system whose domain is the reals. For an uncountable set, even with upper and lower bounds and even if it is totally ordered, we have a similar problem as we encountered immediately above when we tried to map C onto D in an order-preserving way. There is no "next" element in the ordering on the uncountable set. For the countable case this problem was avoided by indexing the elements with the natural numbers (even though the numerical indexing did not preserve the empirical order), and then constructing an order-preserving function that makes use of

[143] Thanks to Michael Levin for guidance on this point.

the interspersion of the rationals. This doesn't work for the uncountable case because we cannot enumerate them[144].

What we'll do instead is associate the elements in the uncountable set with elements in a perspicuously chosen countable subset. Then, using the theorem above (that a countable, totally ordered relational system is isomorphically embedded in $\mathfrak{N}$), we show that the uncountable relational system is isomorphic to $\mathfrak{N}$.

We begin by making the idea of "thoroughly interspersed" more precise, with the concept of *order density*. Let $\mathfrak{A} = \langle A, \geq_{AE} \rangle$[145] be a total order with $G \subseteq A$. Then G is *order dense* in A iff for all $a, c \in A$ such that $a >_{AE} c$, there exists $b \in G$ such that $a \geq_{AE} b \geq_{AE} c$. The rationals are order dense in the reals. Krantz et al. have proven that, if $\mathfrak{A}$ is a total order, then the following two conditions are materially equivalent (Krantz et al. 1971, 40-42):

(i) There is a finite or countable order-dense subset of A.

(ii) There is an isomorphism of $\mathfrak{A}$ into $\mathfrak{N} = \langle \mathbb{R}, \geq \rangle$.

The basic idea for how to construct the function is to consider the elements in the uncountable set A as a limit of elements in the countable, order dense subset G. Then a function is constructed from G to the reals. What makes the order-preservation possible, both in the countable and uncountable

[144] I rely on (Krantz et al. 1971, 40-42) for the explanation of how to construct an isomorphism-determining function from an uncountable empirical relational system to a numerical relational system, as well as for the concept of order-density that follows (although this is a common mathematical concept). As with the countable case, I will not discuss the proof, but see the above cited pages.

[145] The definition for order density is general; it does not apply only to the empirical relation associated with A discussed above. For ease of exposition I just speak in terms of $\geq_{AE}$.

case, is that the range of the function is order dense. The rationals are order dense in the reals (for the uncountable case), and the rationals are order dense in the rationals (for the countable case). Thus, it is necessary and sufficient for isomorphism from an uncountable total order to a numerical total order (whose domain is $\mathbb{R}$) that there is a countable order-dense subset of the domain of the function. It is sufficient for isomorphism from a countable total order to a numerical total order that there is an order dense subset of the range of the function: Note that it is the order density of the rationals that made the construction of f_1 possible.

For the purpose of applying these concepts to a theory of representation, the question is: Can this be made to work with *two* empirical relational systems? We discussed above that we can't use the indexing method because of the order density of the rationals. What we can do instead is make the empirical assumption that B has an order dense subset (I remark on whether this is a justified empirical assumption in 5.4.7). First, I'll show how to connect two non-finite empirical systems for the countable case.

Let $\mathfrak{A} = \langle A, \geq_{AE} \rangle$ be a countable empirical total order. (We could do without the assumption of total ordering, associate equivalence classes and a total order with the weak order, and then the various qualifications discussed above would apply. However it'll be easier on the eyes if we just assume total ordering from the start. In A.4 I'll provide a list that summarizes the various relationships among the different empirical assumptions we might make.) Let $\mathfrak{B} = \langle B, \geq_{BE} \rangle$ be an empirical total order, and let B have an order dense subset (not necessarily a proper subset). Using f_1 as a guide, I'll construct a function $f_4 \colon A \to B$ that is order-preserving (i.e, for all $a, b \in A$, $a \geq_{AE} b$ iff $f_4(a) \geq_{BE} f_4(b)$), by induction.

First, enumerate A according to our assumption of denumerability as $a_1, a_2, \ldots, a_n, \ldots$. Second, enumerate B as well, as $b_1, b_2, \ldots, b_n, \ldots$. Choose any member of B, and name it b^*. This element is going to act in an analogous way as 0 did in f_1. I'll sometimes refer to b^* as the *zero point* of B. Now we need to partition B, as follows. First define a relation $<_{BE}$ from $\geq_{BE}$ in the obvious way: If $not\ (a \geq_{BE} b)$ then $b <_{BE} a$. For all $i = 1,2, \ldots n, \ldots$, if $b_i \geq_{BE} b^*$ then $b_i \in B^+$. If $b_i <_{BE} b^*$ then $b_i \in B^-$. Partitioning B in this way provides an analogue of the positive and negative rationals.

Set

$$f_4\ (a_1) = b^*.$$

Now consider a_n, for which there are three cases.

Case 1: If $a_n \geq_{AE} a_i$, for all $i = 1, \ldots n-1$, then, if for all $i = 2, \ldots n-1$, Case 1 does *not* apply, then set

$$f_4\ (a_n) = b_{n^+},$$

where b_{n^+} is any randomly chosen member of B^+ other than b^*. If Case 1 *does* apply for some $i = 2, \ldots n-1$, then set

$$f_4\ (a_n) = b_{n^{++}}.$$

Define $b_{n^{++}}$ in terms of $b_{n^{max}}$, as follows.

$$b_n{}^{max} = max\{f(a_i)|Case\ 1\ applies\ \&\ i < n\},$$

where 'max' is with respect to the $\geq_{BE}$ ordering. $b_n{}^{++}$ is any member of B^+ that is $\geq_{BE} b_n{}^{max}$ but $\neq b_n{}^{max}$.

Case 2: If $a_i \geq_{AE} a_n$ for all $i = 1, \ldots, n-1$, then, if for all $i = 2, \ldots, n-1$, Case 2 does *not* apply, then set

$$f_4\ (a_n) = b_n{}^{-}$$

where $b_n{}^{-}$ is any randomly chosen member of B^-. If Case 2 *does* apply for some $i = 2, \ldots, n-1$, then set

$$f_4\ (a_n) = b_n{}^{--}.$$

Define $b_n{}^{--}$ in terms of $b_n{}^{min}$, as follows.

$$b_n{}^{min} = min\{f(a_i)|Case\ 2\ applies\ \&\ i < n\},$$

where 'min' is with respect to the $\geq_{BE}$ ordering. $b_n{}^{--}$ is any member of B^- that is $<_{BE} b_n{}^{min}$.

Case 3: If neither Case 1 nor Case 2 applies, then there are integers i, j, with $1 \leq i, j \leq n$, such that $a_i \geq_{AE} a_n \geq_{AE} a_j$. Define

$$a_n^* = min\{a_i | a_i \geq_{AE} a_n \ \& \ i < n\}$$

$$b_n^* = max\{a_j | a_n \geq_{AE} a_j \ \& \ j < n\}$$

Then set

$$f_4\ (a_n) = b_{1st},$$

where b_{1st} is the first b_i (with respect to the enumeration) between $f_4\ (a_n^*)$ and $f_4\ (b_n^*)$; that is, $f_4\ (b_n^*) \geq_{BE} b_{1st} \geq_{BE} f_4\ (a_n^*)$. $f_4\ (a_n^*)$ and $f_4\ (b_n^*)$ are guaranteed to be defined because a_n^* and b_n^* both fall before a_n in the enumeration ordering and *f* is defined by induction. The existence of b_{1st} such that $f_4\ (b_n^*) \geq_{BE} b_{1st} \geq_{BE} f_4\ (a_n^*)$ is guaranteed by the assumption of order density in B.

Suppes and Zinnes proved that f_1 has the following property:

$$aRb \text{ iff } f_1(a) S f_1(b)$$

for all $a, b \in A$. That f_1 is a bijection followed from that (and the assumption that A is totally ordered). Since I defined f_4 in a parallel way, on the assumption that B is order dense, with b^* acting as a zero point as well as B^+ and B^- playing the role of the positive and negative rationals, we can conclude that f_4 is also an order-preserving bijection:

$$a \geq_{AE} b \text{ iff } f_4(a) \geq_{BE} f_4(b)$$

for all $a, b \in A$. Thus, assuming that A is countable and totally ordered and B is totally ordered and order dense, we can conclude that the empirical relational system $\mathfrak{A}$ is isomorphically embedded in the empirical relational system $\mathfrak{B}$[146].

Implicit in the above construction of f_4 is the assumption that B has neither an upper nor a lower bound. Anticipating that that is not a realistic empirical assumption (we will usually but not always be interpreting B as the set of firing rates), we can straightforwardly modify the construction of f_4 to allow both lower and upper bounds. The lower bound on the empirical system of firing rates is the case where the neuron does not fire at all. Further, there is a physiologic upper limit on firing rate, which is a rate above which the neuron cannot fire. We take these into account as follows.

We will define the lower bound on B as $b_0 = b_n$ such that $b_i \geq_{BE} b_n$ but $b_i \neq b_n$, for all $i = 1,2, \ldots n, \ldots$. Then for Case 2 above, we define b_n– as any member of B^- except b_0. We define b_n– – as any member c of B^- such that $c <_{BE} b_{n^{min}}$, $c \geq_{BE} b_0$, and $c \neq b_0$. Then we define the upper bound on B as $b^{super} = b_n$ such that $b_n \geq_{BE} b_i$ and $b_n \neq b_i$ for all $i = 1,2, \ldots n, \ldots$. (The superscript 'super' is to remind us of 'superlative', so as to avoid confusion with $b_{n^{max}}$.) Then for Case 1 above, we define b_{n^+} as any member of B^+ except b^{super}, and define $b_{n^{++}}$ as any member of B^+ that is $\geq_{BE}$ but $\neq b_{n^+}$ but $<_{BE} b^{super}$. Case 3 remains the same.

Under the modified function that allows the empirical assumptions necessary to interpret B as the set of firing rates, there are at least two members of B that are guaranteed not to have an element from A mapped to them (namely, b_0 and b^{super}). If we include them in the domain of the relational system $\mathfrak{B}$ then we will have an isomorphic embedding. However, while they are necessary for defining

[146] More precisely, define $\mathfrak{B}' = \langle B', \geq_{BE} \rangle$ with $B' =$ the image of A under f_4. Then $\mathfrak{A}$ is isomorphic to $\mathfrak{B}'$ and isomorphically embedded in $\mathfrak{B}$. The function f_4 is a bijection from A to B' but not necessarily from A to B.

the function, we can also define $\mathfrak{B}' = \langle B', \geq_{BE} \rangle$, with $B' =$ the image of A under f_4. In that case we have an isomorphism to $\mathfrak{B}'$.

While more complicated from the perspective of measurement theory, the uncountable case can be handled swiftly from our perspective. What made the isomorphism from the countable empirical relational system to the numerical relational system possible is the order-density of the rationals in the rationals. What makes the isomorphism possible from the uncountable empirical system to the numerical system $\mathfrak{N} = \langle \mathbb{R}, \geq \rangle$ is the order-density of the rationals in the reals and the assumption that there is a countable order dense subset of *A*. We can extend the result of Krantz et al., from the material equivalence of 'there is a finite or countable order dense subset of A' and 'there is an isomorphism of $\mathfrak{A}$ into $\mathfrak{N}$', to the following:

(i) It is sufficient for an isomorphism from $\mathfrak{A}$ into $\mathfrak{B}$ for both A and B to have countable order dense subsets, and for $\mathfrak{A}$ and $\mathfrak{B}$ to be total orders.

So long as we make the empirical assumption that there are countable order-dense subsets of both A and B, and that $\mathfrak{A}$ and $\mathfrak{B}$ are total orders, then from the extension above, we are justified in concluding that isomorphism obtains between $\mathfrak{A}$ and $\mathfrak{B}$, even for the uncountable case.

To conclude, I emphasize two things. First, while the technical concepts from measurement theory are certainly useful for constructing a theory of representation, we must be clear about the conceptual assumptions underlying their use. For example, do we have good reason to assume that some empirical relation induces a total order? If not, we cannot define an isomorphism connecting that

system to another. For another example, if we accept non-finite relational systems, we can do nothing at all without the assumption of order density. Is that a legitimate empirical assumption? The second point that I would emphasize is that, in addition to getting clear on the underlying assumptions, we should also be clear that the concepts and results from measurement theory involve one empirical relational system and one numerical relational system, but not two empirical relational systems. To apply the results from measurement theory, we need to take additional steps to determine the legitimacy of those results for our special case. I've begun that project here, especially by constructing f_4, but I have only given an informal argument that f_4 is structure-preserving. A useful next step would be to demonstrate that with a rigorous proof. I end this appendix with a summary of the implications of the various empirical assumptions we might make, which can also be found in 5.4.6.

A.4 Summary of Results

Finite Case

F1. If $\mathfrak{A}$ and $\mathfrak{B}$ are weak orders with finite domains, then $\mathfrak{A}'$ and $\mathfrak{B}'$, the associated relational systems generated by constructing equivalence classes, are total orders.

F2. There exists $f: A/\sim \rightarrow B/\sim$, a bijection, such that $\mathfrak{A}'$ and $\mathfrak{B}'$ are isomorphic only if $[A/\sim] = [B/\sim]$[147].

F3. f defines a homomorphism from $\mathfrak{A}$ to $\mathfrak{B}'$.

[147] Notation: I use '[X]' to denote the cardinality of the set X. The set $A/\sim$ is the set of equivalence classes constructed of members of A.

F4. If $[A/\sim] > [B/\sim]$ then f defines an isomorphism* from $\mathfrak{A}'$ to $\mathfrak{B}'$. If $[A/\sim] < [B/\sim]$ then f isomorphically embeds $\mathfrak{A}'$ in $\mathfrak{B}'$.

F5. If we assume that $\mathfrak{B}$ is a total order, then $\mathfrak{A}'$ is isomorphic to $\mathfrak{B}$, $\mathfrak{A}$ is homomorphic to $\mathfrak{B}$.

F6. If we assume that both $\mathfrak{A}$ and $\mathfrak{B}$ are total orders then $\mathfrak{A}$ and $\mathfrak{B}$ are isomorphic. Similar remarks regarding the cardinality of A and B as those in F4 apply here.

Infinite Countable Case

C1. If $\mathfrak{A}$ and $\mathfrak{B}$ are countable total orders, and if B has an order dense subset then there exists a function $f: A \to B$ that defines an isomorphic embedding of $\mathfrak{A}$ in $\mathfrak{B}$ or an isomorphism of $\mathfrak{A}$ and $\mathfrak{B}$.

C2. If $B = Im(f|A)$[148] then $\mathfrak{A}$ is isomorphic to $\mathfrak{B}$. If $Im(f|A) \subset B$ then $\mathfrak{A}$ is isomorphically embedded in $\mathfrak{B}$.

C3. If we do *not* assume that $\mathfrak{A}$ and $\mathfrak{B}$ are total orders, but instead only assume weak ordering, then $\mathfrak{A}'$ and $\mathfrak{B}'$, the associated relational systems generated by constructing equivalence classes, are total orders. Then if $B/\sim$ has an order dense subset then there exists a function $f: A/\sim \to B/\sim$ that defines $\mathfrak{A}'$ as isomorphically embedded in $\mathfrak{B}'$. Under these assumptions, $\mathfrak{A}'$ and $\mathfrak{B}'$ are isomorphic only if $B/\sim = Im(f|A/\sim)$.

C4. Similar remarks apply here as to the finite case. If we assume total ordering on B but not A then $\mathfrak{A}'$ is isomorphic to $\mathfrak{B}$ and $\mathfrak{A}$ is homomorphic to $\mathfrak{B}$.

[148] Notation: '$Im(f|A)$' denotes the image of A under f.

C5. If there are upper and lower bounds to B (or $B/\sim$, as the case may be), that is, if we assume the existence of b_0 and b^{super} (or their associated equivalence classes), then if $b_0, b^{super} \in B$ then f defines an embedding. If $b_0, b^{super} \notin B$ then f defines an isomorphism, and mutatis mutandis for the function on equivalence classes.

C6. Let A^{RE} partition A, with A countable and A^{RE} finite ('RE' is intended to connote 'range equivalence'; see 5.4.6). Then the function described above for the finite case can be used to map range equivalence classes of A to members of B (with B finite), defining an isomorphism, embedding, etc. from $\mathfrak{A}^{RE} = \langle A^{RE}, \geq_{AR} \rangle$ to $\mathfrak{B}$. The relation $\geq_{AR}$ is defined in terms of $\geq_{AE}$, as $a \geq_{AE} b$ iff $a^{re} \geq_{AR} b^{re}$, where a^{re} and b^{re} are the range equivalence classes associated with a and b, respectively.

C7. Given range equivalence classes as above, the function that determines isomorphism from $\mathfrak{A}^{RE}$ to $\mathfrak{B}$ is associated with a function (that is, $f_1(a) = f_2(a^{re})$) that defines a homomorphism from $\mathfrak{A}$ to $\mathfrak{B}$.

Uncountable Case

U1. If A and B both have countable order dense subsets, and $\mathfrak{A}$ and $\mathfrak{B}$ are both total orders, then $\mathfrak{A}$ is isomorphic to $\mathfrak{B}$.

U2. Let A^{RE} partition A, with A uncountable and A^{RE} countable. Then the function described above for the countable case can be used to map range equivalence classes of A to members of B (with B countable), defining an isomorphism, embedding, etc. from $\mathfrak{A}^{RE}$ to $\mathfrak{B}$.

U3. Given range equivalence classes as above, the function that determines isomorphism from $\mathfrak{A}^{RE}$ to $\mathfrak{B}$ is associated with a function that defines a homomorphism from $\mathfrak{A}$ to $\mathfrak{B}$.

Appendix B: The Cognitive and Neurobiological Mechanisms of Vibrotactile Discrimination

B.0 Introduction

To understand mental representation and its place in the physical world, we need an understanding of *both* its constitution and implementation. If the theoretical conception of representation set forth in this dissertation is a good one, then the neurobiological mechanisms that underlie vibrotactile discrimination are in fact *cognitive* mechanisms. They are physical instantiations of both representation and computation. This makes an understanding of the neural implementation of those states equally as important as understanding their nature from a conceptual perspective.

In this appendix I provide a detailed literature review of the electrophysiological recording paradigm introduced in earlier chapters. I take the work of chapters 7 and 8 to establish that the states I mention below are physical vehicles of representation and computation, so I will use those words as appropriate.

B.1 The Task and its Psychophysics

Vernon Mountcastle and his colleagues developed a novel experimental strategy that combined psychophysical experiments with electrophysiological recordings, allowing the researchers to record the activity of single neurons in a behaving animal. The judicious placement of recording electrodes, combined with a clever experimental design that had macaque monkeys perform a cognitive task requiring the combination of sensory representations, memory, and motor plans, allowed the

researchers to begin probing into the neural encoding mechanisms that underlie the performance of a cognitive task. To my mind, this is nothing other than an experimental paradigm that has begun to shed light on the *nature* of representation, by discovering the neural mechanisms that implement the performance of a cognitive task. Most of the results discussed below come from the lab of Mountcastle's student and colleague, Ranulfo Romo.

The sense of "flutter" arises as a result of vibrating tactile stimuli in the range of about 5-50 Hz. One of the key results discovered early on was that humans and macaque monkeys (henceforth, simply 'monkeys') have similar detection and discrimination thresholds for vibrating flutter stimuli applied to the fingertip (Mountcastle, LaMotte, and Carli 1972, 204; Talbot et al. 1968; Mountcastle, Steinmetz, and Romo 1990; LaMotte and Mountcastle 1975). That is, suppose we were to apply the same stimuli, at different frequencies and amplitudes, to both humans and monkeys, and have them behaviorally signal judgments about them (is it higher or lower than a base stimulus to which it is compared, is it detectable as different than a base stimulus, etc.). Then, map those psychological judgments against the physical properties of the stimulus (this is called a *psychophysical curve*). From the psychophysical curves alone we could not, with any statistical reliability, tell whether the creature who made the judgment was a human or a monkey. In addition to the comparison of human and monkey psychophysical responses, Mountcastle and colleagues also discovered a great deal about which neurological mechanisms implement the sense of flutter. We will return to this shortly.

The basic, classical task (LaMotte and Mountcastle 1975; Mountcastle, Steinmetz, and Romo 1990) is as follows. A seated monkey has its left hand secured, palm up. A stimulator tip is lowered, indenting the glabrous (hairless) skin of one of the monkey's fingertips (it is not vibrating at this point). The monkey then presses a key with its free right hand, and holds the key down. The stimulator then produces a sinusoidal vibration, between 5 and 50 Hz, to the left hand fingertip (this is the *base*

stimulus, or *f1* for first frequency), followed by a delay period (or, *interstimulus interval*), followed again by a second stimulation (the *comparison* or *f2*). At the offset of the comparison stimulus, the monkey releases the key with its right hand, and signals its choice on which frequency was faster by pressing one of two push buttons located at eye level (the medial button signals the comparison is lower, the lateral button signals the comparison is higher). The monkey is rewarded with a drop of juice for correct discrimination.

Importantly, in the classic task the base stimulus was always at the same frequency. Hernandez et al. (1997) wondered whether in this task the animals were *discriminating* the comparison as higher or lower than the base, or simply *categorizing* it as something like "high" or "low". If the animals were discriminating rather than categorizing, they would be able to perform the task with similar levels of accuracy when the base stimulus was changed from trial to trial. However they were not able to do so, but required substantial further training (that is, as if they were learning a new task). Second, during training the base frequency used was always 30 Hz. In subsequent trials, the experimenters randomly switched the base frequency from 20 to 40 Hz. In these cases, the monkeys judged every comparison frequency less than 30 Hz as lower and every comparison greater than 30 Hz as higher, regardless of the base. Third, in separate runs the base stimulus was simply removed; in less than 50 trials the animals began to categorize the stimulus as greater or lesser than 30 Hz, with similar psychophysical curves to the original set. Thus, in the original task it seemed that the monkeys were simply ignoring the base stimulus, and were categorizing the comparison as high or low according to the arbitrary categories learned during training.

The same animals were then retrained, with the base stimulus changing from trial to trial. After training they were able to discriminate with similar levels as to the categorization task, even when the base stimulus was randomly varied and when the difference between base and comparison was

randomly varied. Accurate discrimination is around 75% for differences of 6 and 8 Hz, degrades a little at a 4 Hz difference, and is indistinguishable from chance levels at a 2 Hz difference. Thus, they cannot discriminate frequency difference at that fine of a level.

Some further physical parameters are psychologically relevant. The necessary minimum stimulus duration is about 250 msec. The use of stimuli that are 200 msec or less results in a substantial drop in performance levels. In the original task, the stimuli and interstimulus interval was always 1 sec. However, Hernandez et al. (1997) discovered that with intervals up to 10 sec, performance levels were similar; accuracy did not drop substantially until about a 15 sec delay period. Finally, stimulus amplitude is relevant. The initial indentation is 500µm. When stimuli are presented at different frequencies, human subjects notice a difference in intensity, even though amplitude is the same (LaMotte and Mountcastle 1975). So it is possible in principle to discriminate based on subjective intensity rather than frequency. To correct for this, amplitudes are adjusted so that they are all judged at equal subjective intensity (based on Mountcastle, Steinmetz, and Romo 1990). To confirm that the animals were attending to the frequency and not the amplitude, the researchers varied amplitude by measures much larger than those used in the original paradigm. The monkeys' performance in the discrimination task was identical to when the amplitudes were corrected for subjective intensity, thus providing solid evidence that they performed the task based on frequency discrimination, but neither on amplitude (i.e. subjective intensity) nor categorization.

This updated paradigm, using stimulus amplitudes corrected to equal subjective intensity (human subjects are used to construct these), and base frequencies that vary unpredictably thus forcing discrimination rather than categorization, is the one we will be discussing henceforth. Romo and Salinas (2003) argue that this paradigm provides optimal conditions for exploring neural codes: It involves a nontrivial cognitive task (sensory discrimination), humans and monkeys perform similarly on the task, it

uses simplified sensory stimuli, and through the sort of psychophysical analysis discussed above, we can be assured that the process involves working memory and discrimination, rather than long term memory and categorization.

> In principle, the task can be conceptualized as a chain of neural operations or cognitive steps: encoding the first stimulus frequency (f1), maintaining it in working memory, encoding the second stimulus frequency (f2), comparing it with the memory trace that was left by the first stimulus, and communicating the result of the comparison to the motor system (Romo and Salinas 2003).

B.2 Relevant Neuroanatomy

In glabrous skin, there are at least four different kinds of mechanoreceptor transducer organs and the afferent fibers with which they are associated (there are other cells associated with proprioceptive and nociceptive inputs, but we will not discuss them here) (Vallbo 1995; Gardner, Martin, and Jessell 2000; Gardner and Kandel 2000). They can be distinguished in two dimensions: rapidly or slowly adapting, and superficially or deeply located transducer organs. The slowly adapting receptors will fire, even incessantly, in response to constant pressure, whereas the rapidly adapting receptors will fire only in response to changes in pressure. Mechanical deformations of the skin and hence of the receptor organs results in the opening and closing of ion channels, itself resulting in changes in capacitance and current flow and ultimately voltage changes across the membrane, which results in the firing of action potentials. Rapidly adapting cells are those for whom these mechanical deformations

result in voltage changes, but very quickly the voltage changes return to the baseline state and action potentials no longer fire, even if the skin is still depressed. That is, they "adapt" to the new skin position. Slowly adapting receptors have the opposite property: the change in skin deformation will continually result in changes in current flow and voltage differences, and hence, the firing of action potentials. Rapidly adapting cells thus respond to vibrations and changes, whereas slowly adapting cells respond to constant pressure.

The deeply located transducer organs associated with rapidly adapting primary afferents are known as *Pacinian corpuscles*, whereas the superficially located organs are *Meissner's corpuscles*. Mountcastle and colleagues learned that the sense of flutter and vibration, and their associated frequencies (5-50 Hz and 60-300 Hz, respectively) are transmitted to the central nervous system (CNS) via distinct fiber groups[149]: the superficially located Meissner's corpuscles transmit flutter stimuli whereas the deeply located Pacinian organs and their associated axons transmit vibrating stimuli (that is, greater than 50 Hz) (Mountcastle et al. 1967; Talbot et al. 1968). Further, microstimulation of the peripheral fibers, where a small electrode is inserted into the skin and generates current trains and thus action potential spikes in the afferent fiber, reveals that even a single action potential from a single axon can be reliably perceived, with spatial and qualitative properties matching the type of fiber that was stimulated (Vallbo 1995). That is, if a rapidly adapting fiber associated with a Meissner's corpuscle located at the tip of the third finger was stimulated (in the arm), the human subject will report a sensation as of flutter localized to the receptive field of that fiber, on the third finger, and mutatis mutandis for other types of fibers and locations.

[149] 'Fiber' simply means *axon*. It is customary to refer to very long axons with 'fibers'; a *tract* is a group of fibers. An *afferent fiber* is one traveling towards the central nervous system, and an *efferent fiber* travels away, towards the periphery. A *primary afferent* is the first neuron in the chain of processing, whose endpoints are a transducer organ at the periphery at one end, and another neuron in the central nervous system (either in the spinal cord or further up into the brain) at the other.

The path to the CNS during the flutter discrimination task looks like this. Ambient mechanical energy at the fingertip is transduced by superficially located, rapidly adapting Meissner's corpuscles, which have small, well-defined receptive fields, into action potentials. The primary afferent is the neuron whose cell body is located in the dorsal root ganglion outside the spinal cord. It has two axons: the peripheral axon is connected to the transducer organ at the fingertip, and its central axon travels into the spinal cord, and up the spinal cord all the way to the medulla (the bottom portion of the brain stem). The second-order neuron synapses with the primary afferent, crosses over to the other side of the medulla and travels up to the thalamus, synapsing in the ventral posterior lateral nucleus of the thalamus. The third-order neuron then travels from the thalamic nucleus into the cortex, via the internal capsule, ending in S1, the primary somatosensory cortex (Gardner, Martin, and Jessell 2000; Romo, DeLafuente, and Hernandez 2004), contralateral to the stimulated finger.

Once in the cortex, the circuitry becomes dauntingly complex. It is not at all incautious to say that, at *every* level in cortical and even sub-cortical processing, neural operations are subject to both feedforward and feedback projections, and information processing is done in a massively parallel fashion. The path of cortical processing discussed below is a "suggested but still uncertain sequential cortical processing scheme" (Romo, DeLafuente, and Hernandez 2004, 198; caption to Figure 15.2).

The primary somatosensory cortex is composed of four areas, 1, 2, 3a, and 3b. Each area has a complete topographic map of the body's surface composed of the receptive fields of the respective neurons. Further, the specialization of peripheral fibers seems to continue in S1; neurons are classified in S1 as rapidly adapting, slowly adapting, or Pacinian, because their firing activities are similar to their respective primary afferents (Romo and Salinas 2001, 109). The areas associated with the rapidly adapting circuit here under consideration are areas 1 and 3b. Areas 2 and 3a are involved in other kinds of processing, such as pain and proprioception. Neurons in S1 are organized into columns, with similar

neurons with similar response properties in the same column. We will discuss some cortical microstimulation experiments shortly for which this is an important element. For the task under consideration, after the third-order neuron travels from the thalamus, it synapses with neurons in areas 1 and 3b of S1, which then exit S1 and travel to the secondary somatosensory cortex, or S2.

From S2, fiber tracts travel both to the supplemental motor area-proper (SMA-proper), which is the posterior part of the medial premotor cortex (MPC), as well as to the inferior convexity of the prefrontal cortex (PFC). Further, S2 appears to be serially connected to the ventral premotor cortex (VPC), which may then transmit its output to MPC (Romo, Hernandez, and Zainos 2004). From PFC, fibers travel to the pre-SMA, the anterior part of MPC. From MPC, fibers travel to the primary motor cortex (M1), and from there back down through the cortex, through the internal capsule and through various areas in the brain stem, crossing over at the medulla, back down the spinal cord and out the ventral roots to innervate the right hand, which then presses one of the two push-buttons to signal the monkey's decision, as discussed above. We will not focus on any of the activity downstream of M1, where the motor commands from M1 are transformed into specific muscle motions at the periphery. Instead we will focus on the neural mechanisms implementing the encoding of sensory stimuli, working memory, a decision process, and the sensorimotor transformation resulting in a general motor command in M1.

Here's a quick review. Peripheral activity travels to areas 1 and 3b of S1, then to S2. The outgoing signal from S2 then gets widely distributed, to at least PFC, MPC, and VPC; PFC and VPC both appear to be serially connected to MPC. Then MPC transmits activity to M1, whose activity ultimately results in the monkey's button-pressing behavior signaling its choice. Consistent with many other studies in this and other modalities, these cortical areas are typically associated with cognitive activities in the following way. Primary and secondary sensory areas are involved in sensory processing. PFC is

widely implicated in short-term or working memory processes, and MPC/VPC are considered to be pre-motor areas, which begin the transformation of signals from sensory and memory processes into motor plans. Primary motor areas are associated with the implementation of generalized motor plans, which then get refined into more specific muscle commands, taking into account various feedback mechanisms by the basal ganglia, cerebellum, and spinal cord.

B.3 Sensory Encoding Mechanisms

The experimental paradigm under consideration is designed to allow investigation into the neural bases of several cognitive processes: working memory, comparison and decision procedures, and sensorimotor transformation. However, all subsequent processing is constrained by the original sensory encoding procedures. So we begin by asking how the sensory stimulus is represented in the CNS.

In the early investigations, Mountcastle and colleagues found the rapidly adapting neurons of S1 to be strongly phase-locked to the stimulus. That is, the neurons fired a spike or burst of spikes for each sinusoidal wave of the stimulus (Mountcastle et al. 1969; Mountcastle, Steinmetz, and Romo 1990). However, they did not find neurons that significantly modulated their firing rate as a function of the changing stimulus frequency. From these and other studies with similar findings (LaMotte and Mountcastle 1975; Recanzone, Merzenich, and Schreiner 1992), the hypothesis was proposed that rapidly adapting neurons in S1 encode stimulus frequency in the *temporal structure* of the neural firing, but not in average firing rate. More specifically, the proposed hypothesis is that periodicity of the neural firing constitutes the code for representing stimulus frequency. I'll sometimes refer to this as a

temporal code, which is distinct from a *rate code*. Thus, neurons central to S1 are hypothesized to "read off" the temporal structure of S1 neural firing in order to estimate stimulus frequency.

This hypothesis went unchallenged for several years, however, we should note that it was generally based on a small sample size (only 17 neurons were recorded during the task condition) (Mountcastle et al. 1969), or on anesthetized animals (Recanzone, Merzenich, and Schreiner 1992). Romo and colleagues later returned to this hypothesis to see if it would be confirmed by a larger sample of recorded neurons, while the awake animal was engaged in the vibrotactile discrimination task (Salinas et al. 2000; Hernandez, Zainos, and Romo 2000). What they found instead was that apparently both the temporal code and a rate code are able to adequately encode the stimulus frequency, although it is not clear that the animal, or neural processes downstream of S1, make use of periodicity.

B.3.1 Firing Rate and Periodicity in S1

Single-cell intracortical recordings were made while the trained animal performed the vibrotactile discrimination task. Microelectrodes are inserted into the cortex of the monkey, and aimed at the distal fingertip areas of the part of the cortex under investigation. The stimuli are then placed on the fingertips, at the center of the receptive field for the neuron being recorded. In S1, neurons have small, well-defined receptive fields, so they aim for the center of the distal fingertip area. In S2, neurons have much larger receptive fields, generally spanning all fingertips, sometimes reaching into the forearm and bilaterally. So the stimulus is not aimed at the center of the receptive field (it would be hard to determine if there is a "center"), but is also placed at the fingertip. Similar considerations apply to the more central cortical areas with very large receptive fields. Neurons are selected for study if they react in any way (relative to background noise) upon either base or comparison stimulus, or during the interstimulus interval. Electrode placement is later confirmed by standard histological techniques: The

animal is euthanized, the cortex is thinly sliced and then stained, and the area in which the electrode was placed can be identified by the track marks in place, and then confirmed as being in the desired anatomical location of the targeted cortical area.

Periodicity is the property of exhibiting regular, repeating characteristics. For each spike train elicited during stimulation, a power spectrum was computed using Fourier decomposition[150]. They took the median frequency around the peak power frequency, and used that as a measure of periodicity (Hernandez, Zainos, and Romo 2000, 6192). In plainer language, computing a power spectrum gives you the component sine and cosine functions that any given function is composed of, as well as the "power" or amount that each frequency contributes to the original function. It is done using *frequency bins* which are each a small range of frequencies. The peak power frequency bin is the bin that contributes the most to the original function. Taking the median frequency from that bin simply means finding the center of the bin. Hernandez and colleagues called this a quantitative measure of periodicity, but really it is an estimate of stimulus frequency, based on the fact that the stimulus is periodic. Thus, if the neural response is perfectly phase-locked to the stimulus, then the median of the peak power frequency will be identical to the frequency of the stimulus. Hence, this is not a measure of periodicity in the sense in which we might say "more or less periodic"; rather, it is an estimate of the stimulus frequency.

In addition to computing periodicity as discussed above, they computed the average firing rate simply by counting the number of spikes over the stimulus period (500 ms) and dividing by that time.

[150] Fourier analysis is the process of decomposing a function into component sine and cosine functions. A Fourier decomposition is thus a mapping from functions to functions, in such a way that the original function can be reconstructed from the basis functions. In order to recreate the original function from the basis functions, both the phase (the starting point at the zero x-coordinate of the sine wave) and the amplitude or power (or amount) that each function (and hence, frequency) contributes to the original function, must be specified. A power spectrum density shows how much each particular frequency contributes to the original function. Fourier decomposition can also be thought of as mapping a function in the time domain to a function in the frequency domain, essentially showing how much each frequency contributes to the original function. For example, the power spectrum for a single, 10 Hz frequency sine wave would include 10 Hz at 100%, and every other frequency at 0%.

Then, for each stimulus frequency, they calculated the mean periodicity and mean firing rate over all trials with that frequency. They selected the neurons that had a significant linear fit of the periodicity and/or firing rate as a function of stimulus frequency, and required the slope to be significantly different than zero, as the neurons for study.

The above two measures are ways of correlating firing rate with stimulus frequency and periodicity with stimulus frequency, by determining whether the neural responses are correlated as functions of frequency. Importantly, they also sought to compare neural responses with behavioral responses. They do this by calculating what they call *neurometric curves*, which plot the probability that an ideal observer using (for example) firing rate could correctly discriminate the two stimuli, and then compare that to a psychometric curve, which plots the animal's discriminatory behavior against the frequency difference, using the same dimensions. Neurometric curves for firing rate were computed using the following rule: if there are more spikes during f_2, then f_2 is higher. For periodicity, the rule was, if the periodicity value is higher during f_2, then f_2 is higher. Although the measurement of periodicity is really an estimate of stimulus frequency, for the purposes for which it is used, it is a legitimate measurement. If the neuron is encoding stimulus frequency in the temporal arrangement of its spikes, then since the stimulus is periodic, the ideal observer would be able to determine which of the two presented stimuli are faster from the temporal arrangement of the spikes alone.

Finally, the last measurement is the discrimination threshold. This is the difference between base and comparison that is necessary for the animal (or neuron) to correctly discriminate 75% of the time. As far as I understand it, the 75% threshold is chosen arbitrarily.

The results they found are as follows. Out of 223 rapidly adapting neurons tested in S1 (from both areas 1 and 3b), they found 188 that satisfied the criteria above in terms of responding as a

function of frequency (either in their firing rate or periodicity). Out of the 188, 139 responded with periodic spike intervals, 72 with firing rate modulation as a function of frequency, and 23 satisfied both criteria. The authors note that "it is important to remark that previous studies had not reported neurons with aperiodic, stimulus-dependent firing rate responses" (Hernandez, Zainos, and Romo 2000, 6193). The neurometric discrimination thresholds for the periodic neurons were much lower than the animal's discrimination threshold, whereas the firing rate thresholds were very similar to the animal's discriminatory thresholds. Thus, the neurons with electrical activity that covaries in its temporal regularity as a function of stimulus frequency are *more* discriminating than both the neurons whose firing rate covaries as a function of frequency, as well as the behaving animal.

Using an aperiodic version of the task, where the vibrating stimulus had the same mean number of indentations but the time intervals between amplitude peaks varied randomly, unsurprisingly they found no neurons that significantly covaried, according to their temporal structure, with frequency. However, of all of the neurons that had significant rate modulations during the periodic version of the task, every one had similar rate modulations during the aperiodic version as well. So these neurons seem to be using the common code of firing rate for both the aperiodic and periodic versions of the task. It should be noted that the animals perform equally well on the aperiodic version with no further training needed, even though they are trained on the periodic version of the task.

The authors conclude from this study that it appears that, among the population of rapidly adapting neurons of S1, there are two subpopulations. The one population, which is likely closer to the periphery in terms of its number of thalamic inputs, is more closely phase-locked to the stimulus. Further, the second appears to "read off" the time course of the earlier subpopulation and integrate that temporal code into a rate code. However, even though the temporal code enjoys greater discriminatory ability, and the animal is rewarded for correct discriminations, it seems that the

"monkeys do not *use* this exquisite representation [of the temporal structure of the stimulus] for frequency discrimination" (Hernandez, Zainos, and Romo 2000, 6195, my emphasis).

B.3.2 Further Measures of Periodicity and Comparison of S1 and S2

The study discussed above used only one measure of periodicity, and the major comparison between firing rate and periodicity is in terms of the neurometric versus psychometric discrimination thresholds. The Romo group also completed some further comparisons of periodicity and firing rate, and compared the responses of S1 and S2 neurons with the animal's behavior (Salinas et al. 2000).

In this study, four quantities are derived from the power spectrum: the power at stimulus frequency and at twice the stimulus frequency, the maximum power, and the frequency (x-coordinate) at peak power (the y-coordinate). This last quantity will be denoted with 'PSFP' (for power spectrum frequency at peak). Notice that PSFP in the Salinas et al. study is identical to what the Hernandez et al. (2000) authors simply termed 'periodicity'. The first three quantities above are measures of periodicity, while the last is an estimate of frequency of the stimulus, based on the periodicity of the evoked spike train. Firing rate is calculated in the usual way.

One way to compare the representational capacities of firing rate to periodicity is to separately compare them each to stimulus frequency and see how well they map onto it, as a function of frequency, as was done in the Hernandez (2000) paper. But this is really only an intuitive comparison. To quantify the strength of association of firing rate with stimulus frequency as well as periodicity[151] with stimulus frequency, and to compare that strength in the same units, the authors calculated Shannon's mutual information between each quantity and stimulus frequency.

[151] Note that for every calculation performed with any measure of periodicity, the same was made for every other measure of periodicity, although they do not always report the results. The most usual quantity we will deal with is the PSFP.

I have earlier criticized (2.4.2) each of the notions of information as being non-objective in a way that is uniquely threatening for a naturalistic reduction of mind, including Shannon's mutual information. However, we should be clear about the way that concept is used here. The authors of this study are attempting to compare, in the same terms, the strength of association between distinct quantities, each with stimulus frequency. To do that, they need several antecedent probabilities: they need the probability of the stimulus, the probability of the response, and the conditional probability of the response given the stimulus. For this last, they make the standard assumption that the response (i.e., firing rate or periodicity) admits of a normal distribution. For the probability of the stimulus, since they only use 8 different frequencies, and the study is designed in such a way that every frequency arises an equal number of times, they calculate the probability of the stimulus as 1/8. Clearly, any quantification of the "amount of information" that firing rate (say) can carry, which is based on the assumption that there are only 8 (equally likely) possible frequencies in the universe, is unjustified.

However, three points should be kept in mind. First, these scientists are not attempting to use the concept of information for a philosophical, naturalistic reduction of mental representation, but instead as a method of comparing how well two distinct quantities covary with a third. Second, the assumption that no frequency other than the 8 frequencies used is a "relevant possibility" makes the quantification of information into a non-objective quantity: That is, it is a quantity that depends on a cognitive agent to decide that there are only 8 equally likely, relevant possibilities. However, since the same assumptions of there being 8 equally likely possibilities, normal distributions, etc., are made for both the various measurements of periodicity as well as firing rate, it does seem justified to compare the two, relative to each other. So we can say, for example, that periodicity bears a stronger association with stimulus frequency than firing rate (if that is what bears out). Finally, recall my discussion in 5.5: even though mutual information is not an objective quantity, it can be used as an epistemic guide to the

existence of an objective, lawfully grounded regularity between two kinds of events. Thus, even though such and such a particular quantity of information measured is not objective, the fact that given the same assumptions, we get the same results in a statistically robust way, can be used as evidence that a nomically grounded covariation does indeed exist between firing rate and stimulus frequency, and between periodicity and stimulus frequency, even though we cannot quantify the strength of that association in an objective manner. I therefore include discussion of the information calculations, although these clarifications must be kept in mind.

The results they found were as follows. In S1, firing rate increases as a function of frequency, approximating a linear function. The mean PSFP is very close to identical to the stimulus frequency, for each stimulus frequency presented. This demonstrates that in S1 there is a high level of periodicity, and the neural responses are strongly phase-locked to the stimulus, consistent with the findings of the Hernandez (2000) paper. In principle there was six times more information available in the periodicity of the evoked spike trains than in the firing rate.

In S2, firing rate modulation as a function of frequency also occurs, and also carries a statistically significant amount of information about stimulus frequency. There are some differences between S1 and S2. First, the amount of information carried by firing rate in S2 is less than in S1, by a factor of about 2. Second, periodicity is significantly reduced. Although some (7.5%) neurons carried an amount of information that was significantly different from zero in the periodicity of their spike trains, the mean PSFP was independent of stimulus frequency for each frequency. This contrasts with the PSFP in S1 which was a function of frequency (it was approximately equal to stimulus frequency). While the PSFP did carry some information in S2 (that is, only for those 7.5% of neurons), it was less than that in S1 by a factor of about 10.

An active versus a passive condition was compared. The active condition is the task described above; in the passive condition the otherwise free hand was restrained, no discrimination was made, and no reward was offered. In S1, the information from firing rate was significantly higher in the active as compared to the passive trials, and the mean variability in firing rate across trials was significantly lower (thus, firing rate is a more reliable code when the animal is actively discriminating). Similar changes were found with respect to firing rate in S2.

The differences in periodicity were more subtle in S1. There was no statistically significant change in the information to be found in PSFP across conditions, however, the mean power at stimulus frequency was significantly greater in the active than the passive condition (hence, the spike trains were more periodic in the active condition – tighter phase-locking occurred when the animal was actively discriminating). No changes were found in the periodicity in S2, which was to be expected. Thus, the behavioral context modulated both encoding mechanisms, which therefore favors neither when trying to decide which code is used by more central cortical mechanisms.

So far the experimental findings of this paper don't point one way or the other with respect to the coding mechanisms used. Both periodicity and firing rate are reliable indicators of stimulus frequency, both are functions of frequency, and both are modulated by attentional effects. In principle there is more information available in periodicity, in S1. The final comparison is that between neuronal and behavioral responses. If firing rate is used as an encoding mechanism, then there should be a significant difference between a standardized measurement of firing rate between hits (accurate discriminations) and errors. If periodicity is used, the same should be found with respect to periodicity.

The researchers found that in both S1 and S2, there was no significant difference in any of the three measures of periodicity between the hit and error conditions. When compared across populations

of cells rather than individual neurons, no significant correlation was found between any measure of periodicity and behavior. In both S1 and S2, there was a significant correlation in the population measures, between the difference in standardized rates between the hit and error trials of both types (base greater than comparison and vice versa). Further, in S2 there was a significant difference in standardized rate measures in individual neurons, between hit and error trials, for trials of both types. In S1 there was a significant difference between trials where the base is greater than the comparison. The number of neurons found with a significant difference in the other type of trial (comparison greater than base) however was within that expected by chance.

Thus, if the animal made an error in discrimination, signaling that it found the comparison to be lower than the base when in fact the comparison was higher than the base, there is a significant likelihood that individual neurons in S1 and S2 were firing at a rate that is lower than they would have been firing, given that frequency, had the animal discriminated correctly. This correlation between neural activity and behavior holds for both individual neurons and the behavior of neurons averaged together and considered as a population. Similarly, when the animal signaled that the comparison was higher when in fact it was lower, the behavior of individual neurons, when measured at the individual level and when averaged together into population measures, can predict that the animal will be in error, because those neurons will be firing at a faster rate than they would be, for that frequency, when the animal makes a correct discrimination judgment. By contrast, none of the periodicity measurements, in either S1 or S2, bore out this same correlation between animal behavior and neural behavior, even though the temporal structure of the spike trains covaries with the stimulus frequency.

This covariation between the animal's behavior and firing rate, the lack of covariation between the animal's behavior and periodicity, the similarity in discrimination thresholds between animal and firing rate, the difference in thresholds between animal and periodicity, and finally, the similarity of

firing rate-based neurometric curves with psychophysical curves (Hernandez, Zainos, and Romo 2000), all provide strong evidence that cortical mechanisms use the rate code in frequency discrimination, but not the temporal code. Even though in principle the temporal structure seems to make for a better code, the monkey and/or its brain seems to use the rate code. It is important to remember however that the recordings are taken from trained animals who are good at performing this task. The fact that periodicity is not used in this task does not imply that it is not used elsewhere. For example, humans can readily distinguish periodic from aperiodic stimuli, so it seems that the temporal structure of the spike trains would have to be used in that case.

B.3.3 Subpopulations in S2

An important difference in neural processing arises in S2, which was not present in S1. The neurons in S1 all have positive "slopes". That is, with respect to those neurons whose firing rate is modulated by stimulus frequency, in every case, increasing stimulus frequency results in increasing firing rate in an approximately monotonic fashion. In S2, first, the majority of the neurons do not significantly covary their periodic time structure with the stimulus frequency (see above: only 7.5% do). Instead, the vast majority appear to encode stimulus frequency with a rate code. However, slightly less than half fire preferentially with lower stimulus frequencies. That is, they have a negative slope: increasing stimulus frequencies are associated with decreasing firing rates, in an approximately monotonic fashion (Salinas et al. 2000).

This separation of neurons into functionally segregated subpopulations may play an important role in neural processing, as it is also found in other more central areas (to be discussed shortly). Further, Romo et al. (2003) have argued that having sets of neurons with opposite tuning properties may in fact be beneficial for stimulus encoding and for frequency discrimination, by mitigating the

problem of correlated noise in populations of neurons. Random noise is generated constantly, since the movements of ions across cell membranes, the opening of some ion channels, and vesicle release, are stochastic processes. Further, random correlations of that noise across populations of cells decreases the coding efficiency of that population by decreasing the signal to noise ratio (Zohary, Shadlen, and Newsome 1994). However, when the random noise fluctuations are correlated across oppositely tuned populations, the noise cancels itself out, thus decreasing the amount of noise, increasing the signal to noise ratio, and increasing the coding efficiency of the population of neurons considered as a whole (with both positive and negative populations as subsets of the larger, more efficient population). Thus, as a network property, having oppositely tuned populations appears to be an adaptation that allows for increased efficiency in information processing.

B.3.4 Artificial Percepts Generated Through Cortical Microstimulation

Thus far, we've discussed some fairly good reasons to believe that the firing rate of the rapidly adapting neurons of S1 and their associated counterparts in S2 underlies the cognitive task of somatosensory discrimination of flutter stimuli: There is a significant correlation between firing rate modulations as a function of stimulus frequency, there is a good fit between psychometric and neurometric thresholds generated by firing rate, and there are correlations between standardized measures of firing rate, both at the individual and population levels, with the animal's behavior. Additionally, lesion studies demonstrate that without S1, animals have severe impairments in discrimination and categorization tasks (LaMotte and Mountcastle 1979; Zainos et al. 1997). Hence, the activity of the neurons in S1, which has been shown to be correlated in a regular way with stimulus frequency, is necessary for the tactile discrimination task. Providing even stronger evidence,

microstimulation studies have demonstrated that firing rate in S1 is *sufficient* to generate the entire set of neural events that underlie the flutter discrimination task.

Romo and colleagues (Romo et al. 1998; Romo et al. 2000) used implanted electrodes not only to record, but also to inject pulses of current into S1. While the monkeys performed the vibrotactile discrimination task, electrical stimulation pulses, oscillating at the same frequency as the mechanical stimulation would have been, randomly replaced the mechanical stimulus. Thus, in some trials the monkeys performed the standard task with natural stimuli, and in others, direct cortical microstimulation of the rapidly adapting neurons of S1 was used. The authors found that the monkeys were able to perform the discrimination task with accuracy levels that were indistinguishable from the natural, mechanical stimuli alone.

The microstimulation consisted of injected bursts, with 2 spikes per burst, at an amplitude of 65 µamps or more. With less than 40 µamps, the monkeys just waited as if no stimulus had been presented, and at amplitudes between 40 and 65 µamps, the monkeys performed the task, but discrimination was at chance levels. They also used aperiodic electrical stimuli as a comparison, and found that there was a slight but significant difference. Monkeys were able to discriminate at 84% correct for periodic but 81% for the aperiodic condition (Romo et al. 1998).

Since the base frequency changed from condition to condition, the monkeys would not be able to successfully perform the task unless they discriminated, as opposed to categorized the stimulus (Hernandez et al. 1997). Further, the animals were continually switched from natural to artificial stimuli, with no change in performance. The authors conclude that such high performance "is consistent with the induction of an artificial percept … Thus, the microstimulation patterns used may elicit flutter

sensations referred to the fingertips that are not unlike those felt with mechanical vibrations" (Romo et al. 1998, 389-390).

Notably, in the 1998 study the base stimulus was always mechanical, while the comparison stimulus was switched between mechanical and direct cortical stimulation. In a follow-up, Romo and colleagues (Romo et al. 2000) also used electrical stimulation for the base stimulus, as well as for both the base and the comparison. This is an important difference: In the original study using electrical stimulation for the comparison only, the results suggest that direct cortical stimulation can be used to induce an artificial percept that can be compared in working memory to a previous (natural) percept. By contrast, showing successful discrimination behavior with an artificial base stimulus shows not only that artificial stimulation is sufficient to induce something like a sensation, but also that this can be stored in working memory, and that working memory trace can be compared to a subsequent (natural) stimulus. Finally, that the animal can perform the task based entirely on artificial stimulation shows that the induced activity in S1 alone is sufficient to cause the entire chain of neural/cognitive events leading to successful discrimination, including sensory representation, working memory, comparison and decision, sensorimotor transformation, motor plans, and motor output.

They found just that. The monkeys were able to perform at levels well above chance with an artificial base stimulus, with no significant difference in accuracy levels from the solely mechanical version. Further, the entire cognitive task could be performed with artificially injected current alone. There were however some differences in this latter case. For artificial base and comparison as compared to mechanical base and comparison, the animals were correct 80% vs. 89% of the time, where that difference is significant. Also, the psychophysical threshold (the smallest frequency difference they could detect 75% of the time), was very slightly smaller for mechanical (2.88 Hz) than artificial stimuli (3.73 Hz), with the difference significant.

Finally, providing further evidence of the specialization of this rapidly adapting circuit, it should be noted that this only works when the microelectrode stimulates the rapidly adapting columns, but not the slowly adapting columns. When stimulation is at the slowly adapting columns, the monkeys respond as if they recognize stimulation, but are at chance levels for discrimination. When the electrode is at the border between slowly adapting and rapidly adapting columns, their behavior is better than chance but less than it is with rapidly adapting columns.

Thus, modulations in firing rate are correlated with stimulus frequency and with animal behavior. They are necessary for performance of the task, but further, stimulation of the rapidly adapting neurons of S1 (those anatomically connected to the Meissner's corpuscles) is sufficient to initiate the entire chain of cognitive and neural events associated with sensory flutter discrimination, at accuracy levels indistinguishable from the natural, mechanical version of the task.

B.4 Working Memory

To successfully discriminate the first from the second tactile stimulus, and decide which has a greater frequency, the animal must maintain a mnemonic trace of the first stimulus. Further, it must be held in an informational code that can be readily compared with a sensory encoding of the second stimulus. In other words, ideally we would think that the mnemonic trace and the second sensory encoding are both in the "same language". As discussed above, the delay period can be extended to as much as 10-15 seconds before performance levels begin to drop appreciably (Hernandez et al. 1997), thus, the working memory component of the task is relatively short-lived.

Previous studies have demonstrated a strong correlation between activity in the prefrontal cortex (PFC) and working memory, for auditory (Bodner, Kroger, and Fuster 1996) and visual (Funahashi, Bruce, and Goldman-Rakic 1989; Miller, Erickson, and Desimone 1996) modalities. There appears to be a sub-specialization: tasks involving spatial properties (as opposed to object identification) find implementation of working memory in the dorsolateral areas of the PFC, whereas tasks not involving a spatial component, but do involve object identification, seem to be implemented more ventrally, in the inferior convexity (Wilson, O'Scalaidhe, and Goldman-Rakic 1993; Rao, Rainer, and Miller 1997). To investigate its role, and to get an understanding of the behavior of individual neurons that may implement working memory, the Romo group recorded from PFC while trained monkeys engaged in the flutter discrimination task (Romo et al. 1999). Consistent with the hypothesized functional specialization, they did not find neurons in the dorsolateral PFC that modulated their activity from baseline during any part of the task. On the contrary, they found several hundred (439) in the inferior convexity of the PFC that did respond during the task.

To get a handle on the relationship between neural activity and stimulus frequency, they recorded the activity of each neuron that fired significantly different from baseline during the delay period, when no stimulus was presented. They used a regression analysis[152] to see if the changes in firing rate can be modeled as a function of stimulus frequency (f_1); results are as follows. Of the 493 neurons, 65% were found to fire (during the delay period) as a monotonic function of stimulus frequency, either as a linear or sigmoidal function. Approximately half of this monotonic population had

[152] Regression analysis is a statistical tool that is used to judge how well one can predict the value of a dependent variable, given the value of an independent variable. The conceptual underpinning of this technique is that regression analysis seeks to determine whether regular and reliable correlations exist between different variables. In the example above, if there is a reliable covariation between the value of the frequency and the value of the firing rate, then if we knew the value of the frequency, we would be able to predict the value of the firing rate as a function of frequency.

a positive slope and the other half had a negative slope. About 40% of the total population exhibiting a monotonic response could be fit as a linear function of stimulus frequency, and the rest as sigmoids.

The above described responses during the delay period were not fixed, however. Rather, according to when in the delay period they fired, the neurons can be categorized as *early, persistent*, or *late*. Early neurons fire (as a function of f_1) during the first third of the time period but not during the remainder, late neurons fire in the final third but not the first two thirds, and persistent neurons fire as a function of f_1 throughout the delay period. 23% of the active neurons recorded were persistent, 34% early, and 33% were late. Importantly, the duration of the delay period was fixed at 3 sec for the majority of the trials. To test whether the temporal dynamics of these neural responses are fixed or are relative to the time scale of the task, the authors also studied some of the same neurons while the animals performed the task with a 6 sec delay period. They found that the behavior of the late neurons is not time-locked to the beginning of the trial, but shifts in proportion to the delay period, waiting longer to fire strongest when the delay period is lengthened (Romo et al. 1999).

Finally, this activity should not be interpreted as preparatory or anticipatory motor responses. The same experiment was performed with a stimulus set designed such that the correct motor response could not be predicted from the base stimulus alone. That is, the probability of the comparison being lower than the base was .5 in that stimulus set, and it would therefore be impossible to successfully perform the task (at better than chance levels) by simply noting the base stimulus and then using that to determine which button to press. In this control experiment the animal performed similarly as before and the results were similar. Thus, the activity found in PFC during the delay period, which is a function of the first stimulus frequency, reflects a working memory component of the task.

While widely implicated in working memory, the PFC is not the only cortical area that is active during the delay period. Hernandez et al. (Hernandez, Zainos, and Romo 2002) later found a population of neurons in the medial premotor cortex (MPC) whose firing during the delay period could be adequately fit as a monotonic function of f_1. As in PFC, this population was split approximately in half in terms of the slopes: Some fired at greater rates when f_1 was higher, and some at greater rates when f_1 was lower. However, while the temporal characteristics (early, persistent, or late) of PFC were approximately split into thirds, in MPC the majority (60%) were late neurons. Further, Romo et al. (Romo, Hernandez, and Zainos 2004) found similar results in ventral premotor cortex (VPC), another pre-motor area that is believed, like MPC, to participate in linking sensory and memory events with motor actions. Finally, delay period activity is found even as early in cortical processing as S2, though not in S1 (Salinas et al. 2000; Hernandez, Zainos, and Romo 2000). The delay period neurons of S2 are split approximately in half in terms of their slopes, however, all of the delay neurons found in S2 are early neurons. Thus, information about the first stimulus is not held in either S2 or S1 throughout the delay period.

B.5 Comparison and Decision Procedures

The task under consideration is performed successfully when the animal holds a mnemonic trace of the first stimulus, f_1 throughout the delay period, and then compares that memory with the second stimulus, f_2. Upon comparison, the animal then decides which of the two frequencies was greater. Where and how does this comparison and decision procedure take place?

Gold and Shadlen (2001) have hypothesized that a simple subtraction procedure among populations of neurons could be used to decide among competing sensory hypotheses. To study this quantitatively, the Romo group again had trained monkeys perform the discrimination task while using single-cell recording techniques in various brain areas (Hernandez, Zainos, and Romo 2002; Romo, Hernandez, Zainos, Lemus et al. 2002; Romo, Hernandez, and Zainos 2004). They again used linear regression analysis to see if the neural firing rates could be fit as a function of the stimulus. They used multiple linear regression analysis, fitting firing rate to the following equation:

$$R = a_1 f_1 + a_2 f_2 + c$$

where *c* is a constant, f_1 and f_2 are the frequencies of the base and comparison stimulus, respectively, and a_1 and a_2 are coefficients that determine the strength of the relationship between *R* (firing rate) and stimulus frequency. Four values are of particular importance here. When a_1 is significantly different from 0, then there is a correlation between firing rate and the frequency of the base stimulus; when $a_1$is different from 0 yet a_2 is at 0, then the firing rate is solely a function of f_1. Mutatis mutandis for a_2 and f_2. When both a_1 and a_2 are significantly different from 0, the firing rate correlates with some combination of the base and comparison stimulus. Finally, when $a_1 = -a_2$, then firing rate is now correlated with neither f_1 nor f_2, but with the *difference*, $f_2 - f_1$.

Romo and colleagues (2002) performed this experiment while recording from S2, and computed the multiple linear regression analysis as above, over a sliding time window during the comparison period (i.e., during the 500 ms when f_2 is presented). They found 208 neurons with activity significantly different from baseline firing. The neural behavior during the comparison in S2 is not static: in the

beginning (first 200 ms), the neurons fire either as a function of f_1 or f_2 (that is, either a_1 or a_2 is significantly different from 0). During the final 300 ms however, the regression analysis shifts to the diagonal axis (where $a_1 = -a_2$), hence, the firing rate is no longer correlated with either f_1 or f_2, but with the difference, $f_2 - f_1$. Of the 208 neurons here under consideration, during the initial 200 ms, 12% were f_1-dependent, 48% were f_2-dependent, 17% were $f_2 - f_1$-dependent, and the rest could not be unambiguously classified. The responses during the final 300 ms were slightly more complicated. 20% were unambiguously functions of $f_2 - f_1$, and 13% were unambiguously functions of f_2. On an individual level, the remaining 67% were in an intermediate group: their firing was a function of both f_2 and f_1. From a population measure, however, the firing rate of the population as a whole, toward the end of the comparison period, became more clearly a function of $f_2 - f_1$.

Importantly, the Romo group found, as before, two populations of neurons (Romo, DeLafuente, and Hernandez 2004). One population fired more strongly when $f_2 - f_1$ is positive (i.e., $f_2 > f_1$), and another fired more strongly when $f_2 - f_1$ is negative. This is consistent with the Gold and Shadlen (2001) subtraction hypothesis: two subgroups exist with opposite response slopes, and the overall decision procedure can be computed by a competition between the two (or a subtraction operation).

It is also important to note that information about f_1 is not maintained in either S1 or S2 throughout the delay period. Thus, finding neurons in S2 that fire as a function of f_1 during the comparison period provides extremely strong evidence of a feedback mechanism from PFC, MPC, or VPC (or another area, as yet unknown, that maintains an encoding of f_1 in working memory). Thus far, it seems that S2 is involved in the comparison process and decision procedure. What other areas do so as well?

As discussed above, the pre-motor areas VPC and MPC are believed to participate in linking sensory and memory processes with motor plans for output. This, along with their anatomical connections to S2, makes them good candidates for further investigation in the vibrotactile discrimination task. The Romo group did just that, and found similar results as above. Namely, in MPC, 139 of the 264 neurons with strongest responses during the comparison period fired as a function of $f_2 - f_1$ throughout the entire comparison period. 81 initially encoded f_1 and then shifted to the diagonal, and 29 initially encoded f_2 and then shifted over to encode $f_2 - f_1$ (Hernandez, Zainos, and Romo 2002). In VPC, similar responses were found (Romo, Hernandez, and Zainos 2004). Importantly, as in S2, opposite populations are found, which fire more strongly either when $f_2 - f_1$ is positive or when it is negative. This dual-encoding scheme seems to be a prominent feature of neural computation, as we have found it in sensory encoding, working memory, and during the comparison procedure. Apparently, these $f_2 - f_1$-dependent signals are to be found even in the PFC (Romo, DeLafuente, and Hernandez 2004).

Finally, it is important to note that the entire set of neural/cognitive events is to be found, to a greater or lesser degree, in each of the areas thus far mentioned (excluding S1 and M1, to which we turn next). Sensory encoding, working memory, a comparison process (i.e., where firing rate is a function of both f_1 and f_2 but not of $f_2 - f_1$), and a subtraction/decision procedure (where firing rate is a function of $f_2 - f_1$) are to be found in S2, PFC, VPC, and MPC, although each area does appear to have specializations. S2 for example does not encode the base stimulus throughout the entire delay period, but PFC does; MPC and VPC mostly encode the base during the latter portion of the delay period. Most of the comparison process in S2 is strictly comparison, rather than $f_2 - f_1$ – dependent, whereas in VPC and MPC, most of the neurons active during the comparison period are purely $f_2 - f_1$ – dependent

throughout the comparison period. Finally, there are sensory encodings during the base stimulus period in each of S1, S2, PFC, VPC, and MPC.

B.6 Motor Plans

What is the role of primary motor cortex (M1)? Does it simply receive a signal that constitutes the output of the decision process, or does it participate in that computational decision as well? M1 is essentially silent during the base and delay period, and its pattern of activity during the comparison period is very similar to that of S2, MPC, and VPC, in that the firing rate of its neurons is a function of $f_2 - f_1$ (Romo, DeLafuente, and Hernandez 2004). This, along with response latency data (Romo, Hernandez, Zainos, Brody et al. 2002; Romo, Hernandez, and Zainos 2004) showing that the comparison activity in M1 occurs later than that type of activity in S2, MPC, or VPC, suggests that M1 receives the output of the decision procedure, but perhaps does not participate in the computation itself.

However, it should be noted that, as in the other cortical areas, M1 has two populations, each of which are selectively sensitive to $f_2 > f_1$ or $f_1 > f_2$ (Romo, DeLafuente, and Hernandez 2004). Additionally, using a different task in which monkeys were trained to categorize (not discriminate) the speed of a mechanical probe moving across the finger (Salinas and Romo 1998), the differential activity in M1 seems to point towards a different conclusion.

This task is similar to the discrimination task, except that monkeys simply signal whether a given stimulation is higher or lower than 20 Hz. They perform with high levels of accuracy, far greater than chance. During the task, Salinas and Romo recorded from M1, and found that the majority of the neurons (about 85%) responded during motor output, but not differentially between different arm

movements or according to the category chosen. However, the remaining 15% did respond differentially to the different category choices, but not during (i) passive stimulation with no motor output, (ii) passive arm movement, nor (iii) during a visually guided task with the same movements, where a visual cue directed the monkey towards which button to press, even though the same vibrating stimulus was presented.

Interestingly, the firing rates had a sigmoidal shape: for a neuron that "preferred" higher speeds, its firing rate was essentially the same for stimulus speeds of 22-30 Hz. For a neuron that "preferred" lower speeds, its rate was essentially the same for stimulus speeds of 12-20 Hz (see Salinas and Romo 1998, figures 3 and 4). Thus, as found earlier, there are two subpopulations, each of which is selective for either high or low speeds. Finally, through analyzing error patterns, and the differences between error and hit trials, we get the following hypotheses. An ideal, purely motor neuron that is selective for lateral movement would fire at a high rate for lateral and low rate for medial movements, and vice versa for a motor neuron selective for medial movements. A pure sensory neuron would fire selectively for high or low speeds, regardless of motor output. Analysis of the differences between hit and error trials however did not bear out either of these hypotheses. Instead, the rates correlated with a combination of movement direction and speed category. "[T]he differential neurons are selective for the speed categories, but because their activity has an impact on the motor output, their firing rates correlate with both category and movement" (Salinas and Romo 1998, 509).

Given the results from the above study, it appears that there is a subpopulation of neurons in M1 that are neither purely sensory nor purely motor, and that appear to participate in the neural computation/decision process itself, which may drive the larger population. That subpopulation is again broken down into further subpopulations, each selective for a different category. Interestingly, the sigmoidal shape of the firing rate as a function of tactile speed suggests that these neurons encode, or at

least correlate with, arbitrary, learned categories (“high” or “low”). Whether that analysis should be applied to the tactile discrimination task is uncertain, since the flutter task is a discrimination task, involving working memory and the comparison of a sensory representation with a short-term memory, whereas the categorization task involves comparison with a learned category stored in long-term memory. However, M1 does appear to play a role in the decision procedure for at least the categorization task, and it does have differential activity selective for the different decisions the animal may make (i.e., base greater than comparison or vice versa). Whether that differential activity participates in the comparison and decision procedure, or simply receives a copy of a decision already made, is unclear.

Appendix C: Specifying Relational Systems by Neurometric Discrimination Thresholds

C.0 An Alternate Method of Typing Biological Relational Systems

According to SPT, relational systems must be independently specified based on considerations from the relevant biological and physical sciences. In chapters 7 and 8 I accepted idealizing assumptions about firing rate and frequency, which allowed me to make use of isomorphisms between totally ordered relational systems whose domains were order dense and with the power of the continuum. However, I do not claim that that will be the best manner of specifying relational systems in every case. SPT is a pliable theory, generalizable to other cases with different kinds of relational systems. In this final appendix I consider a manner of specifying relational systems based on the discovered neurometric discrimination thresholds, and show how SPT can handle this.

As neural activity becomes less phase-locked to the stimulus (and hence less periodic), the discriminatory capacity of that activity decreases. As discussed in B.3.1, neurometric thresholds for S1 neurons, calculated using periodicity, are lower (i.e., more discriminating) than neurometric thresholds calculated using firing rate. Further, the firing rate-generated neurometric thresholds closely match psychophysical thresholds, whereas periodicity-generated neurometric thresholds do not. Using firing rate, the ideal observer would behave much like the animal does, whereas using periodicity, the ideal observer would be more discriminating. Essentially, the animal, and firing rate, cannot discriminate frequency differences of less than 3 Hz. Let us consider an alternate method of defining relational systems based on these experimental findings. The alternate suggestion is going to be based on range equivalence classes (see 5.4.6), defined in terms of neurometric thresholds.

Let's begin by defining range equivalence classes and their associated relational systems. Let A be the set of frequencies, with the power of the continuum, and A^{RE} be a countable partition of A. Then let $\mathfrak{U}^{RE} = \langle A^{RE}, >_{AR} \rangle$, and define the relation $>_{AR}$ in terms of $>_A$, as $a >_A b$ iff $a^{re} >_{AR} b^{re}$, where a^{re} and b^{re} are the range equivalence classes associated with a and b (elements of A), respectively. The average neurometric discrimination threshold based on firing rate for the neurons in S1 studied by (Hernandez, Zainos, and Romo 2000) is 3.37 $\pm$ 1.82 Hz, whereas the average neurometric threshold for neurons that vary in their periodic temporal structure is 0.79 $\pm$ 1.22 Hz[153]. It should be kept in mind that neurometric and psychometric thresholds are calculated as the difference in frequency that the neuron/monkey discriminates 75% of the time. For simplicity, let's only deal with the rate-based neurometric threshold, and set it at 3 Hz.

Given this 3 Hz discrimination threshold, we can define the members of A^{RE} as follows. So long as we continue to assume that measurement of frequency along a continuum is justified, then we get the result that $\mathfrak{U}$ is isomorphic to $\mathfrak{N}^+$, hence, the members of A can be mapped one-one with the nonnegative real numbers, with a bijective function, c. But since c is bijective, it has an inverse, so we can use c' to map continuous ranges from the number line back to continuous ranges of elements of A, in 3 Hz increments, starting at 5 Hz, since we are effectively operating in the 5-50 Hz range. We then have A^{RE}, which is finite, and we can define $>_{AR}$ and $\mathfrak{U}^{RE}$ as described above.

Next we define the physiological relational system $\mathfrak{B}$ or an appropriate range equivalence-based analogue of it. First, map A to B according to r_3: [154]

$$r_3(s) = 22 + 0.7s$$

[153] For comparison, the average psychometric threshold reported is 3.07$\pm$0.34 Hz. So the animal cannot reliably discriminate at levels less than about 3 Hz, nor can firing rate. However, periodicity thresholds were reported as low as 0.20 Hz.

[154] Recall from chapter 7 that r_3 describes the discovered correlation between firing rate in subpopulation-2 of S1 and tactile frequency.

Then, given that the ordering induced by $>_A$ remains within each range equivalence class, we can take the median element of each class, and call it a_m^{re}. Take $r_3(a_m^{re})$, for every $a^{re} \in A^{RE}$. Those elements of B comprise a new class, which we can call B^*. B^* and A^{RE} are equinumerous (since $r_3: A \rightarrow B$ generates an isomorphism, and hence is bijective), and the old ordering, $>_B$, can be used to order B^*. Now, define $r_8: A^{RE} \rightarrow B^*$, as follows. If $r_3(a_m^{re}) = b \in B^*$, then $r_8(a^{re}) = b$. This new function r_8 defines an isomorphism from $\mathfrak{A}^{RE}$ to $\mathfrak{B}^* = \langle B^*, >_B \rangle$. We need to connect frequencies (the members of A), not equivalence classes, to B^*. However, since $>_{AR}$ induces a total order, we won't be able to get a homomorphism from $\mathfrak{A}$ to $\mathfrak{B}^*$. However, we can construct a Δ/Ψ –morphism to handle this.

Define $\mathfrak{A}^{\Delta/\Psi} = \langle A, =, >_A \rangle$, and let $\mathfrak{B}^{*\Delta/\Psi} = \langle B^*, =, >_B \rangle$. Then we use r_8 to define a new function $r_9: A \rightarrow B^*$, as follows. If $a \in a^{re}$ then $r_9(a) = r_8(a^{re})$. r_9 defines a Δ/Ψ–morphism between $\mathfrak{A}^{\Delta/\Psi}$ and $\mathfrak{B}^{*\Delta/\Psi}$. Let the identity relation be an element of Δ, and let $>_A \in \Psi$, and further, there are no other elements of Δ or Ψ. Note that $r_9(a) >_B r_9(b)$ implies that $a >_A b$, but the implication does not go the other way. Hence, r_9 counter-preserves but does not preserve $>_A$. On other hand, $a = b$ implies that $r_9(a) = r_9(b)$, hence, r_9 preserves identity. Since these are the only respective elements of Δ and Ψ, r_9 preserves all of the relations in Δ and counter-preserves all of the relations in Ψ, and hence defines a Δ/Ψ -morphism from $\mathfrak{A}^{\Delta/\Psi}$ to $\mathfrak{B}^{*\Delta/\Psi}$, and this gives us structural preservation between $\mathfrak{A}^{\Delta/\Psi}$ and $\mathfrak{B}^{*\Delta/\Psi}$.

The elements of $\mathfrak{B}^*$ each map to a 3 Hz range of frequencies. For example, consider the range equivalence class a^{re} defined as [11,14) Hz[155], whose median (a_m^{re}) is 12.5 Hz. According to r_3, 12.5 Hz maps to 30.75 spikes/sec, thus, according to r_8, a^{re} maps to 30.75 spikes/sec. Assuming a reasonable

[155] I've decided to make the lower end of the range inclusive and the upper end a limit, so that we don't get any elements of A in more than one element ofA^{RE}. The need for A^{RE} to partition A is not arbitrary, but which bracket is open and which is closed is. But nothing turns on it.

argument could be made for the teleofunction hypothesis which would claim that neurons in subpopulation-2 have the teleofunction of covarying, according to r_8, with energy states at the periphery, we would then assert the following. A firing rate of 30.75 spikes/sec by neurons in subpopulation-2 of S1 f-predicates the property of vibrating at 12.5 $\pm$ 1.5 Hz.

Without the jargon, I suggest that we can define ranges of energy states, in accordance with the experimental finding that the discrimination threshold for firing rate is about 3 Hz. Then, take the median element of that range, and map it to a firing rate according to the experimentally determined r_3 (say, *x* spikes/sec). Then we can say that *x* spikes/sec f-predicates a 3 Hz range of frequencies, in accordance with the discovered discrimination threshold.

One obvious fault with this proposal is that only the firing rates in B^* are assigned a content, and that would, at least prima facie, seem to be an unacceptably arbitrary result. We can however accommodate the other elements of B, as follows. Define range equivalence classes again, but this time for B, using r_3. Since $r_3(5) = 25.5$ spikes/sec, and $r_3(8) = 27.6$ spikes/sec, we'll let b_1^{re} be composed of every member of B, between 25.5 and 27.6 spikes/sec, ordered by $>_B$. To do this, we again need to make the assumption that a continuous measurement of B is justified, providing a bijection to $\mathbb{R}^+$, and its inverse providing the continuous ranges, as we did above for A^{RE}. The remaining elements of B^{RE} are defined in turn, where the members of B that correspond (under r_3) to the endpoints of the 3 Hz increments defining A^{RE}, define the endpoints of each $b^{re} \in B^{RE}$. Notice that the average and median element of each b^{re} is a member of B^*. Also notice that B^{RE} is equinumerous with A^{RE}. Finally, define $\mathfrak{B}^{RE} = \langle B^{RE}, >_{BR} \rangle$, with $>_{BR}$ defined exactly parallel to the definition of $>_{AR}$.

We may now define r_{10} by mapping the first member (in the $>_{AR}$ ordering) of A^{RE} to the first member (ordered by $>_{BR}$) of B^{RE}, and so on up. The function r_{10} defines an isomorphism from $\mathfrak{A}^{RE}$ to $\mathfrak{B}^{RE}$. We may define a second function, $r_{11}: A \rightarrow B^{RE}$, as follows. If $a \in a^{re}$ then $r_{11}(a) = r_{10}(a^{re})$.

The function r_{11} defines a Δ/Ψ -morphism from $\mathfrak{A}^{\Delta/\Psi}$ to $B^{RE^{\Delta/\Psi}}$, with $B^{RE^{\Delta/\Psi}} = \langle B^{RE}, =, >_{BR} \rangle$, and this gives us structural preservation between $\mathfrak{A}^{\Delta/\Psi}$ and $B^{RE^{\Delta/\Psi}}$.

To define our representation function, however, we'll have to do something different, because r_{11} maps frequencies to range equivalence classes, not frequencies to rates. Try this: for any element $b \in B$, if $b \in b_1^{re}$, then b f-predicates the property predicated by the median element of b_1^{re}. Or in other words, every element of b_1^{re} gets mapped to by whatever maps to the median of b_1^{re}. The problem is that we no longer have a function: each element of, say, a_1^{re}, maps to the infinite number of elements of b_1^{re}, and this cannot be rectified by redefining the function from B^{RE} to A^{RE} because the same issue will arise.

For the purposes of defining a correspondence relation, which determines f-predicative content, the fact that this is not a function is unproblematic. The functions r_8 through r_{11} are functions, and can be used for whatever mathematical work needs to be done. The only problem is that, on the above suggestion, we will not only have multiple contents for each firing rate, but we will also have multiple firing rates that have the same contents. However, I've argued in 5.4.8 that neither one of these scenarios should be considered a problem.

The point of the above exercise is to demonstrate that, while we started with the idealization assumptions (to define continuous ranges), ultimately we were able to abandon them and only deal with a finite number of elements in our representation functions. Further, we were able to account for the experimental determination of discrimination thresholds.

Structural preservation theory does not decide whether we should use range equivalence classes or the earlier method to define the relevant relational systems here, because these are independent questions about how to type brain states and energy states (see 5.4.7.2 for discussion). However, even though it doesn't provide answers to such questions, SPT provides a structured

theoretical framework within which fruitful questions about representation can be asked. It provides us with a conceptual framework for thinking about representation and for interpreting experimental results. I take it to be an advantage of the theory that it is able to account for several different reasonably plausible methods of defining brain states and the energy states that we take them to represent. Should it turn out that further conceptual and experimental work makes it more pressing that we deal with range equivalence classes, structural preservation theory has the resources to do so. However, at this point I would not advocate using the range equivalence method discussed above, at least for the cases here under discussion.

First, discrimination thresholds, as discussed in the experimental paradigm under consideration, are calculated based on whether the neuron/animal correctly discriminates 75% of the time. But why 75%? Why not 95%, or something else? By changing the chosen cutoff point, we change the discrimination threshold, thus making the threshold itself relative to a seemingly arbitrary decision. Second, consider again the experimentally determined function defining the relationship between rate and frequency:

$$r(s) = 22 + 0.7s + \sigma\epsilon.$$

The final term on the right hand side accounts for noise, which occurs in every neural system. Ultimately, ionic movement across membranes is a stochastic process, many of the mechanisms that open and close ion channels are stochastic processes, and vesicle release into the synaptic cleft is a stochastic process. As a result, there will always be some amount of "random" electrical activity, which we call noise. Initially, I had suggested that we ignore the noise term, since noise is by definition not a signal and hence not relevant to understanding representational content. However, if we consider how the discrimination thresholds are calculated, we can see that the thresholds are probably a result of

noise, rather than a result of indeterminate but bounded content, as the range equivalence method would imply.

To get the firing rate-based neurometric threshold, for example, the rule used by the ideal observer is, if there are more spikes during the comparison stimulus than during the base stimulus, then the comparison is higher. By adding in random spikes not caused by the stimulus which do not covary with frequency according to the representation function, the ideal observer's rule will be progressively les reliable as the difference between frequencies gets smaller, thus making each individual spike more important to the rule's outcome. Thus, the discrimination threshold is not a result of bounded, "range-equivalent" content, but because of a noisy signal.

Third, part of the motivation for using range equivalence classes here was to see if we can excise some of the idealizing assumptions. Even though we ended up with finite relational systems, we never really abandoned the idealization assumptions, since we need to make those assumptions in order to define the range equivalence classes in the first place. In all then, at least in this case, we should not type brain states and energy states by range equivalence classes. However, structural preservation theory does have the resources to account for this should it become necessary in other circumstances.

Bibliography

Allen, C., and M. Bekoff. 1997. *Species of mind: The philosophy and biology of cognitive ethology*. Cambridge: MIT Press.

Armstrong, D.M. 1968. *A materialist theory of the mind*. London and New York: Routledge and K. Paul, Humanities Press.

———. 1981. *The nature of mind, and other essays*. Ithaca, NY: Cornell University Press.

Austin, J.L. 1975. *How to do things with words*. 2nd ed. Oxford: Clarendon Press.

Baker, L.R. 1987. *Saving belief: A critique of physicalism*. Princeton, NJ: Princeton University Press.

Bechtel, W. 2001. Representation: From neural systems to cognitive systems. In *Philosophy and the neurosciences: A reader*, edited by W. Bechtel, P. Mandik, J. Mundale and R. S. Stufflebeam. Malden, Mass.: Blackwell Publishers.

Bekoff, M., and D. Jamieson, eds. 1996. *Readings in animal cognition*. Cambridge: MIT Press.

Bickle, J. 2003. *Philosophy and neuroscience: A ruthlessly reductive account*. Dordrecht and Boston: Kluwer Academic Publishers.

Bodner, M., J. Kroger, and J.M. Fuster. 1996. Auditory memory cells in dorsolateral prefrontal cortex. *Neuroreport* 96:1905-1908.

Bond, R.J., and W.J. Keane. 1999. *An introduction to abstract mathematics*. Pacific Grove, CA: Brooks/Cole.

Brentano, F. 1874. *Psychologie vom Empirischen Standpunkt*: Leipzig.

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

Cantwell Smith, B. 2002. Cummins - or something isomorphic to him. In *Philosophy of mental representation*, edited by H. Clapin: Clarendon Press/Oxford.

Churchland, P.M. 1979. *Scientific realism and the plasticity of mind*, *Cambridge studies in philosophy*. Cambridge ; New York: Cambridge University Press.

———. 1981. Eliminative materialism and the propositional attitudes. *The Journal of Philosophy* 78 (2):67-90.

———. 1986. Some reductive strategies in cognitive neurobiology. *Mind* 95 (379):279-309.

———. 1988. *Matter and consciousness : a contemporary introduction to the philosophy of mind*. Rev. ed. Cambridge, Mass.: MIT Press.

———. 1989. *A neurocomputational perspective: The nature of mind and the structure of science*. Cambridge, MA: MIT Press.

Churchland, P.S. 1986. *Neurophilosophy: Toward a unified science of the mind-brain*. Cambridge, Mass.: MIT Press.

Cover, T. M., and J.A. Thomas. 1991. *Elements of information theory*, *Wiley series in telecommunications*. New York: Wiley.

———. 2006. *Elements of information theory*. 2nd ed. Hoboken, N.J.: Wiley-Interscience.

Craver, C.F. 2007. *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford and New York: Oxford University Press.

Cummins, D., and C. Allen, eds. 1998. *The evolution of mind*. Oxford: Oxford University Press.

Cummins, R., and G. Schwarz. 1988. Radical Connectionism. In *Spindel Conference 1987: Connectionism and the Philosophy of Mind*, edited by T. H. J. Tienson: Supplement, Southern Journal of Philosophy.

Cummins, R. 1989. *Meaning and mental representation*. Cambridge, Mass.: MIT Press.

———. 1996. *Representations, targets, and attitudes*. Cambridge, Mass.: MIT Press.

Dawkins, R. 1976. *The selfish gene*. Oxford: Oxford University Press.

Dennett, D.C. 1969. *Content and consciousness*. New York: Humanities Press.

———. 1982. Styles of mental representation. *Proceedings of the Aristotelian Society* 83:213-226.

———. 1983. Intentional systems in cognitive ethology: The "Panglossian Paradigm" defended. *The Behavioral and Brain Sciences* 6:343-390.

———. 1987. Evolution, error, and intentionality. In *The Intentional Stance*, edited by D. C. Dennett. Cambridge: A Bradford Book. MIT Press.

———. 1987. Reflections: Real patterns, deeper facts, and empty questions. In *The Intentional Stance*. Cambridge: A Bradford Book. The MIT Press.

———. 1987. *The intentional stance*. Cambridge, Mass.: MIT Press.

Devitt, M. 1994. The methodology of naturalistic semantics. *The Journal of Philosophy* 91 (10):545-572.

———. 1996. *Coming to our senses: A naturalistic program for semantic localism*. Cambridge and New York: Cambridge University Press.

———. 1997. *Realism and truth*. 2nd ed. Princeton, N.J.: Princeton University Press.

Devitt, M., and K. Sterelny. 1999. *Language and reality: An introduction to the philosophy of language*. 2nd ed. Cambridge, Mass.: MIT Press.

Dretske, F.I. 1981. *Knowledge and the flow of information*. 1st MIT Press ed. Cambridge, Mass.: MIT Press.

———. 1983. Precis of *Knowledge and the Flow of Information*. *Behavioral and Brain Sciences* 6:55-63.

———. 1986. Misrepresentation. In *Belief: Form, content, and function*, edited by R. Bogdan. New York and Oxford: Oxford University Press.

———. 1988. *Explaining behavior: Reasons in a world of causes*. Cambridge, Mass.: MIT Press.

———. 1995. *Naturalizing the mind*. Cambridge, Mass.: MIT Press.

———. 2002. A recipe for thought. In *Philosophy of mind: Classical and contemporary readings*, edited by D. J. Chalmers. New York and Oxford: Oxford University Press.

Eliasmith, C. 2000. How neurons mean, Philosophy-Neuroscience-Psychology Program, Washington University in St. Louis, St. Louis.

Eliasmith, C., and C. H. Anderson. 2003. *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. Cambridge, Mass.: MIT Press.

Feigl, H. 1970. The 'orthodox' view of theories: Remarks in defense as well as critique. In *Minnesota studies in the philosophy of science, vol. IV*, edited by M. Radner and S. Winokur. Minneapolis: University of Minnesota Press.

Fletcher, L. 2008. Information, relevance, and objectivity. *Indiana Undergraduate Journal of Cognitive Science* 3:3-20.

Fodor, J.A. 1975. *The language of thought*. New York: Crowell.

———. 1984. Semantics, Wisconsin style. *Synthese* 59:231-250.

———. 1987. Making mind matter more. *The Journal of Philosophy* 84 (11):642.

———. 1987. *Psychosemantics : The problem of meaning in the philosophy of mind*. Cambridge, Mass.: MIT Press.

———. 1990. *A theory of content and other essays*. Cambridge, Mass.: MIT Press.

———. 1990. A theory of content II, the theory. In *A Theory of Content and Other Essays*, edited by J. A. Fodor. Cambridge and London: A Bradford Book. MIT Press.

———. 1994. *The elm and the expert: Mentalese and its semantics*. Cambridge: MIT Press.

———. 1998. *Concepts: Where cognitive science went wrong*. Oxford and New York: Clarendon Press and Oxford University Press.

———. 1999. Diary. *London Review of Books* 21:19.

Frege, G. 1952. *Translations from the philosophical writings of Gottlob Frege*. New York: Philosophical Library.

Funahashi, S., C.J. Bruce, and P.S. Goldman-Rakic. 1989. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology* 61:331-349.

Gardner, E.P., and E.R. Kandel. 2000. Touch. In *Principles of Neural Science*, edited by E. R. Kandel, J. H. Schwartz and T. M. Jessell: McGraw-Hill.

Gardner, E.P., J.H. Martin, and T.M. Jessell. 2000. The bodily senses. In *Principles of Neural Science*, edited by E. R. Kandel, J. H. Schwartz and T. M. Jessell: McGraw-Hill.

Garfield, J.L. 1988. *Belief in psychology: A study in the ontology of mind*. Cambridge, Mass.: MIT Press.

Gerstner, W., and W.M. Kistler. 2002. *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge, U.K. ; New York: Cambridge University Press.

Godfrey-Smith, P. 1994. A continuum of semantic optimism. In *Mental representation: A reader*, edited by S. P. Stich and T. A. Warfield: Blackwell.

Gold, J.I., and M.N. Shadlen. 2001. Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Science* 5:10-16.

Goodman, N. 1968. *Languages of art: An approach to a theory of symbols*. Indianapolis,: Bobbs-Merrill.

———. 1976. *Languages of Art*. 2nd ed. Indianapolis/Cambridge: Hackett.

Grice, H.P. 1957. Meaning. *The Philosophical Review* 66 (3):377-388.

Haugeland, J. 1985. *Artificial intelligence: The very idea*. Cambridge, Mass.: MIT Press.

Hernandez, A., E. Salinas, R. Garcia, and R. Romo. 1997. Discrimination in the sense of flutter: New psychophysical measurements in monkeys. *The Journal of Neuroscience* 17 (16):6391-6400.

Hernandez, A., A. Zainos, and R. Romo. 2000. Neuronal correlates of sensory discrimination in the somatosensory cortex. *Proceedings of the National Academy of Sciences USA*:6191-6196.

———. 2002. Temporal evolution of a decision-making process in the medial premotor cortex. *Neuron* 33 (6):959-972.

Horgan, T. 1992. From cognitive science to folk psychology: Computation, mental representation, and belief. *Philosophy and Phenomenological Research* 52 (2):449-484.

———. 1994. Computation and mental representation. In *Mental representation: A reader*, edited by S. P. Stich and T. A. Warfield. Oxford: Blackwell.

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

———. 1968. Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology (London)* 195:215-243.

Kitcher, P. 1981. Explanatory unification. *Philosophy of Science* 48:507-531.

Koch, C. 1999. *Biophysics of computation: Information processing in single neurons*. New York: Oxford University Press.

Krantz, D.H., R.D. Luce, P. Suppes, and A. Tversky. 1971. *Foundations of measurement*. 3 vols. Vol. 1. New York: Academic Press.

Kripke, S.A. 1980. *Naming and necessity*. Cambridge, Mass.: Harvard University Press.

LaMotte, R.H., and V.B. Mountcastle. 1975. The capacities of humans and monkeys to discriminate between vibratory stimli of different frequency and amplitude: A correlation between neural events and psychological measurements. *Journal of Neurophysiology* 38:539-559.

———. 1979. Disorders in somesthesis following lesions of parietal lobe. *Journal of Neurophysiology* 42:400-419.

Levin, M. 1997. Plantinga on functions and the theory of evolution. *Australasian Journal of Philosophy* 75 (1):83-98.

Lewis, D.K. 1966. An argument for the identity theory. *The Journal of Philosophy* 63 (1):17-25.

———. 1973. *Counterfactuals*. Cambridge: Harvard University Press.

Loewer, B. 1983. Information and belief. *Behavioral and Brain Sciences* 6 (75-76).

Mandik, P., M. Collins, and A. Vereschagin. 2007. Evolving artificial minds and brains. In *Mental states, volume 1: Nature, function, evolution*, edited by A. C. Schalley and D. Khlentzos. Amsterdam and Philadelphia: John Benjamins Publishing Company.

McGinn, C. 1989. *Mental content*. Oxford and New York: Blackwell.

Miller, E.K., C.A. Erickson, and R. Desimone. 1996. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *The Journal of Neuroscience* 16:5154-5167.

Miller, I. 1984. *Husserl, perception, and temporal awareness*. Cambridge, Mass.: MIT Press.

Millikan, R.G. 1984. *Language, thought, and other biological categories: New foundations for realism*. Cambridge, Mass.: MIT Press.

———. 1986. Thoughts without laws; Cognitive science with content. *The Philosophical Review* 95 (1):47-80.

———. 1989. Biosemantics. *The Journal of Philosophy* 86 (6):281-297.

———. 1989. In defense of proper functions. *Philosophy of Science* 56 (2):288-302.

———. 1990. Compare and contrast Dretske, Fodor, and Millikan on teleosemantics. *Philosophical Topics* 18 (2):151-161.

———. 1993. Propensities, exaptations, and the brain. In *White Queen Psychology and Other Essays for Alice*, edited by R. G. Millikan. Cambridge and London: A Bradford Book. MIT Press.

———. 1993. *White Queen psychology and other essays for Alice*. Cambridge, Mass.: MIT Press.

———. 1995. Pushmi-pullyu representations. *Philosophical Perspectives* 9:185-200.

———. 2001. What has natural information to do with intentional representation? In *Naturalism, evolution, and mind*, edited by D. Walsh. Cambridge: Cambridge University Press.

———. 2004. *Varieties of meaning*. Cambridge, Mass.: MIT Press.

Mountcastle, V.B., R.H. LaMotte, and G. Carli. 1972. Detection thresholds for stimuli in humans and monkeys: Comparison with threshold events in mechanoreceptive afferent nerve fibers innervating the monkey hand. *Journal of Neurophysiology* 35:122-136.

Mountcastle, V.B., M.A. Steinmetz, and R. Romo. 1990. Frequency discrimination in the sense of flutter: Psychophysical measurements correlated with postcentral events in behaving monkeys. *The Journal of Neuroscience* 10:3032-3044.

Mountcastle, V.B., W.H. Talbot, I. Darian-Smith, and H.H. Kornhuber. 1967. Neural basis of the sense of flutter-vibration. *Science* 155:597-600.

Mountcastle, V.B., W.H. Talbot, H. Sakata, and J. Hyvarinen. 1969. Cortical neuronal mechanisms in flutter-vibration studied in unanesthetized monkeys: Neuronal periodicity and frequency discrimination. *Journal of Neurophysiology* 32:452-484.

O'Brien, G., and J. Opie. 2004. Notes toward a structuralist theory of mental representation. In *Representation in mind: New approaches to mental representation*, edited by H. Clapin, P. Staines and P. Slezak: Elsevier.

Pierce, C.S. 1931-1958. *The collected papers of C.S. Pierce, vols. 1-8*. Edited by A. Burks, C. Hartshorne and P. Weiss. Cambridge: Harvard University Press.

Prinz, J.J. 2002. *Furnishing the mind: Concepts and their perceptual basis*. Cambridge, Mass.: MIT Press.

Putnam, H. 1960. Minds and machines. In *Dimensions of mind*, edited by S. Hook. New York: New York University Press.

———. 1975. The meaning of 'meaning'. In *Language, mind, and knowledge*, edited by K. Gunderson. Minneapolis: University of Minnesota Press.

———. 1975. *Mind, language, and reality*. Vol. 2. Cambridge and New York: Cambridge University Press.

———. 1975. What is mathematical truth? In *Mathematics, matter, and method: Philosophical papers*, edited by H. Putnam. Cambridge: Cambridge University Press.

Pylyshyn, Z.W. 1984. *Computation and cognition: Toward a foundation for cognitive science*. Cambridge, Mass.: MIT Press.

Quine, W. V. 1948. On what there is. *The Review of Metaphysics* 2 (21).

———. 1951. Two dogmas of empiricism. *The Philosophical Review* 60:20-43.

Ramsey, W. 2007. *Representation reconsidered*. Cambridge and New York: Cambridge University Press.

Rao, S.C., G. Rainer, and E.K. Miller. 1997. Integration of what and where in the primate prefrontal cortex. *Science*:821-824.

Recanzone, G.H., M.M. Merzenich, and C.E. Schreiner. 1992. Changes in the distributed temporal response properties of S1 cortical neurons reflect improvements in performance on a temporally based tactile discrimination task. *Journal of Neurophysiology* 67:1071-1091.

Rieke, F. 1997. *Spikes: Exploring the neural code*. Cambridge, Mass.: MIT Press.

Romo, R., C.D. Brody, A. Hernandez, and L. Lemus. 1999. Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature* 399:470-473.

Romo, R., V. DeLafuente, and A. Hernandez. 2004. Somatosensory discrimination: Neural coding and decision-making mechanisms. In *The Cognitive Neurosciences*, edited by M. Gazzaniga. Cambridge: A Bradford Book. MIT Press.

Romo, R., A. Hernandez, and A. Zainos. 2004. Neuronal correlates of a perceptual decision in ventral premotor cortex. *Neuron* 41 (1):165-173.

Romo, R., A. Hernandez, A. Zainos, C.D. Brody, and L. Lemus. 2000. Sensing without touching: Psychophysical performance based on cortical microstimulation. *Neuron* 26:273-278.

Romo, R., A. Hernandez, A. Zainos, C.D. Brody, and E. Salinas. 2002. Exploring the cortical evidence of a sensory-discrimination process. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 357 (1424):1039-1051.

Romo, R., A. Hernandez, A. Zainos, L. Lemus, and C.D. Brody. 2002. Neuronal correlates of decision-making in secondary somatosensory cortex. *Nature Neuroscience* 5 (11):1217-1225.

Romo, R., A. Hernandez, A. Zainos, and E. Salinas. 1998. Somatosensory discrimination based on cortical microstimulation. *Nature* 392 (6674):387-390.

———. 2003. Correlated neuronal discharges that increase coding efficiency during perceptual discrimination. *Neuron* 38 (4):649-657.

Romo, R., and E. Salinas. 2001. Touch and go: Decision-making mechanisms in somatosensation. *Annual Review of Neuroscience* 24:107-137.

———. 2003. Flutter discrimination: Neural codes, perception, memory and decision making. *Nature Reviews. Neuroscience* 4 (3):203-218.

Rosch, E. 1973. On the internal structure of perceptual and semantic categories. In *Cognitive development and the acquisition of language*, edited by T. E. Moore. New York: Academic Press.

———. 1975. Cognitive representation of semantic categories. *Journal of Experimental Psychology* 104:192-233.

———. 1978. Principles of categorization. In *Cognition and categorization*, edited by E. Rosch and B. Lloyd. Hillsdale, N.J.: Erlbaum.

Salinas, E., A. Hernandez, A. Zainos, L. Lemus, and R. Romo. 1998. Cortical recording of sensory stimuli during somatosensory discrimination. *Society for Neuroscience Abstracts* 24:1126.

Salinas, E., A. Hernandez, A. Zainos, and R. Romo. 2000. Periodicity and firing rate as candidate neural codes for the frequency of vibrotactile stimuli. *The Journal of Neuroscience* 20 (14):5503-5515.

Salinas, E., and R. Romo. 1998. Conversion of sensory signals into motor commands in primary motor cortex. *The Journal of Neuroscience* 18 (1):499-511.

Searle, J.R. 1983. *Intentionality: An essay in the philosophy of mind*. Cambridge and New York: Cambridge University Press.

———. 1992. *The rediscovery of the mind*. Cambridge, Mass.: MIT Press.

Shannon, C.E., and W. Weaver. 1949. *The mathematical theory of communication*. Urbana: University of Illinois Press.

Shepard, R., and S. Chipman. 1970. Second order isomorphism of internal representations: Shapes of states. *Cognitive Psychology* 1:1-17.

Stich, S.P., and T.A. Warfield. 1994. *Mental representation: A reader*. Oxford and Cambridge: Blackwell.

Suppes, P., and J.L. Zinnes. 1963. Basic measurement theory. In *Handbook of Mathematical Psychology*, edited by R. D. Luce, R. R. Bush and E. Galanter. New York: John Wiley and Sons, Inc.

Swoyer, C. 1991. Structural representation and surrogative reasoning. *Synthese*:449-508.

Talbot, W.H., I. Darian-Smith, H.H. Kornhuber, and V.B. Mountcastle. 1968. The sense of flutter-vibration: Comparison of the human capacity with response patterns of mechanoreceptive afferents from the monkey hand. *Journal of Neurophysiology* 31:301-334.

Tarski, A. 1954. Contributions to the theory of models, I, II. *Indagationes Mathematicae* 16:572-588.

Turing, A.M. 1950. Computing machinery and intelligence. *Mind* 59 (236):433-460.

Usher, M. 2001. A statistical referential theory of content: Using information theory to account for misrepresentation. *Mind & Langauge* 16 (3):311-334.

Vallbo, A.B. 1995. Single-afferent neurons and somatic sensation in humans. In *The cognitive neurosciences*, edited by M. Gazzaniga. Cambridge: A Bradford Book. MIT Press.

Van Fraassen, B.C. 1980. *The scientific image*. Oxford: Clarendon Press.

———. 1985. Empiricism in the philosophy of science. In *Images of Science: Essays on Realism and Empiricism, with a Reply from Bas C. Van Fraassen*, edited by P. M. Churchland and C. A. Hooker. Chicago: University of Chicago Press.

Wilson, F.A.W., S.P. O'Scalaidhe, and P.S. Goldman-Rakic. 1993. Dissociation of object and spatial processing domains in primate prefrontal cortex. *Science* 260:1955-1958.

Wittgenstein, L. 1953. *Philosophical investigations*. Oxford: Blackwell.

Zainos, A., H. Merchant, A. Hernandez, E. Salinas, and R. Romo. 1997. Role of primary somatic sensory cortex in the categorization of tactile stimuli: Effects of lesions. *Experimental Brain Research* 115:357-360.

Zohary, E., M.N. Shadlen, and W.T. Newsome. 1994. Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature* 370:140-143.