INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

Mental Content in a Physical World: An Alternative to Mentalese

Christopher David Viger
Department of Philosophy
McGill University, Montreal
October 1998

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfilment of the requirements of the degree of
Doctor of Philosophy.

ABSTRACT

In an attempt to show how rational explanation of human and animal behaviour has a place in the scientific explanation of our physical world, Fodor advances the language of thought hypothesis. The purpose of this dissertation is to argue that, contrary to the language of thought hypothesis, we need not possess a linguistic internal representational system distinct from any natural language to serve as the medium of thinking. I accept that we have an internal representational system, but by analyzing Fodor's theory of content, I show Fodor's argument that the internal system must be as expressive as any natural language, which he uses in arguing that the internal system is the linguistic medium of thought, is unsound. Distinguishing an informational theory of content from a causal theory of content, which Fodor conflates, I argue that internal representations, whose content is determined by the information they carry, cannot be related in a way that corresponds to semantic associations between terms in natural languages, reflecting actual associations of items in the world. Furthermore, provided certain animal cognition, which is homogeneous with human cognition, can be explained without requiring that the internal system possess anything corresponding to the logical connectives, the internal system *need not* possess anything corresponding to the connectives. I give such an explanation of animal cognition by developing an approach to content in the Rylean/Dennettian tradition, based on the notion of embodied cognition, in which animals embody the hypotheses they entertain in virtue of their total dispositional state, rather than explicitly representing them. It follows that there are two features of natural languages, semantic associations of terms and possessing logical connectives, that the internal system need not have. Hence a rational interpretation of linguistic behaviour need not be derived from an intentional interpretation of the transformations on internal representations, from which it follows that the internal system need not be the medium of thought, which can be a natural language. Also, if the internal system need not possess anything corresponding to the logical connectives, it need not be as expressive as natural languages, and hence it need not be linguistic.

RÉSUMÉ

En essayant de démontrer comment les explications rationnelles des comportements humain et animal ont leur place dans l'explication scientifique de notre monde physique, Fodor émet l'hypothèse du langage de la pensée. Cette thèse a pour but de débattre que, contrairement à l'hypothèse du langage de la pensée, nous n'avons pas besoin de posséder un système de représentation linguistique interne distinct de tout langage naturel, qui est le véhicule de la pensée. J'accepte que nous ayons un système de représentation interne, mais en analysant la théorie du contenu de Fodor, je démontre que l'argument de Fodor que le système interne doit être aussi expressif que tout langage naturel--argument qu'il utilise pour promouvoir l'idée que le système interne est le véhicule linguistique de la pensée--est mal fondé. En distinguant une théorie ionformationelle du contenu d'une théorie causale du contenu--contrairement à Fodor, qui les combine--j'explique qu'on ne peut pas établir un rapport entre les représentation internes, dont le contenu est déterminé par l'information qu'elles communiquent, d'une maniere qui corresponde aux associations sémantique qui existent entre les termes des langages naturels, reflétant les associations actuelles d'items dans le monde. De plus, étant donné qu'on peut expliquer une certaine cognition animale (homogène à la cognition humaine) sans exiger que le système interne possède quoi que ce soit qui corresponde aux conjonctions logiques, le système interne n'a pas besoin de possèder quoi que ce soit qui corresponde aux conjonctions logiques. Je présente une telle explication de la cognition animale en développant une façon d'aborder le contenu selon la tradition de Ryle et Dennett, basée sur la notion de la cognition incarnée selon laquelle les animaux incarnent les hypothèses qu'ils nourrissent en raison de leur état naturel total, au lieu de les représenter explicitement. Il s'ensuit qu'il y a deux caracteristiques des langages naturels, nommément, c'est-à-dire les associations sémantiques des termes et la possession de conjonctions logiques, dont le système interne n'a pas besoin. Par conséquent une interprétation rationelle du comportement linguistique n'a pas besoin de provenir d'une interprétation intentionelle des transformations des représentations internes, d'où il s'ensuit que le système interne n'a pas besoin d'être le véhicule de la pensée, lequel peut être un langage naturel. Aussi, si le système interne n'a pas besoin de posséder quoi que ce soit, aucun élément qui corresponde aux conjonctions logiques, il n'a pas besoin d'être aussi expressif que les langages naturels, donc il n'a pas besoin d'être linguistique.

ACKNOWLEDGEMENTS

First, I would like to thank my supervisors Dr. Paul Pietroski and Dr. David Davies for their thoroughness and care in advising me on this project. Their professionalism and help have assisted me tremendously.

Also, thanks to SSHRC, FCAR, and McGill University for financial support.

A special thanks to Dr. Daniel Dennett for accepting me as a visiting fellow at Tufts University at a time when this project was beginning to take shape, for his patience and encouragement in allowing me to finish the dissertation as a research associate at Tufts, and for inspiring and helping shape my philosophical interests.

The assistance Dr. Andrew Brook has given me, as friend and adviser, goes beyond the call of duty. He has been especially supportive when I most needed it.

Dr. Robert Stainton has been very helpful in reading and commenting on earlier drafts of this work and helping me clarify many issues, particularly concerning linguistics.

I would also like to thank Nadine Labrèche M.Sc. for helping me with French translation and for translating the abstract.

My parents Dorothy and Chopin Viger have always been supportive of my efforts, and taken an active interest in my work. That support and encouragement during this project is especially appreciated. They were always there for me.

My parents-in-law Sandra and Dr. Daniel Gorman have helped in many ways to make writing this dissertation possible, yet despite the demands placed on them they have been nothing but encouraging. Their efforts are greatly appreciated.

Finally, without the love and support of my family, I could not have completed this project. My children Dan-Paul and Torin were a constant source of inspiration. And despite the strain and upheaval involved in raising two small children and moving, while I was writing this dissertation, my wife, Judith Gorman M.Sc., has not only held up the fort, she has found the energy to help me persevere. I owe this work to her.

TABLE OF CONTENTS

INTRODUCTION

Fodor (1975) advances the hypothesis that there is a language of thought. The hypothesis is offered as an attempt to explain the place of the mental, rationality in particular, in a physical world of causes and effects. The language of thought is much like a machine language in a computer, on Fodor's view. The actual operations of the language of thought are syntactic, but they have an intentional interpretation according to which they are rational transformations of symbols having intentional content. It is important in assessing Fodor's language of thought hypothesis to keep in mind Fodor's main project of trying to secure the place of rational (intentional, psychological) explanation within scientific discourse, in light of challenges from sceptics and eliminativists to the effect that intentionality has no place in the scientific explanations of our physical world. I am in agreement with Fodor that rational explanation does have a place in scientific discourse, and much of this project is devoted to uncovering what commitment to this position entails.

Fodor's argument for the language of thought hypothesis proceeds in two stages. First, he argues that thinking occurs in an internal representational system, which I refer to as the weak hypothesis. Then, based on empirical evidence, Fodor makes several detailed claims about the nature of that representational system; most importantly, Fodor argues that the internal representational system is a linguistic system, hence a *language* of thought. Other features of the internal representational system, according to Fodor, are that it is private, it is innate, it is very rich--in the sense that it contains a representation of everything for which we have a concept--,

transformations of internal representations have an intentional interpretation as truth-preserving rules, the internal system is distinct from any natural language, and thinking animals possess the same internal representational system as humans. I refer to the claim that the internal representational system has the characteristics Fodor attributes to it as the strong hypothesis.

Fodor offers three arguments that the language of thought must be distinct from natural languages; i.e. that it *must be mentalese*. The aim of this dissertation is to challenge that claim. Specifically, since Fodor argues that there *must* be a linguistic representational system distinct from natural languages to serve as the medium of thought, in challenging the language of thought hypothesis, I simply argue that there *need not* be a mentalese. I accept the weak hypothesis that there is an internal representational system, and Fodor's argument that this system is innate. I also accept that at least some human thought occurs in a linguistic medium. However, I disagree that the internal representational system is the linguistic medium of thought. My position is that the innate internal representational system need not be linguistic, and that for competent (natural) language-users it need not be the medium of thinking, which can be a natural language. I develop the position that the first words we learn correspond to our internal representations, as Fodor argues; however, in continuing to learn a natural language it is possible to add structure to the natural language terms that need not be present in the internal representational system. Two ways that structure can be added to the natural language terms are: (i) learning to use the logical connective terms of a natural language from which terms that are already known can be combined

to form new terms; (ii) associating terms to reflect associations of items in the world to which the terms refer. Once a natural language has been learned, it can supersede the internal representational system as the medium of thought. Furthermore, since structure is imposed on the natural language terms by natural language itself, e.g. by the use of the terms for the logical connectives, the rationality of transformations of natural language terms need not be derived from any intentional interpretation of the operations of the internal representational system. "It must not be forgotten that *the semantical characterization of overt verbal episodes is the primary use of semantical terms, and that overt linguistic events as semantically characterized are the model for the inner episodes introduced by the theory*" (Sellars 1956/1997, p.105, emphasis in original).

In chapter 1, I simply present Fodor's arguments that there is a language of thought having the characteristics he attributes to it. My aim in this chapter is to make clear just what I am challenging in challenging the claim that there must be a mentalese, and what the arguments for the claim are. I begin the challenge in earnest by critiquing Fodor's theory of content for symbols in the language of thought, in chapters 2 and 3. I argue that the key condition for content according to Fodor, asymmetric dependence, is not satisfied by any symbol, so that the theory fails to give content to internal representations corresponding to the content of our intuitive semantic types. I use my critique of Fodor's theory of content to several ends, in chapters 4 and 5. First, on Fodor's theory of content, associations between terms in natural languages reflecting associations that occur between items in the world cannot obtain between internal representations of those items. Also, Fodor's theory of content does not apply to the

logical connectives. It is compatible with Fodor's position on content that the internal representational system does not possess anything corresponding to the logical connectives. Thus, provided it is possible to explain certain cognitive functions of animals that are homogeneous with the cognitive functions of language-using humans, without requiring that the internal representational system possess anything corresponding to the logical connectives, the internal system need not possess anything corresponding to the logical connectives. In chapter 6, I offer such an explanation of animal cognition. I base this explanation on a general approach to content, that I develop in chapter 5 as an alternative to Fodor's theory of content, motivated by the way in which his theory fails. The alternative approach is based on the notion of embodied cognition, presented by Andy Clark. It is a dispositional account, in the tradition of Ryle, Armstrong, Stalnaker, and particularly, Dennett. Having given the required explanation of animal cognition, it follows that the internal representational system need not possess anything corresponding to the logical connectives. Then, since the internal representations cannot be associated in a way corresponding to the way that the items they represent are associated in the world, and since the internal representational system need not possess anything corresponding to the logical connectives, it follows that the intentionality of natural languages, particularly that transformations of natural language terms are truth preserving, need not be derived from the intentionality of the internal representational system. But, then, contrary to Fodor's language of thought hypothesis, the internal representational system need not be the medium of thinking. Also, since the internal system need not possess anything

corresponding to the logical connectives, it need not be as expressive as natural languages, in which case the internal system need not be linguistic. I conclude by suggesting how my alternative approach to content might be used to develop an account of abstraction and how it might be that thought is truth preserving.

All of the arguments in this dissertation are my original work. Where someone has helped me develop my line of reasoning, I have made a note to that effect. Also, where I have used someone else's position in developing my argument, I have made it clear in the text.

CHAPTER 1

THE LANGUAGE OF THOUGHT

1.1 Introduction

In this chapter, I lay out Fodor's position that there is a language of thought, mentalese, an internal linguistic medium of thinking, distinct from natural languages. I begin by presenting Fodor's (weak) hypothesis that thinking occurs in an internal representational system. I then present the characteristics Fodor claims the internal representational system to possess (the strong hypothesis), most notably that it is a linguistic representational system distinct from natural languages. Fodor appeals to many of these characteristics in arguing that the linguistic medium of thought is distinct from natural languages. Next, I consider Fodor's argument for the weak hypothesis that thinking occurs in an internal representational system. I conclude the chapter by giving a detailed presentation of Fodor's three arguments that the language of thought must be distinct from natural languages. I respond to the first two arguments in the course of presenting them and I sketch my response to the third argument, but its full presentation must await chapter 4, since it requires material developed in chapters 2 and 3 concerning Fodor's theory of content for symbols in the language of thought.

1.2 What is the Hypothesis that there is an Internal Representational System? (Weak Hypothesis)

The language of thought hypothesis is a hypothesis about the nature of *rational* thought, not restricted to thinking in humans (Fodor 1975, pp. 29, 198-201). Fodor does not clarify what he means by rational thought and he does not define it: "I don't

propose to quibble about what's to count as thinking" (Fodor 1975, p.56). However, it seems to be contrasted with arational brain processes, such as brain activity that causes mental states without being intentionally characterizable itself, that is, without being something to which we assign content.[1] "The events which fix such [mental] states have no interpretation under that assignment of formulae which works best overall to interpret the etiology of our mentation. ... Some mental states are, as it were, brute incursions from the physiological level" (Fodor 1975, p.200). Thus, Fodor acknowledges that there are mental events whose occurrences do not result from processes with an intentional characterization, but these are not instances of what Fodor takes to be rational thinking[2] so they do not fall within the scope of the claim that there is a language of thought (Fodor 1975, pp.200-201).[3]

Though Fodor does not explain what he means by rational thinking, he does indicate what his hypothesis concerns by considering examples of the kinds of processes he takes to be typical cases of rational thought: considered action, concept learning, and perceptual integration, which is the process of forming a coherent description of the distal environment based on immediate percepts caused by the environment. Fodor's

---

[1] There are cases Fodor would not consider instances of rational thought, even though the relevant brain processes are intentionally characterizable. For example, an actress who worries about forgetting her lines, which makes her nervous, such that her nervousness causes her to forget her lines, does not rationally think her way to forgetting her lines.

[2] That is, the term 'mental' is not coextensive with the term 'rational' for Fodor; rather 'mental' means 'psychological' construed in a broad sense.

[3] For those already familiar with the language of thought hypothesis, rational thinking just is the operating of the language of thought, which accounts for some but not all of our mental states, according to Fodor. Of course, Fodor cannot put it this way in arguing that thinking is the operating of a language of thought.

analysis of these examples for a common element leads him to conclude that rational thinking is a process of inductive generalization that proceeds by forming hypotheses and testing them against available data. This is an empirical claim based on evidence from psychology and the models that psychological theories employ to explain the data from studying considered action, concept learning, and perceptual integration. I now present a brief sketch of Fodor's analyses of these rational thought processes; at this stage I do not engage Fodor's claims, since I am merely clarifying what the language of thought hypothesis is about.

According to Fodor, concept learning is determining that certain environmental conditions obtain, as indicated by some response.[4] The model for this process is that hypotheses about the environmental conditions are formed and tested. Each hypothesis is assigned a probability of obtaining, which is used to fix an order for testing the hypotheses. A hypothesis is tested by producing a response when the conditions specified in the hypothesis obtain. Some kind of positive indicator to the response, such as a food reward for animals or the word "yes" for humans, counts as confirming evidence for the hypothesis. A hypothesis that is confirmed by the available data serves as the grounds for an inductive inference that the concept to be learned has in its extension just those things satisfying the conditions specified by the hypothesis. The argument for this account of concept learning is that it best fits with the data showing that there is a non-arbitrary "relation between what is learned and the experiences that

[4] Fodor makes the case that the particular response is irrelevant and so is not what is being learned. (Fodor 1975, pp.35-6, footnote 6). Nonetheless, some response must be possible in the appropriate conditions if the concept has been learned.

occasion the learning" (Fodor 1975, p.38). Specifically, the experiences that occasion learning some concept are confirming evidence for the hypothesis that grounds the inductive generalization.

Considered action is a process of representing a given situation, the behavioural options an organism has in that situation, and the expected consequence of each behavioural option. This is a process of hypothesis formation and confirmation in which each hypothesis has the form that in the given situation a particular behaviour will have a consequence of a certain value, from which a preference ordering is then assigned to each consequence. The data used to formulate the hypotheses are derived from past experiences, and the inductive generalization is the choice of the behaviour that is most highly preferred in the given situation. The argument for this account of considered action, over say a behaviourist account, is that it is able to explain how it is that decisions are often based on possible outcomes, which are not environmental events. Whatever it is that conditions a response, it has to be able to cause that response, according to behaviourists; hence the mere possibility of some event cannot determine an organism's behaviour. Furthermore, behaviourists deny that internal mental states about the possibility of some event can cause behaviour. But then there is nothing in the behaviourists' ontology by which behaviour can be determined based on what an organism judges to be its best option in a given situation. Any model that does allow that organisms choose among various options in determining their behaviour is a form of hypothesis formation and confirmation, according to Fodor.

Models of perceptual integration follow the same pattern. Hypotheses must be

formed about the distal source of proximal stimulations. Typically there are many possible environments that could cause a specified sensory input. So perception involves hypothesizing about the nature of the environment and assigning a probability to each hypothesis by integrating sensory information with background information about the environment. An inductive generalization about the nature of the environment is made based on the probabilities each hypothesis receives. Fodor's defense of this position rests on the empirical fact that any information an organism gets about its environment is mediated by some sensory mechanism. Sensory mechanisms are sensitive to the physical properties of environmental events, but the ensuing perceptual judgments are typically not in terms of physical properties. That is, we typically perceive events individuated as types other than physical types, even though our sensory mechanisms only detect physical properties. This suggests that, at some level, possible perceptual categories are associated with sensory inputs. The association of perceptual categories with sensory inputs is a process of hypothesis formation and confirmation, based on previously confirmed associations.

In processes of hypothesis formation and confirmation, the data are represented and the hypotheses are formulated in terms of a set of symbols, also referred to as internal representations, that determine the conceptual and perceptual spaces of a creature. The processes of hypothesis formation and confirmation that constitute rational thinking are computational processes, where computations are manipulations of symbols based on rules concerned exclusively with the syntax of the symbols, i.e. non-semantic properties of the symbols. The computations are performed on tokens of

symbol types; symbol tokens are transformed into tokens of other symbol types. That is to say, the computations that constitute rational thinking are systematic transformations of tokens of internal representation types. The internal representation types, their tokens, and the rules by which internal representations are transformed constitute an internal representational system.[5]

1.3 What are the Characteristics of the Internal Representational System? (Strong Hypothesis)

The principal characteristic of the internal representational system that serves as the medium for rational thought, according to Fodor, is that it constitutes a *language*. But what exactly is being claimed in saying that the internal representational system is a language? It does not follow that just because a set of symbols operates according to rules that this system does constitute a language. In order for a representational system to be a *linguistic* representational system it must have certain properties. The key features of linguistic representational systems for my purposes are that they are productive and systematic.[6] The productivity of linguistic representational systems is simply that there is no limit to the number of expressions that are well-formed in a linguistic representational system. The systematicity of linguistic systems is that a system which can express certain propositions can express certain other propositions.

---

[5] I use the term "system" here, as opposed to "scheme" which is characterized purely in terms of syntax, because, though syntax determines the structure of the system, internal representations do have content.

[6] As Kaye points out, "no well-developed theory yet distinguishes linguistic from non-linguistic representations" (Kaye 1995, p.93). Thus, I take productivity and systematicity merely as necessary conditions for a system to be linguistic.

For example, any system that can express, "Mary is tall," and "Bob is strong" can express "Mary is strong".[7] Clearly natural languages possess these features. What explains the productivity and systematicity of natural languages is that "natural languages have a syntax which is compositional and recursive. That is, every natural language has minimal elements, and rules for combining and recombining these elements, such that the meaning of larger expressions is systematically dependent upon the meaning of its[8] parts, and how they're combined. ... natural languages have a **combinatorial semantics**" (Stainton 1996, p.114, emphasis in original). Thus, the productivity of natural languages is a result of the fact that the primitive meaningful elements can be combined, ad infinitum, in determinate ways to produce new meaningful elements. Natural languages are systematic because the rules by which primitive meaningful elements are combined to form complex meaningful elements ensure that if a system can form certain expressions, it can form many more with the same structure. Henceforth, I will take productivity and systematicity as necessary features for a system to be a linguistic representational system. So Fodor's claim that the internal representational system is a language of thought entails that the internal representational system is productive and systematic.[9]

---

[7] Note that because the resources by which natural languages are systematic are compositional, the "other propositions" that can be expressed are those that are expressed using the same structure. Though compositionality *explains* the systematicity of natural languages it need not be assumed for systematicity. See below.

[8] I take it that "its" should read "their", but I have followed the text in the quotation.

[9] The productivity and systematicity of the internal representational system follow as an immediate consequence of Fodor's anti-bootstrapping argument (1.9), in which he argues that the

It is consistent with the weak hypothesis that different creatures could possess different languages of thought, even within a single species. However, it follows from Fodor's reasoning that there is a single language of thought common to humans and animals that think. That is, humans and animals think by transforming the elements of a single set of symbols according to the same rules. Fodor's reasoning which entails this claim is based on empirical evidence suggesting homogeneities between the thought processes of humans and thinking animals (Fodor 1975, pp.57-8, especially note 3). Briefly, the idea is that the homogeneities are explicable if humans and animals share a common internal representational system, and with no serious alternative explanation on offer, Fodor draws the obvious conclusion; there is *one* language of thought.[10] Furthermore, Fodor argues that the language of thought must be innate in order to enable humans to learn the natural languages they speak. I consider these claims in detail below when I present Fodor's arguments that the language of thought cannot be a

---

internal representational system is at least as expressive as any natural language. Regardless of whether or not it has a combinatorial semantics, a system that is at least as expressive as any natural language must be productive and systematic because, by supposition, it can express whatever natural languages can express. Of course, it does not follow that the system has a combinatorial semantics. However, since animals and humans have only finite cognitive resources, the language of thought cannot be productive, and hence not as expressive as natural languages, if it does not have a combinatorial semantics.

[10] It might be wondered why animals could not have an internal representational system that is a subset of the human representational system. The problem with this suggestion for Fodor is that he argues that the language of thought cannot be a natural language because animals think (1.5). But if Fodor accepted that the internal representational system of animals was a smaller system than the system humans use, the conclusion would not follow. In fact, this suggestion is essentially my response to Fodor's argument that animals think (1.6). What Fodor is not committed to is that *any* creature that thinks has the same language of thought as humans. Aliens, for example, could have a different language of thought because there need not be homogeneities between their thinking and ours.

natural language.

Another characteristic of the language of thought is that the rules according to which internal representations are transformed are *biologically* determined by the possible state changes in the brain. The reason is that while the language of thought hypothesis is logically consistent with materialism, dualism, or idealism, Fodor places materialist constraints on the view, specifically token physicalism. For Fodor, every mental event is token identical with a physical event, thus the possible changes in brain states correspond to the possible transformations of the internal representations.[11] Fodor's claim that the transformations of internal representations occur according to rules amounts to the claim that the innate biological structure of the brain which determines the possible states of the brain can be intentionally characterized as determining under what conditions further symbols can be tokened given that a specific symbol is already tokened.

The rules according to which the language of thought operates are generally truth preserving. "The key to the nature of cognition is that mental processes preserve semantic properties of mental states; trains of thought, for example, are generally truth preserving, so if you start thinking with true assumptions you will generally arrive at conclusions that are also true" (Fodor 1987, p.154). However, since transformations of representations are token identical with physical transformations that occur in the brain, it is incumbent on Fodor to explain how the rules qua biologically determined physical

---

[11] In fact, a *subset* of the possible changes in brain states corresponds to the possible transformations of internal representations, since not all neural events are mental events.

state changes are generally truth preserving. Fodor's explanation is that organisms are simply built to make it the case that constraints on certain of their physical state changes are semantically evaluable as truth preserving transformations of internal representations. This explanation is motivated by taking very seriously an analogy between the language of thought and machine languages in computers.

A machine language is the internal system of representations in which computers perform calculations. That is, computers are built to compute in a machine language. According to Fodor, the language of thought is also an internal system of representations that creatures are built to use in order to perform computations that constitute thinking. A machine language performs a series of operations on symbols according to rules, where only the syntactic features of a symbol determine what can be done with it. However, computers are constructed so that the syntactic operations they perform correspond to semantic inferences. That is, computers are built so that their physical processes have a particular semantic interpretation. According to Fodor, the language of thought functions in the same way. The state changes that brains can undergo are determined by physical (neural) properties, but some of these transformations have a (naturalistic) semantic interpretation as truth preserving semantic inferences because of the way creatures are built.[12]

---

[12] "[A computer has] the following property: the operations of such a machine consist entirely of transformations of symbols; in the course of performing these operations, the machine is sensitive solely to the syntactic properties of the symbols; and the operations that the machine performs on the symbols are entirely confined to alterations of their shapes. Yet the machine is so devised that it will transform one symbol into another if and only if the symbols so transformed stand in certain *semantic* relations; e.g., the relation that the premises bear to the conclusion in a valid argument. ... Computers are a solution to the problem of mediating between the causal

The language of thought is a private language, in the sense that there are no public criteria or conventions for applying the symbols or terms of the language (Fodor 1975, p.68). Creatures that have a language of thought are simply built to use symbols in some ways, but not others that are logically possible. As I stated above, the manipulations of representations that a creature can perform are constrained by the way it is built, because such manipulations are token identical with physical brain processes. Creatures transform representations according to rules not by learning a set of rules and applying them; rather the physical state changes they can undergo, as determined by their physical makeup, *constitute* rules. Thus it is not public criteria or conventions that dictate the use of symbols, it is the private inner workings of creatures determined by their physical designs. The language of thought is *not* private in the sense that its terms refer to things that only its speaker can experience, such as sense data, etc.. Symbols in the language of thought can be representations of anything, including external objects and states of affairs.

Fodor argues that the language of thought is distinct from natural languages, such as English or Japanese.[13] His conception of the relation between the language of thought and natural languages is motivated by the analogy between the language of thought and machine languages. In addition to a machine language, computers operate in many other languages such as Basic, Pascal, and Fortran. These other languages are

---

properties of symbols and their semantic properties. So *if* the mind is a sort of computer, we begin to see how you can have a theory of mental processes that... explains how there could regularly be nonarbitrary content relations among causally related thoughts" (Fodor 1990, pp.22-3).

[13] Fodor's three arguments for this position are presented below.

compiled into the machine language, computations are performed in the machine language, and the results are decompiled back to the user-language. Similarly, whatever language a person might speak, the computations they perform that constitute thinking are conducted in the language of thought. Natural languages merely express those thoughts. Hence the language of thought must be a very rich system, which is to say it must possess a vast range of symbols.[14] The language of thought must contain symbols--not necessarily primitive--of everything for which we have a concept, since we could not think of something for which we had no representation if thinking is simply a process of transforming representations. It follows that, in particular, the language of thought must be at least as rich as any natural language because the thoughts expressed in a natural language are formulated in the language of thought (Fodor 1975, pp.82, 133).[15]

1.4 Argument for the Existence of an Internal Representational System

Fodor's argument that there is an internal representational system which serves as the medium for thinking is a very simple one:

(PR 1) Cognitive processes, such as considered action, concept learning, and perceptual integration, are computational processes. (PR 2) Computation presupposes a

---

[14] The view as presented is neutral concerning what proportion of symbols in the language of thought are primitive. The claim is simply that the language of thought is powerful enough to generate a vast range of symbols. Of course, Fodor's current position is that virtually all of the symbols are primitive, which can be taken as an additional characteristic of the language of thought; however, I will not develop this theme since an analysis of lexical decomposition is not pertinent to my argument.

[15] An analysis of this claim is a major portion of this dissertation.

representational system in which the computations can be performed. (Conclusion) It follows that any organism capable of cognitive processing must possess a representational system in which it performs the computations that constitute those cognitive processes.

Note that this argument merely supports the claim that thinking occurs in an internal representational system; it does not establish that an internal representational system must be a linguistic system. For this dissertation, I am granting the argument, though one could challenge either premise. Premise 2 holds only when certain assumptions are made about computational processes. Roughly, Fodor's view supposes the explicitness of the representations that, say, a connectionist system would not exhibit.[16] As Fodor says: "according to RTM [the representational theory of mind], mental processes are transformations of mental representations. The rules which determine the course of such transformations may, but needn't, be themselves explicitly represented. But the mental contents (the 'thoughts', as it were) that get transformed *must be* explicitly represented or the theory is simply false" (Fodor 1990, pp.23-4, emphasis in original). Fodor's justification of premise 1 rests on the fact that the only plausible models we have of processes of rational thought, such as considered action, concept learning, and perceptual integration, are computational models.[17] Not only does

---

[16] The intermediate processing steps of a connectionist system have no semantic interpretation, so the processing cannot be interpreted as transformations of internal representations. Thus, the operations of a connectionist system would model "brute incursions from the physiological level" rather than rational thinking, on Fodor's view.

[17] Fodor takes it that alternative accounts of cognitive processes, such as behaviourism, have been adequately refuted.

it not follow from this fact that no other kinds of models are possible, Fodor is also making the assumption that we are ontologically committed to the entities of our theoretical models (Fodor 1975, pp.51-2); one could dispute Fodor's weak hypothesis by challenging the assumption. Furthermore, Fodor's analysis of cognitive processes does not demonstrate that models of these processes are computational; rather, it makes the case that they are processes of hypothesis formation and confirmation. In order to infer premise 1 from his analysis[18], Fodor requires a connection between processes of hypothesis formation and confirmation and computational processes. Specifically, Fodor requires that cognitive processes of hypothesis formation and confirmation just are computational processes. Fodor himself does not make explicit any connection between these processes and he often uses the notions interchangeably; however, I think it is possible to argue for the required connection as follows.

Processes of hypothesis formation and confirmation are computational processes because organisms must represent not only actual circumstances but also non-actual, possible situations or outcomes with a probability assigned to each possibility. Those probabilities are computed, as is the degree to which a hypothesis is confirmed. Fodor

---

[18] Fodor explicitly denies that he is attempting to demonstrate that cognitive processes are computational (Fodor 1975, p.39, note 10); rather, he claims that he is simply assuming that they are--as a working hypothesis because it is presupposed by cognitive psychology--and investigating the consequences of this assumption. However, he does attempt to justify the assumption. In fact, the main theme of the entire first chapter (Fodor 1975) seems to be that considered action, concept learning, and perceptual integration are computational processes. For example, Fodor argues as follows: "My present point is that there is only one kind of theory that has ever been proposed for concept learning--indeed, there would seem to be only one kind of theory that is conceivable--and this theory is incoherent unless there is a language of thought. In this respect, the analysis of concept learning is like the analysis of considered choice; we cannot begin to make sense of the phenomena unless we are willing to view them as computational..." (Fodor 1975, p.36).

cites some evidence which indicates that the process of assigning a probability to a hypothesis is sensitive to the form in which the hypothesis is couched. For example, logically equivalent hypotheses are not, in general, assigned the same probability (Fodor 1975, pp.39-40, 57).[19] This is good evidence that the representation of a hypothesis is being syntactically manipulated in determining a probability assignment, which is what our notion of a computational process requires. The process of assigning a probability to some hypothesis begins with a representation which encodes the hypothesis and ends with a representation of the expected probability of the hypothesis; the assignment is simply a systematic[20] transformation of one representation into another, which just is the notion of computational processes we introduced. So it follows that processes of hypothesis formation and confirmation are computational processes as we have understood them; hence, cognitive processes are computational, as the argument for the weak hypothesis requires.

Fodor's analysis of considered action, concept learning, and perceptual integration, together with the argument that processes of hypothesis formation and confirmation are computational not only shows that cognitive processes are computational, but that their being rational processes at all is tied essentially to their being computational. Rationality seems to inherently require the representation and evaluation of various possibilities; a creature is rational when it adopts its best guess of

---

[19] Fodor cites Wason and Johnson-Laird 1972 as the source of this data.

[20] The transformation is systematic in the sense that the expected probability is that of the hypothesis whose representation is being transformed.

its best option in a given situation.[21] To do so it must have some kind of representation of its options and it must have a way to evaluate them. Thus an investigation of rational thought processes seems to suggest that they must be processes of hypothesis formation and confirmation, and empirical results support the claim that processes of hypothesis formation and confirmation are computational processes. It follows that rational thought processes must be computational processes.

Recall, that the entire analysis of rational thought processes serves to justify the assumption that these processes are computational. Then since computation presupposes a representational system in which to perform the computations, any organisms that exhibit rational thought processes must possess an internal system of representations.

1.5 First Argument that the Language of Thought is not a Natural Language (Animals Think)

As Fodor points out, an attractive way of satisfying the demands that we have a representational system for performing computations that constitute cognitive processes is to suggest that the required representational system is a natural language.[22] The appeal is that "it allows the theorist both to admit the essential role of computation (and hence of representation) in the production of behavior and to resist the more scarifying implications of the notion of a language of thought" (Fodor 1975, p.56). Supposing the medium in which we think is a natural language does not commit us to the existence of anything to which we are not pretheoretically committed anyway. However, Fodor

---

[21] This point is critical and will be discussed at length in chapters 4 and 6.

[22] Sellars (1956/1997), Harman (1974), and Field (1978) advance this view.

argues that this attractive line of reasoning leads to an untenable position. "The obvious (and, I should have thought, sufficient) refutation of the claim that natural languages are the medium of thought is that there are nonverbal organisms that think" (Fodor 1975, p.56). In particular, Fodor argues that the processes of considered action, concept learning, and perceptual integration are achievements of infrahuman organisms and preverbal children. Since these are computational processes they presuppose a representational system, and the representational systems of infrahuman organisms and preverbal children cannot be natural languages.

An interesting underlying assumption in the above argument is that infrahuman organisms and preverbal children think rationally by much the same processes as language-using humans. If the processes were not similar it would not follow that just because language-using humans and infrahuman organisms and preverbal children think that language-using humans do not think in a natural language. It could be the case that verbal humans think in natural languages while other thinking organisms use a different, simpler representational system for thinking. Only if verbal humans and other organisms think by much the same processes can it follow that natural languages cannot be the medium of our thinking. Fodor's defense of the claim that language-using humans and other organisms *do* think by virtually the same processes is based on empirical results. Empirical evidence suggests that "there are homogeneities between the mental capacities of infraverbal organisms and those of fluent human beings which, so far as anybody knows, are inexplicable except on the assumption that infraverbal

psychology is relevantly homogeneous with our psychology" (Fodor 1975, p.57).[23] In particular, what seems to be homogeneous is the representational system that verbal humans and other organisms use for thinking. Some evidence that the representational systems are the same is that nonverbal organisms seem to be sensitive to the same forms of hypotheses, i.e. the same representations, as verbal humans. For example, nonverbal organisms and verbal humans have greater difficulty learning disjunctive concepts than negative or conjunctive concepts (Fodor 1975, pp.57-8). The similarity in competencies between nonverbal organisms and verbal humans is evidence of similar representational systems. Clearly to the extent that verbal humans and nonverbal organisms use the same representational system for thinking, that system cannot be natural language.

1.6 Response to the Argument that Animals Think

One possibility that Fodor does not consider is that verbal humans share relatively simple representational processes with nonverbal organisms, explaining the homogeneities between our mental capacities; but in addition, verbal humans possess a much richer representational system, which might explain the much greater mental capacities of verbal humans. The augmented representational system of verbal humans would have to be as rich as any natural language, since it would have to express the predicates of any natural language, as Fodor claims[24]; but this view would not commit us to supposing that chimpanzees and rats have representational systems as rich as a

---

[23] The only evidence Fodor cites uses preverbal children as subjects. The source of the evidence is Fodor, Garrett, and Brill (1975).

[24] This argument is presented below.

natural language just because they exhibit some of the cognitive capacities that we have; nonlanguage-using organisms could have a relatively simple representational system they use for thinking--simple enough that the rules governing the use of the representations need not be linguistic. Verbal humans could possess a representational system sufficiently similar to the representational systems of nonverbal organisms to account for homogeneities in thought, but that system could be vastly augmented in verbal humans. The question is then open as to whether the augmented representational system of verbal humans could be a natural language. The argument that we share certain capacities with nonverbal organisms is irrelevant in determining whether we use natural language as a medium for thinking if some simple, non-linguistic representational system is sufficient for explaining those capacities.[25] The additional elements augmenting our internal representational system are, by supposition, unique to verbal humans.[26] "The short and simple answer to this [argument] is that the natural (spoken) language theory of thought can happily accept the claim that many thoughts do not occur in (mental tokens of) natural language" (Kaye 1995, p.102).

---

[25] Kaye makes the same point. "This interesting finding [the homogeneities of thought between verbal humans and non-verbal organisms] suggests that the representational systems of adults, infants and animals are similar, but does not imply that animals' and infants' representations are similar to our *linguistic* representations. ... the finding may show only that adult humans share some non-linguistic representations with infants and animals" (Kaye 1995, p.104, emphasis in original).

[26] Another way of putting this response is to point out that Fodor is begging the question in claiming that animals think. For example, in "Thought and Talk" Davidson denies this view (Davidson 1984, pp.155-170). In particular, there is no reason to suppose that just because animals demonstrate similar behaviours (intentionally characterized) to humans in certain situations that animals require a linguistic representational system. I return to this issue in (4.8), and a full discussion of it comprises chapter 6.

1.7 Second Argument that the Language of Thought is not a Natural Language (Innateness)

Fodor does not consider the question of an augmented representational system being a natural language, because he does not even entertain the possibility of an augmented representational system. But he offers a second line of reasoning that the language of thought is distinct from all natural languages, by arguing that the language of thought is innate. According to Fodor, in order to learn a language, an organism must learn the predicates of that language, which requires at least learning some determination of--though not a procedure for determining--the extensions of those predicates. However, a determination of the extension of a predicate is itself just another predicate. Thus, learning a language requires having a stock of predicates, coextensive with the predicates of the language being learned. Furthermore, not *every* predicate can be learned, since an organism learns even its first predicate in terms of some coextensive predicate[27]; some predicates must be innate. Now since all natural language predicates are learned, the innate predicates must be distinct from natural language predicates. The first predicates of a natural language that a child learns simply cannot be learned in terms of coextensive natural language predicates because the child does not yet know any coextensive natural language predicates.[28]

Learning what the predicates of a language mean involves learning a

---

[27] This point is developed and defended below.

[28] Of course, this reasoning does not preclude some natural language predicates being learned in terms of other natural language predicates, once sufficiently many natural language predicates have been learned.

> determination of the extension of these predicates. Learning a determination of the extension of the predicates involves learning that they fall under certain rules (i.e., truth rules).[29] But one cannot learn that *P* falls under *R* unless one has a language in which *P* and *R* can be represented. So one cannot learn a language unless one has a language. In particular, one cannot learn a first language unless one already has a system capable of representing the predicates in that language *and their extensions*. And, on pain of circularity, that system cannot be the language that is being learned. But first languages *are* learned. Hence, at least some cognitive operations are carried out in languages other than natural languages (Fodor 1975, pp.63-4, emphasis in original).

To complete the argument that the language of thought is distinct from any natural language it remains to be shown that the innate predicates in terms of which natural language predicates are learned are predicates of the language of thought. Fodor argues that we learn the predicates of a natural language by learning some determination of the extension of those predicates. Learning a determination of the extension of some predicate is, like all cognitive processes, a process of hypothesis formation and confirmation (Fodor 1975, p.59). Since natural language predicates are learned in terms of the innate predicates, these innate predicates must be the ones used in formulating the hypotheses to learn the natural language predicate. But since the language of thought just is the system of representations in which hypotheses are formulated and confirmed, the innate predicates in terms of which natural language predicates are learned must be predicates of the language of thought, as required. Now since the argument above showed that the innate predicates are distinct from natural

---

[29] Fodor defines a truth rule as an inductive generalization based on hypothesis formation and confirmation that determines the extensions of predicates for some language (Fodor 1975, p.59). A more thorough presentation of the notion of a truth rule is given below.

language predicates, it follows that the predicates of the language of thought are distinct from natural language predicates; hence, the language of thought is distinct from any natural language.

The key to the innateness argument as to why the language of thought cannot be a natural language is that natural languages are learned. In order to learn the predicates of a natural language, one has to learn (at least) a determination of the extension of those predicates. Since learning is a cognitive process, it involves formulating hypotheses and confirming them.[30] Fodor calls the hypotheses by which we learn a determination of the extension of a natural language predicate, truth rules. A truth rule for a predicate *P* is of the form:

(1) $\ulcorner P_y \urcorner$ is true iff x is G.

A truth rule is true if its substitution instances are true, where a substitution instance is obtained by:

> 1. Replacing the angles by quotes. ...
> 2. Replacing 'P_y' by a sentence whose predicate is *P* and whose subject is a name or other referring expression.
> 3. Replacing 'x' by an expression which designates the individual referred to by the subject of the quoted sentence (Fodor 1975, p.59, footnote 5).

It is clear from this definition that *G* and *P* are coextensive whenever (1) is true. Note also that the expression on the right-hand side of a truth rule need not be from the same language as the sentence quoted on the left. In fact, the point of Fodor's innateness

---

[30] Again Fodor's argument for this claim is that the only plausible models we have of learning is as a cognitive process of hypothesis formation and confirmation.

argument is that, at least in the initial stages of language learning, the expressions on the left and right-hand sides of a truth rule cannot be stated in the same language. The reason is that "and this point is critical, *G* in formula (1) is *used*, not mentioned. Hence, if learning *P* is learning a formula of form (1), then an organism can learn *P* only if it is already able to use at least one predicate that is coextensive with *P*, viz., *G*" (Fodor 1975, p.80). Now since in the initial stages of learning a language one is not able to use any predicate of that language, the right-hand expression must be from a language different from the language to which the predicate in the left-hand expression belongs. "Trivially, one cannot use the predicates that one is learning in order to learn the predicates that one is using" (Fodor 1975, p.82). In particular, the predicates in terms of which first natural languages are learned must not belong to any natural language.

As Fodor states, it is a "critical" point of the innateness argument that *G* is used in formula (1) and not mentioned. Thus, some independent reason that *G* must be used in (1) is required. Fodor does not explicitly offer such a reason, but one is readily available. Learning *P* could not be guaranteed if *G* were merely mentioned in formula (1) and not used. If *G* were merely mentioned, *G* could be a predicate whose extension the learner did not know. All she would learn is that *G* and *P* are coextensive; in fact, *G* and *P* could be predicates of the same language. Learning a rule similar to (1) except that *G* is mentioned and not used would not enable the learner to predicate *P* of anything if she could not predicate *G* of anything. And if there is nothing of which she

could predicate P, she would not have learned P.[31] Moreover, learning is not just formulating hypotheses; one of the hypotheses must be confirmed. Suppose a learner has formulated several hypotheses about a predicate P of the form H_i: $\ulcorner P_y \urcorner$ is true iff $\ulcorner G_{i,x} \urcorner$ is true[32], in which the G_i are merely mentioned and that the learner cannot use the G_i. Nothing can count as confirming evidence of one hypothesis over another, because she is unable to determine of anything whether it is in the extension of G_i, for any i. And since the language of thought is private, nothing about the way other people use the natural language term P can help her choose between the possible predicates G_i in the language of thought. Only by comparing how people use P to the way the G_i are *used* can she determine which G_i is coextensive with P, and by supposition she cannot use the G_i. So unless she can use the predicates in a language of thought, she cannot learn the predicates of the natural language.

---

[31] Fodor seems to forget this point in discussing verificationism:

> If, e.g., it is true that 'chair' means 'portable seat for one', then it is plausible that no one has mastered 'is a chair' unless he has learned that it falls under the truth rule '$\ulcorner$y is a chair$\urcorner$ is true iff x is a portable seat for one'. But someone might well know this about 'is a chair' and still not be able to *tell* about some given object (or, for that matter, about any given object) whether or not *it* is a chair. He would be in this situation if, e.g., his way of telling whether a thing is a chair is to find out whether it satisfies the right-hand side of the truth rule, and if he is unable to tell about *this* (or any) thing whether it is a portable seat for one (Fodor 1975, p.62, emphasis in original).

If someone really is unable "to tell about *this* (or any) thing whether it is a portable seat for one", then she is unable to use the predicate 'is a portable for seat for one', for surely using a predicate entails predicating it of things, hence she could not learn 'is a chair' in terms of 'is a portable seat for one'. Nonetheless, requiring that G is used in (1) does not commit Fodor to the position that the meaning of a predicate is given by a procedure for determining the extension of the predicate, only that to learn its meaning entails being able to use it, though, of course, not infallibly.

[32] I am using the notation $G_{i,x}$ to mean that G_i is a predicate whose argument is x.

1.8 Response to the Innateness Argument

The simple response to the innateness argument is that while it commits us to an innate internal representational system, this is something to which we were already committed in responding to the argument that animals think. Humans must possess an internal representational system similar to the internal representational systems of other thinking animals in order to explain the homogeneities in thought. Fodor's innateness argument makes a compelling case that early language acquisition is mediated by this internal representational system. What Fodor's argument does not demonstrate is that once a natural language is learned it cannot be the medium for thinking, nor that the internal representational system is a linguistic system. Learning the first predicates of a natural language could be much as Fodor presents it. Natural language predicates could be learned by formulating hypotheses that pair the natural language predicates with predicates belonging to the internal representational system, provided these hypotheses *use* the innate predicates. However, because of our finite mental capacities we cannot have infinitely many primitive innate predicates. Then, since there are infinitely many natural language predicates, not all of them can be coextensive with a primitive innate predicate. Some of our natural language predicates must be coextensive with combinations of other predicates. Now a correct hypotheses for learning a natural language predicate that is coextensive with a combination of some other predicates must use, not merely mention, the *combination* of those other predicates. However, to use a combination of predicates, they must actually be combined, and actually combining predicates *generates a new predicate*, because infinitely many combinations are possible

and our innate stock of primitive internal representations must be finite. Now nothing in Fodor's innateness argument precludes the possibility that combinations of predicates used to learn a natural language predicate are combinations of *natural language predicates*. In fact, since the combinations actually formed must be coextensive with predicates of natural language in order that those predicates might be learned, natural language imposes structure on the representational system that language-users employ. It is thus at least possible that the structure imposed is imposed on those elements of natural language that have already been learned, and not on the innate elements of the representational system. In fact, it could be the case that the internal representational system is not linguistic in that it is not productive or systematic, or is so to only a very limited degree. In this case, early language acquisition would serve simply to introduce natural language predicates coextensive with the predicates in the internal representational system. However, once a stock of natural language predicates had been learned, natural language would impose structure on them by which they could be combined to form other predicates that enabled the learning of more natural language, combinations that the elements of the internal representational system could not enter into.[33] At such a point, natural language could supersede the internal representational system as the medium of thought. Nonetheless, the homogeneities between our thinking and that of preverbal organisms would remain because what structure is present in the internal representational system would be embedded in the structure of natural language

[33] I develop this suggestion in chapter 4.

in virtue of the predicates in natural language coextensive with the predicates of the internal representational system.

It is important to notice that unless the predicates on the right-hand side of a truth rule are used, as Fodor's argument requires, it would not be necessary to actually generate combinations of predicates. All that would be required would be a representation of the combinations by which they could be mentioned in hypotheses for learning natural language predicates. In that case, natural language would not impose any structure on the internal representational system. However, since the predicates must be used, natural language does impose a structure on representations used in thinking. In particular, it is natural language that determines what combinations of predicates are actually produced in the formation of hypotheses for learning natural language.[34]

Fodor's innateness argument does not preclude the possibility that thinking occurs in natural language. It merely shows that early language acquisition must employ the internal representational system that we share with preverbal organisms. And though truth rules employed in early language acquisition must be expressed in the internal representational system, this expressive power is not indicative of a linguistic representational system, since every truth rule has exactly the same form, differing only in what predicates occur in each rule. In particular, the power of the internal representational system to express truth rules does not entail that the internal

---

[34] I will assume that natural languages do not differ at the level of logical form. Given this the specific natural language being learned is irrelevant to the structure imposed.

representational system is either productive or systematic.[35] However, my reasoning as to how natural language could be the medium of thought entails that, beyond early language acquisition, natural language actually increases the expressive power of our internal representational system. Fodor's third argument that the language of thought cannot be a natural language argues against exactly this possibility.

1.9 Third Argument that the Language of Thought is not a Natural Language (Anti-Bootstrapping)

The anti-bootstrapping argument argues *against* the possibility that "a foothold in the language having once been gained, the child then proceeds by extrapolating his bootstraps: The fragment of the language first internalized is itself somehow essentially employed to learn the part that's left. This process eventually leads to the construction of a representational system more elaborate than the one the child started with" (Fodor 1975, p.83). Specifically, the anti-bootstrapping argument tries to demonstrate that the language of thought is at least as expressive as any natural language, by denying the possibility of using one part of a natural language in order to learn another part of that natural language *that cannot be expressed in terms of the first part*. The reasoning is as follows.

Every predicate we learn is learned in terms of a coextensive predicate. Thus, we can only use the portion of a natural language we have learned to learn new predicates. That is, we cannot use a natural language to learn a predicate that is not expressible in terms of the portion of that natural language we already know, because

---

[35] I return to this point in chapter 6.

we cannot formulate a truth rule containing a predicate coextensive with the one to be learned; by supposition we do not know a predicate that is coextensive with the one to be learned. Now this reasoning extends to show that the language of thought must be at least as expressive as any natural language. The first natural language predicates we learn must be learned in terms of the language of thought, since by supposition we do not know any coextensive natural language predicates that we could use in a truth rule. So the first natural language predicates we learn can be expressed in the language of thought. Once we have learned some natural language predicates, any other natural language predicate we learn is either learned in terms of a coextensive predicate in the language of thought or a coextensive predicate in natural language. If it is learned in terms of a predicate in the language of thought, then clearly the language of thought can express that predicate. However, even if it is learned in terms of a natural language predicate we have already learned, *that* predicate is coextensive with some predicate in the language of thought. Hence the predicate we are learning is coextensive with a predicate in the language of thought, and so can be expressed in the language of thought. In general, if a natural language predicate *P* is learned in terms of another predicate *Q* in the same natural language[36], the language of thought can still express *P*, for the predicate *Q* in terms of which *P* is learned, being a natural language predicate, must itself have been learned. Again *Q* might have been learned in terms of a coextensive predicate in the language of thought or a coextensive predicate in natural

---

[36] I assume that *P* and *Q* are in the same natural language because the issues Fodor raises concern *first* language acquisition.

language. But, since we can only learn finitely many natural language predicates and the first predicates we learn *must* be learned in terms of some coextensive predicate in the language of thought, there must be a finite chain of coextensive predicates connecting any predicate we learn to a predicate in the language of thought. Now, since the biconditional used in truth rules is transitive, it follows that for every predicate we learn there is a coextensive predicate in the language of thought; hence, the language of thought is at least as expressive as any natural language. Fodor puts it thus:

> ...I have been saying that one can't learn *P* unless one learns something like '$\ulcorner P_y \urcorner$ is true iff *Gx*', and that one can't learn *that* unless one is able to use G. But suppose G is a predicate (not of the internal language but) in the same language that contains *P*. Then G must itself have been learned and, *ex hypothesi*, learning G must have involved learning (for some predicate or other) that G applies iff *it* applies. The point is that this new predicate must either be a part of the internal language or 'traceable back' to a predicate in the internal language by iterations of the present argument. In neither case however does any predicate which belongs to the same language as *P* play an essential role in mediating the learning of *P*. ... Nothing can be expressed in a natural language that can't be expressed in the language of thought (Fodor 1975, pp.83-4, emphasis in original).

As I mentioned above (1.3), that the internal representational system is a linguistic system follows as a consequence of the anti-bootstrapping argument. Natural languages possess all of the properties of linguistic representational systems, because *they are* linguistic representational systems. In particular, they are productive and systematic. Since the internal representational system is at least as expressive as any natural language, it must also be productive and systematic Thus, a consequence of the anti-bootstrapping argument is that the internal representational system is a *language* of

thought.[37]

This same reasoning actually serves as an independent argument for a version of the language of thought hypothesis. Human thought is expressed in natural languages, which are productive and systematic. This indicates that the medium underlying thought must be productive and systematic; so there must be a language of thought. "Linguistic capacities are systematic, and that's because sentences have a constituent structure. But cognitive capacities are systematic too, and that must be because *thoughts* have constituent structure. But if thoughts have constituent structure, then LOT is true" (Fodor 1987, pp.150-151, emphasis in original). This version of the language of thought hypothesis is not one I propose to challenge. Notice, however, that this argument does not identify the language of thought with the internal representational system; that identification requires the anti-bootstrapping argument, which I argue, in chapter 4, is unsound. Thus, I accept that there is an internal representational system, though I conceive its character differently from Fodor (5.7). I also accept that there is a language of thought. What I deny is that the internal representational system must be the language of thought. I argue that the internal representational system need not be linguistic, and that a natural language can be the language of thought.

---

[37] Technically, Fodor needs to show that any other feature necessary for being a linguistic system is one that the internal representational system has in virtue of being at least as expressive as natural languages which have the feature. For my purposes, productivity and systematicity are the only features I need consider.

1.10 Response to the Anti-Bootstrapping Argument

My short response to the anti-bootstrapping argument is that it is entirely possible that natural language predicates learned in terms of coextensive predicates in the internal representational system can be combined to form new predicates that are not expressible in the internal representational system, because predicates in the internal representational system cannot be so combined. It is then possible to learn natural language predicates that are not coextensive with any predicate in the internal representational system in terms of such combinations, from which it follows that natural languages have more expressive power than the internal representational system.[38] A full development of this response comprises chapter 4.

I turn now to an analysis of Fodor's theory of content for symbols in the language of thought. This analysis serves three purposes in the dissertation. First, I show that a result I establish in demonstrating why Fodor's theory of content fails contradicts the conclusion of the anti-bootstrapping argument, which suggests that the anti-bootstrapping argument is unsound. Fodor can only resist this conclusion by abandoning the notion that the content of symbols in the language of thought is determined by the causal relations between tokens of a symbol and the world. This move is open, but it is very unclear how Fodor can make it without giving up naturalism (2.2). A second purpose of my analysis is that in being very clear how

[38] Another possibility is that some terms of natural language are not learned. In my full response to the anti-bootstrapping argument (chapter 4), I claim that the logical connectives are such terms and that they allow us to learn predicates in a natural language that are not expressible in the language of thought.

symbols in the language of thought get their content, on Fodor's story, we can see that the content of the logical connectives is determined differently from that of other symbols. This exposes an implicit premise in the anti-bootstrapping argument as false. Once the premise is made explicit, we can see that the internal representational system need not possess anything corresponding to the logical connectives, and so need not be as expressive as natural languages, provided that animal cognition can be explained without requiring the internal representational system possess anything corresponding to the logical connectives. The third purpose of my analysis of Fodor's theory of content is that by understanding just why Fodor's theory fails, it is possible to take a different approach to content that avoids these difficulties. I present such an alternative approach to content in chapter 5, and I use this approach in chapter 6 to explain the behaviour of animals and preverbal children without requiring that the internal representational system is linguistic.

CHAPTER 2

FODOR'S THEORY OF CONTENT

2.1 Introduction

As we saw in chapter 1, Fodor's token physicalism together with the hypothesis that there is a language of thought, leads Fodor to the position that certain transformations of neural states just are transformations of symbols in an internal representational system.[1] The arguments we have considered so far are intended to show that there is an internal representational system, and to demonstrate some of the characteristics of that system. However, since token physicalism is not entailed by the weak hypothesis, but rather is a constraint Fodor places on the view, it is incumbent on Fodor to explain how it could be that neural states are endowed with intentional content. The sufficient conditions for neural states to have content that Fodor proposes provide the framework of a causal theory, which, like any causal theory of content, is subject to the criticism known as the disjunction problem. In this chapter, I present the naturalistic constraint, which Fodor imposes on materialist theories of content in order to respond to sceptics or outright eliminativists regarding intentional explanation, and what that constraint entails for Fodor's p ject. I explain the disjunction problem for causal theories of content, and present Fodor's theory of content, which includes the notion of asymmetric dependence as an attempt to solve the disjunction problem. I consider an objection to Fodor's theory due to Ned Block, to which, I argue, Fodor has

---

[1] As we saw in (1.3), Fodor's version of token physicalism is a thesis about both events and states.

an adequate response; however, I develop objections closely related to Block's objection, which I argue show that Fodor's theory cannot solve the disjunction problem.

2.2 The Naturalistic Constraint

> 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 upon their list. But *aboutness* surely won't; intentionality simply doesn't go that deep. It's hard to see, in face of this consideration, how one can be a Realist about intentionality without also being, to some extent or other, a Reductionist. If the semantic and intentional are real properties of things, it must be in virtue of their identity with (or maybe their supervenience on?) properties that are themselves *neither* intentional *nor* semantic. If aboutness is real, it must be really something else (Fodor 1987, p.97, emphasis in original).

"It counts as conventional wisdom in philosophy that (*i*) the intentional/semantical predicates form a closed circle and (*ii*) intentional states are intrinsically holistic" (Fodor 1990, p.51). I will refer to (i) and (ii) as the standard assumptions concerning intentionality. The first standard assumption, (i), is that it is not possible to formulate sufficient conditions[2] for realizing semantic/intentional predicates in an exclusively non-semantic/non-intentional language. That is, any sufficient conditions for satisfying a semantic/intentional predicate must be stated in a vocabulary

---

[2] Fodor never explicitly addresses the issue of what constraints are to be placed on the sufficient conditions, but he implicitly requires that they have some kind of explanatory power. That is, the conditions should reveal in some way why the neural states in question have the content that they do. For example, a sufficient condition for content is having a body molecularly identical to mine in a world molecularly identical to this one. (I got this example from Pietroski, personal communication, though he attributes it to Block.) However, such a condition has no explanatory power and so is not a solution of the kind Fodor requires to the problem of how neural states could have content.

that includes other semantic/intentional predicates. The point of assumption (ii) is that it is not possible to have isolated or singular intentional states. Anything that can be in one intentional state must be able to be in many other intentional states.[3] Fodor rejects both of the standard assumptions, thereby setting himself the goal of producing a theory of content "according to which (i) and (ii) are both false" (Fodor 1990, p.82, footnote 3); what Fodor calls a naturalized semantics.

The naturalistic constraint on theories of content, obtained by rejecting (i), is that sufficient conditions for realizing semantic/intentional predicates must be formulated in some *non-semantic/non-intentional* language. It is worth noting that holding the naturalistic constraint together with token physicalism commits Fodor to a kind of *type* identity theory. The reason is that, given his token physicalism, Fodor needs to explain how a particular neural event (tokening) can be a tokening of an intentional event type, and because of the naturalistic constraint, the explanation must be given in a *non-intentional* vocabulary. That is, Fodor's theory must show that a neural tokening has the content P, if Z, where Z is described in non-intentional language. Any tokening of a neural event type satisfying the conditions specified by Z has the content P. But then Z specifies a non-intentional type, all of whose tokenings have content P; i.e. the tokenings of type Z are a subset of the tokenings of type P.

---

[3] A strong motivation for holism is that intentional explanations of behaviour, i.e. explanations of behaviour in terms of beliefs and desires, seem to require that intentional states do not occur in isolation. Thus, in adopting atomism, the burden is on Fodor to show how the naturalistic sufficiency conditions he offers for content are adequate for grounding intentional explanation. My main criticism of Fodor's theory of content is that the sufficiency conditions cannot ground intentional explanations.

Hence, the naturalistic constraint commits Fodor to something much stronger than mere token identity, as expressed in, say, Davidson's position of anomalous monism.

Fodor's motivation for adopting the naturalistic constraint is to fend off scepticism regarding intentional explanation that he sees arising from the two standard assumptions concerning intentionality.[4] The most extreme challenge to intentional explanation comes from the eliminativists. Physicalists hold that every event can be described in the language of the natural sciences. In order to maintain intentional explanation, some connection between semantic/intentional predicates and predicates of the natural sciences is required. But the eliminativists take the assumption that the semantic/intentional predicates form a closed circle, as grounds for denying a connection[5] between the semantic/intentional predicates and the predicates of the natural sciences, and conclude that intentional explanation is to be eliminated from all scientific discourse. Fodor answers the challenge from eliminativism in the most direct way possible; he rejects assumption (i) that the semantic/intentional predicates form a closed circle; i.e. he adopts the naturalistic constraint. Thus, the naturalistic constraint is central in Fodor's overall project of defending intentional explanation. Without the naturalistic constraint, his view is in danger of collapsing into eliminativism.[6] In the

---

[4] While Fodor's theory of content rejects both of the standard assumptions concerning intentionality, the discussion in this chapter, except for a short section near the end of this chapter, focuses on the first assumption that the intentional/semantic predicates form a closed circle.

[5] This is a bit too strong. There could be some connection between semantic/intentional predicates and predicates of the natural sciences, but not one of interest to the sciences.

[6] The denial of assumption (i) is not the only way to establish a connection between semantic/intentional predicates and the predicates of the natural sciences. Davidson's position of

following discussion it will be important to keep in mind Fodor's strong commitment to the naturalistic constraint.

2.3 Causal Theories of Content and the Disjunction Problem

Very roughly, a causal theory of content says that the content of a symbol is the property of being whatever it is that causes the symbol to be tokened. For instance, if it is dogs that cause a symbol to be tokened, then the meaning of that symbol is the property of being a dog. The extension of the symbol "dog" is all and only dogs "[s]ince the extension of a symbol is just the set of things that have the property that the symbol expresses" (Fodor 1990, p.59). Of course, this characterization is much too crude to be a position anyone actually holds, but it serves to highlight the crucial difficulty confronting causal theories of content. Sometimes symbols are falsely tokened; that is, something that is *not* in the extension of the symbol causes it to be tokened. However, if the content of a symbol were simply the property of being whatever it is that causes the symbol to be tokened, it should not be possible to have false tokenings. Instead, a symbol tokened by things that have different properties would have as its content a disjunctive property. For example, if a fox causes a tokening of the symbol "dog", then it seems that the symbol "dog" should mean the property of being a dog *or* the property of being a fox, according to causal theories. But the content of the symbol "dog" is not the disjunctive property. Foxes *falsely* token the symbol "dog". The burden for causal theories of content is to explain false tokenings of

---

anomalous monism and Dennett's intentional stance accept (i) without denying intentional explanation, though they are interpretationist and so incompatible with Fodor's Intentional Realism.

symbols in terms of the causes of those tokenings.[7] Because, according to causal theories, the content of symbols is ostensibly a disjunctive property, the problem of explaining false tokenings in terms of their causes is known as the disjunction problem.[8]

Causal theorists have attempted to solve the disjunction problem[9], and it is crucial for them to do so, for without a solution their theories get the content of symbols wrong, which is entirely unsatisfactory. "Indeed, it is *so* unsatisfactory that the question whether a natural semantics is possible has recently come to be viewed as identical in practice to the question whether the disjunction problem can be solved within a naturalistic framework" (Fodor 1990, p.60). The exact nature of the disjunction problem depends on the kind of causal theory being offered, so I will present the specific issues that Fodor must address once I have presented his theory of content.

2.4 What is Fodor Offering?

Before presenting Fodor's theory, it is worth considering exactly what it is that Fodor is offering. He is not presenting a theory of content in the sense that he is claiming that the conditions he specifies are, in fact, the ones that obtain. Instead he is

---

[7] In fact, the disjunction problem is a special case of a more general problem of error that any theory of content must address, not just causal theories. That is, any theory must have an account of how we make the kinds of errors we do; those errors lead to the disjunction problem for causal theories of content.

[8] Another way of putting the disjunction problem is as follows: when things having different properties cause tokenings of a symbol, in virtue of what are those tokenings tokenings of the same symbol?

[9] Fodor (chapter 3, 1990) reviews some of the major causal theories of content and their attempts to solve the disjunction problem.

offering naturalistically specified sufficient conditions for symbols in the language of thought to have the content we intuitively ascribe to people's mental states: a possibility argument. If it were the case that the conditions he presents did obtain, then content could be accounted for in non-intentional terms. By only providing sufficient conditions, Fodor's aim is to show that the ostensibly daunting task of giving a naturalistic account of content is realizable, thereby responding to critics who argue that a naturalistic account is impossible. Of course, providing sufficient conditions does not guarantee that a naturalistic account of content is correct; thus, what Fodor is offering is rather modest, too modest, in fact. I will argue that the conditions Fodor offers could be sufficient for content only if there were no semantic associations between symbols in the language of thought. However, in that case little would remain of psychology, which is incompatible with Fodor's main project of preserving psychological explanation that originally motivated his attempt to give a theory of content.

2.5 Fodor's Theory of Content

Fodor's theory of content is a causal theory given in terms of nomic relations among properties. Fodor offers the following conditions as being sufficient for a symbol "X" in the language of thought to mean the property of being an X:

> 1. 'Xs cause "X"s' is a law.
> 2. Some "X"s are actually caused by Xs.
> 3. For all Y not=X, if Ys qua Ys actually cause "X"s, then Ys *causing* "X"s is asymmetrically dependent on Xs *causing* "X"s (Fodor 1990, p.121).

These conditions are naturalistic; they are stated in non-intentional terms and it would be possible for something to be in one intentional state with being able to be in any

others.

Condition 1 is actually an abbreviated way of saying that there is a nomic relation between a property of Xs (in virtue of which Xs cause "X" tokenings) and the property of being a cause of "X" tokenings (Fodor 1990, p.102 and p.121). It is important to notice that there could be, and in general there will be, several nomic relations between properties in virtue of which Xs cause "X" tokenings and the property of being a cause of "X" tokenings. The law that 'Xs cause "X"s' is a generalization of the nomic relations between properties of Xs and being a cause of "X" tokenings not statable in the terms in which the nomic relations are stated.[10] Notice that as long as there is at least one nomic relation between a property of Xs and the property being a cause of "X" tokenings, the generalizing law that Xs cause "X"s holds. This will be important for our later discussion, since it is crucial for asymmetric dependence. I will henceforth use the notation X → "X" to denote that 'Xs cause "X"s' is a law. So for example, stating that 'dogs cause "dog" tokenings' is a law, denoted dog → "dog", is shorthand for referring to all of the nomic relations between properties in virtue of which dogs cause "dog" tokenings and the property of being a cause of "dog" tokenings.

The first two conditions in Fodor's theory of content state that some X actually caused a tokening of the symbol "X", in virtue of one of its properties which is nomically related to the property of being a cause of "X" tokenings. These conditions

---

[10] Fodor clearly states that this is his position in "Special Sciences" (Fodor 1981, especially p.133).

alone do not avoid the disjunction problem, however, since some non-Xs might share a property with Xs that is nomically related to the property of being a cause of "X" tokenings. Fodor introduces the notion of asymmetric dependence, condition 3, to preclude non-Xs from being included in the extension of the symbol "X".

2.6 The Disjunction Problem for Fodor's Theory and Asymmetric Dependence

Merely being a reliable cause of "X" tokenings, for some symbol "X" in the language of thought, cannot be sufficient for being included in the extension of "X". The reason is that "it's a truism that *every* token of a symbol (including the false ones) is caused by something that has some property that is sufficient to cause a tokening of the symbol" (Fodor 1990, p.59). Now since some Ys, non-Xs, can have some of the properties that Xs possess, including properties in virtue of which Xs reliably cause "X" tokenings, Ys can also reliably cause "X" tokenings. So if being a reliable cause of "X" tokenings, i.e. satisfying Fodor's conditions 1 and 2, were sufficient for being included in the extension of "X", some Ys that are non-Xs would have to be so included. In that case, the symbol "X" would mean the property of being an X *or* the property of being a Y. For example, the property of being a colourless liquid reliably causes tokenings of the symbol "water". According to just conditions 1 and 2, rubbing alcohol, which has the property of being a colourless liquid, would have to be included in the extension of the symbol "water".[11]

Because Fodor's theory is given in terms of nomic relations among properties,

---

[11] If you are worried about the smell of rubbing alcohol, imagine someone with a cold. Other examples of this kind include leather and nylon, wool and acrylic, etc..

the disjunction problem for Fodor's theory arises only in cases of systematic error.[12] If it is not the case that 'Ys cause "X"s'[13] is an instantiated law, then conditions 1 and 2 preclude Ys from being in the extension of "X", even if there are occasions on which Ys cause "X" tokenings. However, the reason that we make *systematic* errors is that there are certain properties that are sufficient for reliably tokening a symbol. Anything having one of those properties can reliably cause tokenings of the symbol, whether or not it is in the extension of that symbol, because there is a nomic relation between each such property and the property of being a cause of tokenings of that symbol. Now if the extension of a symbol were to include everything that could reliably cause the symbol to be tokened, then all of the error cases would have to be considered as veridical. To preserve our intuitions that error cases are, in fact, false tokenings, Fodor's theory requires a condition that precludes non-Xs from being in the extension of the symbol "X", even when a particular non-X has some property that stands in a nomic relation to the property of being a cause of "X" tokenings.

The asymmetric dependence condition is meant to exclude all non-Xs from the extension of the symbol "X", thereby solving the disjunction problem. Intuitively, Fodor's idea is that errors depend on correctness, but not vice versa. In particular, Xs have a number of properties that are sufficient for reliably causing tokenings of the symbol "X". Non-Xs possessing some of these same properties or properties arbitrarily close to them--close enough to be mistaken for them--can also reliably cause tokenings

---

[12] Henceforth, I shall only be concerned with systematic errors.

[13] Of course, I am assuming that the Ys are non-Xs.

of the symbol "X", since these properties are sufficient for causing "X" tokenings. However, it is only because Xs have these properties that the properties are sufficient for reliably causing "X" tokenings. Non-Xs only reliably cause "X" tokenings by being mistaken for Xs.[14] So non-Xs' causings of "X" tokenings depend on Xs' causings of "X" tokenings, but not vice versa. For example, if it were not the case that water causes "water" tokenings in virtue of being a colourless liquid, then rubbing alcohol would not cause "water" tokenings.[15] The dependence is asymmetric, and it is in virtue of this asymmetry that Xs are in the extension of the symbol "X" and non-Xs are not.

Looked at more formally, asymmetric dependence is a relation between laws, specifically the laws 'non-Xs cause "X"s' and 'Xs cause "X"s'. Any property of non-Xs nomically related to the property of being a cause of "X" tokenings in virtue of which 'non-Xs cause "X"s' is a law is, or arbitrarily closely approximates, a property of Xs. Hence, in virtue of that same nomic relation, "Xs cause "X"s' is a law. And since this is true for every nomic relation connecting properties of non-Xs with the property of being a cause of "X" tokenings, supposing counterfactually that the X → "X" law did not hold--that is, if none of the nomic relations connecting properties of Xs with the property of being a cause of "X" tokenings held--, then no non-X → "X" connection would hold either. On the other hand, since no non-X has every property in virtue of which Xs cause "X"s--if it did it would be a special kind of X, in the way that small

---

[14] This assumption is implicit in Fodor's argument. My final point of this chapter is that the assumption is false.

[15] Unless, of course, rubbing alcohol shares some other property with water in virtue of which water causes "water" tokenings.

horses are still horses[16]--breaking the non-X → "X" law for some non-X would not break the X → "X" law. So the non-X → "X" law depends on the X → "X" law, but not vice versa. As a result of this asymmetric dependence, all and only Xs are included in the extension of the symbol "X".

2.7 Block's Objection[17]

A consequence of Fodor's theory is that breaking the X → "X" law breaks the Y → "X" law, for all Y. Thus, counterfactually, were it not the case that "dog" meant the property of being a dog, it wouldn't mean anything on Fodor's story. It is this consequence that Block challenges. Block begins his objection by questioning exactly what it is that Fodor means by the symbol "X" in the language of thought. Block suggests that there are exactly two possible readings of the symbol "X", neither of which will satisfy Fodor's conditions. The two possibilities are that either "X" refers to some uninterpreted string #X#[18], i.e. it is a syntactic item, or it has a semantic value. Block objects that in the first case, the necessary counterfactuals for asymmetric dependence are not satisfied; while in the second case, the counterfactuals can hold as required for asymmetric dependence but this account makes use of semantic notions

---

[16] Again, this is an implicit assumption in Fodor's argument that I challenge later in the chapter.

[17] This objection is presented in Fodor 1990, pp.111-114. It is worth noting that the only account of Block's position is Fodor's presentation of that position.

[18] A word appearing in double quotes refers to a symbol in the language of thought; and a word between number signs is an uninterpreted string; i.e., a symbol viewed syntactically. So for example, "dog" is a symbol in the language of thought, and #d^o^g# is an uninterpreted string. The same convention will hold with variables in place of words.

unavailable to Fodor. Fodor accepts that the second option in which the symbol "X" is individuated semantically is not open to him because of the naturalistic constraint, so that the symbol "X" must be an uninterpreted string #X#. "Block is, of course, perfectly right that for the purposes of a naturalistic semantics the only nonquestion-begging reading of "cow" is #c^o^w#. Henceforth be it so read" (Fodor 1990, p.112). But when we read the symbol "X" as the uninterpreted string #X#,[19] Block argues, then Fodor's theory that #X# means the property of being an X because Y → #X# laws depend asymmetrically on the X → #X# law is just false. Let's look more closely at how the objection goes.

Fodor is committed to the non-semantic reading of the symbol #c^o^w# as an uninterpreted string. Given this reading, Block points out that "there is surely a possible world in which cows don't cause #c^o^w#s but trees do, viz., *the world in which #c^o^w# means tree*" (Fodor 1990, p.111, emphasis in original). However, according to Fodor's asymmetric dependence condition, there should not be any possible worlds in which only non-cows have properties in virtue of which they reliably cause #c^o^w# tokenings. This follows because breaking the cow → #c^o^w# connection is meant to break every Y → #c^o^w# connection on Fodor's story. But with a non-semantic reading of the symbol #c^o^w#, it does not seem to be the case that breaking the cow → #c^o^w# connection precludes the possibility of having some Y → #c^o^w#

---

[19] The reader will now be aware that the notations "X" and #X# both indicate a symbol in the language of thought on this reading. For the remainder of this chapter, I will use the #X# notation to emphasize that Fodor is committed to reading the symbol as an uninterpreted string, that is, purely syntactically.

connection.[20] The symbol #c^o^w# qua uninterpreted string has no meaning until it is connected to the world via causal links on Fodor's story. But it must be possible for those causal links to be different from the ones in our world if the symbol really is an uninterpreted string--otherwise the symbol's meaning would be fixed by the necessary causal connections between it and the world and this account would again appeal to a semantic reading of the symbol. Thus, the nomic relations between properties of cows and the property of being a cause of #c^o^w# tokenings is a contingent fact about our world. It follows that there are possible worlds in which these nomic relations do not obtain and something else has properties nomically related to the property of being a cause of #c^o^w# tokenings. But such a world is one in which only non-cows cause #c^o^w#s, so the symbol #c^o^w# does not satisfy the asymmetric dependence condition. Block concludes that since nothing particular to the symbol #c^o^w# has been used in this discussion, the conclusion will follow for any symbol. That is to say, there are no symbols that satisfy the asymmetric dependence condition when symbols are taken to be uninterpreted strings, as Fodor must read them.

2.8 Fodor's Response to Block

Fodor responds to Block by introducing a ceteris paribus clause into the asymmetric dependence condition. The original idea was that were it not for the X → #X# connection, then there would be no Y → #X# connection either. Block's objection

---

[20] The reader may notice a confusion between worlds in which no X → #X# connection obtains and ones that differ from the actual world insofar as the X → #X# relation obtaining in the actual world is broken. This confusion is Block's and not mine, and it is what Fodor exploits in order to reply to Block's objection.

to this original idea is that, it *could* be the case that every nomic relation between properties of Xs and the property of being a cause of #X# tokenings did not obtain, yet still there is a nomic relation between some property of a non-X and the property of being a cause of #X# tokenings. This would be the case if #c^o^w# meant the property of being a tree. By introducing a ceteris paribus clause into the asymmetric dependence condition, Fodor aims to rule out the problem situations Block imagines. Fodor's amended proposal is that *all else being equal*, if there were no X → #X# connection then there would be no Y → #X# connection, but not vice versa. In order to understand exactly how the amended asymmetric dependence condition is supposed to overcome the difficulties to which its predecessor falls prey, it is important to emphasize that the connections we are considering--X → #X#, Y → #X#-- are "nomic relations among properties rather than causal relations among individuals" (Fodor 1990, p.100). What this means is that the relevant counterfactual situation for considering the asymmetric dependence condition is one in which there is no nomic relation connecting any property in virtue of which Xs cause the symbol #X# to be tokened with the property of being a cause of #X# tokenings; there is no X → #X# law. Now Ys cause #X#s by being mistaken for Xs.[21] That is to say, any property in virtue of which Ys cause #X#s is a property that Xs have, or a property arbitrarily[22] close to one Xs have. So, all else

---

[21] As stated above, this assumption is central to Fodor's argument, and challenging it will be the basis of my objection later in the chapter. Of course, I am ignoring the possibility that a symbol has two meanings, which is irrelevant to this discussion.

[22] The reason the Y property must be arbitrarily close to the X property is that it is one leading to error, and if it were possible to discriminate the X and Y properties no error would occur. This does not mean that there is *no* way the properties can be discriminated, only that in the

being equal, in the relevant counterfactual situation there cannot be a Y → #X# connection.

If there were a nomic relation connecting a property of Ys with the property of being a cause of #X# tokenings, two possibilities would arise. First, the property in virtue of which Ys cause #X#s could be a property that Xs also possess, in which case, according to that very relation, there would be regular instances in which Xs would cause #X#s, contrary to the counterfactual assumption. The other possibility is that the property in virtue of which Ys cause #X#s could be one arbitrarily close to a property Xs possess. Now since the properties are arbitrarily close, it would be impossible in practice to distinguish between instantiations of the two properties in the relevant circumstances. Given the Y → #X# law, Xs would reliably cause #X#s precisely because we could not in practice distinguish between instantiations of the properties; in virtue of the same nomic relation covered by the Y → #X# law, Xs would reliably cause #X#s, again contrary to the counterfactual assumption. In short, every Y → #X# connection is subsumed under the X → #X# connection, since it is only in virtue of having certain properties that Ys can cause #X#s, and Xs have those same properties or properties arbitrarily close to them. Notice that without the ceteris paribus clause, a property in virtue of which Ys cause #X#s would not have to be one that Xs share, and so it would be possible to establish a Y → #X# connection under the less stringent counterfactual condition that simply no X → #X# connection exists. This is how Block generated his

---

relevant conditions they cannot be.

counterexample to Fodor's original proposal, but for the reasons just given, Fodor is able to preclude the unwanted cases by introducing a ceteris paribus clause.

2.9 Analyzing Fodor's Response in Terms of Possible Worlds

Curiously enough, Fodor mischaracterizes his own position when he considers what must be the case in terms of possible worlds given the amended asymmetric dependence condition. Fodor claims that "what's required is just that there be worlds where cows cause "cows" and non-cows don't; and that they be nearer to our world than any world in which some non-cows cause "cows" and no cows do" (Fodor 1990, p.113).[23] This would require that worlds in which we never systematically[24] mistake anything for a cow be closer to the actual world than ones in which #c^o^w# means tree, and we never systematically mistake cows for trees. Intuitively this position seems to be wrong. However, its incorrectness is irrelevant; this is *not* the position Fodor has to defend, because it does not correctly capture the amended asymmetric dependence condition.

Fodor's claim is that, all else being equal:

[C1] If the nomic relations subsumed by the X → #X# law did not obtain, then there would not be a non-X → #X# law either;

and, it is *not* the case that

---

[23] I take both instances of "cows" in this sentence to be errors for "cow"s, though I have followed the text in the quotation. This point is substantiated by the fact that "cow"s is used in the same context in the previous sentence in the text (Fodor 1990, p.113).

[24] Notice that we could still make mistakes in this situation provided they were not systematic; i.e. provided there were no nomic relations between properties of non-cows and the property of being a cause of "cow" tokenings.

[C2] If non-X → #X# laws did not obtain, then the X → #X# law would fail.

In translating these conditions into the language of possible worlds, Fodor seems to conflate the situations for testing them. The assumption in C1 is that there is no X → #X# connection. In the context of the example, this means that possible worlds in which only cows cause #c^o^w#s are irrelevant to the discussion. Similarly, in C2 there are no non-cow → #c^o^w# laws, so worlds in which only some non-cows cause #c^o^w#s should not be considered. The comparison that Fodor suggests between worlds in which only cows cause #c^o^w#s and worlds in which only non-cows cause #c^o^w#s cannot establish anything for Fodor, since the first class of worlds is precluded by C1 and the second class is precluded by C2. Of course, each of these classes is relevant in considering Fodor's main claim that C1 holds while C2 does not. But two separate comparisons must be made, each involving one of these classes of possible worlds, rather than a single comparison between the two as Fodor suggests.

Fodor's amended asymmetric dependence condition is a claim that, all else being equal, one counterfactual conditional, C1, holds, while another, C2, fails. In order to cash this out in terms of possible worlds, each condition must be considered separately. According to condition C1, there is no nomic relation between any property of Xs and the property of being a cause of #X#s; there is no X → #X# connection. In the language of possible worlds, we must not consider any worlds in which Xs reliably cause #X#s. The question of whether or not C1 holds thus becomes: In the class of possible worlds in which there is no nomic relation between any property of Xs and the property of being a cause of #X# tokenings, are the worlds in which some non-Xs

reliably cause #X#s or those in which they do not reliably cause #X#s nearer to the actual world? Fodor's position that C1 holds is the view that, in this counterfactual situation, worlds in which non-Xs do not reliably cause #X#s are nearer to the actual world than ones in which they do. Similarly, the condition C2 stipulates that non-X → #X# connections cease to obtain. Whether or not C2 holds translates into the question: In the class of possible worlds in which non-Xs do not reliably cause #X#s, are the worlds in which Xs reliably cause #X#s or the ones in which they do not closer to the actual world? In this case, Fodor's view that C2 fails translates to the position that worlds in which Xs do reliably cause #X#s are nearer to the actual world than ones in which they do not.

Before analyzing Fodor's claims about the counterfactual conditions C1 and C2, a methodological point is in order. In order to determine which possible worlds are nearer to the actual world, Fodor considers how many nomic relations existing in the actual world would have to be broken and how many new ones added to arrive at a particular possible world (Fodor 1990, p.113). The fewer the changes, the nearer the world. This is exactly the interpretation of a ceteris paribus clause we would expect. Thus, a natural metric imposed on possible worlds by this account is the number of changes required to transform the actual world into the counterfactual situation under consideration. I will call this the C-metric, for 'changes metric'. Notice that the C-metric is not only sensitive to changes involving nomic relations and so is more fine-grained than what Fodor considers. For our purposes, however, the two notions coincide, since we will only need to consider the elimination and introduction of nomic

relations. We now use this machinery to evaluate Fodor's claims that C1 holds and C2 fails.

I will first restate, in terms of possible worlds, Fodor's argument that C1 holds, and show that this adequately answers Block's objection. Without the ceteris paribus clause, the asymmetric dependence condition states that there are no possible worlds in which Xs do not reliably cause #X#s, but some non-Xs do. Block's objection that #c^o^w# could mean tree correctly shows this to be false. However, according to the amended asymmetric dependence condition, the mere existence of the worlds Block imagines are not problematic for Fodor. Block's #c^o^w#-means-tree worlds must actually be closer to our world than Fodor's #c^o^w#-means-nothing worlds in order for Block to sustain his objection to Fodor's position. Now in the C1 counterfactual situation, the X → #X# connection does not obtain. But any property in virtue of which some non-X causes #X#s in the actual world is a property Xs also have or is arbitrarily close to a property Xs have; so, a non-X → #X# connection could only be in place if Xs reliably caused #X#s, which, by counterfactual assumption, they do not. Thus, it follows that any property in virtue of which non-Xs cause #X#s in the actual world cannot support a non-X → #X# connection in the counterfactual situation. So minimally, the possible worlds we are considering must be different from our world to the extent that the X → #X# connection and everything it subsumes are not in place. Fodor's #c^o^w#-means-nothing worlds exactly satisfy this minimal requirement. Block's #c^o^w#-means-tree worlds, on the other hand, additionally require the instantiation of some new law, Y → #X#. "What the present theory claims is that, in the world that's

just like ours except that cow → "cow" and everything nomologically dependent on it are gone, X → "cow" is false for all X .... Well if this is what you mean by 'the nearest possible world in which cow → "cow" is gone', then, clearly Block's world doesn't qualify. To get Block's world, you have to both break cow → "cow" and stipulate tree → "cow" " (Fodor 1990, p.113). More changes are required to the actual world in order to obtain Block's #c^o^w#-means-tree worlds than Fodor's #c^o^w#-means-nothing worlds; so, according to the C-metric, Fodor's worlds are nearer to the actual world, as required.

2.10 Preliminaries to Related Objections

Block's objection suggests other ways in which Fodor's position can be challenged. Two assumptions underpin Fodor's use of the asymmetric dependence condition in his account of content:

[A1] Xs possess properties in virtue of which they cause #X#s that non-Xs do not possess nor do non-Xs have properties that approximate them arbitrarily closely; and,

[A2] non-Xs do not possess properties in virtue of which they cause #X#s that Xs neither possess nor approximate arbitrarily closely.

A1 is the basis for an argument that C2 fails; A2 is used to justify C1, as we have just seen. It is possible to question whether C2 fails, as Fodor requires, by challenging A1. Also, since there are semantic associations between symbols in the language of thought, a symbol can be tokened, *non-falsely*, by something that is not in its extension, providing grounds for denying A2 that Fodor's response to Block does not address. Before considering these objections, however, we need to reconsider exactly how the

asymmetric dependence condition is meant to solve the disjunction problem.

As presented, Fodor's asymmetric dependence condition does not solve the disjunction problem. The asymmetric dependence condition is intended to determine which among certain possibilities for the content of a symbol is the actual content. Recall exactly what the possibilities for the content of a symbol are. If a fox causes a tokening of the symbol #d^o^g#, then it seems that the symbol #d^o^g# should mean the property of being a dog *or* the property of being a fox, according to causal theories of content, when, in fact, #d^o^g# means the property of being a dog. So the two possibilities for the content of #d^o^g# are: (a) the property of being a dog; or, (b) the property of being a dog or the property of being a fox. Whatever Fodor offers must pick out (a) for the content of #d^o^g# over (b). In particular, what the theory is *not* required to do is choose between: (a') the property of being a dog; and, (b') the property of being a fox. But it is the latter properties that Fodor's asymmetric dependence condition is designed to choose between. In general, the question posed by the disjunction problem is: why does #X# mean (the property of being an X) and not (the property of being an X $\vee$ the property of being a Y)?[25] The problem is not that there is a disjunction of possible meanings, it is that one of the possible meanings is a disjunction *that includes the property in the other possibility*. By only considering Y not=X in the asymmetric dependence condition, Fodor has not addressed the heart of the disjunction problem. Furthermore, given the possibilities, the dependence is not

[25] The parentheses are used to make clear what the possibilities are. #X# could mean (the property of being an X) or (the property of being an X $\vee$ the property of being a Y). The possibilities are not (the property of being an X) or (the property of being a Y).

asymmetric. Break the X → #X# connection and you might break the (X ∨ Y) → #X# connection. But certainly breaking the (X ∨ Y) → #X# connection breaks the X → #X# connection. So, the dependence is symmetric and Fodor's conditions do not solve the disjunction problem.

There is a way to amend Fodor's conditions so that despite the symmetrical dependence of possibilities, the meaning of #X# comes out right. Fodor can simply place a condition on when a symbol's meaning is a disjunction of properties. Intuitively, a disjunction of properties should not symmetrically depend on one of the disjuncts. That is, if the meaning of some symbol #Z# is a disjunction of properties, say the property of being an X or the property of being a Y, then just breaking the X → #Z# connection or the Y → #Z# connection should not break the (X ∨ Y) → #Z# connection. The reason is that there are properties of Ys which Xs do not possess and properties of Xs which Ys do not possess that are nomically related to the property being a cause of #Z# tokenings.[26] So, if breaking the X → #Z# connection does break the (X ∨ Y) → #Z# connection, then the meaning of #Z# is not a disjunction of properties. The only possibility remaining is that #Z# means the property of being an X. In our example, breaking the dog → #d^o^g# connection breaks the (dog ∨ fox) → #d^o^g# connection, so the meaning of the symbol #d^o^g# is not a disjunction of properties; #d^o^g# means the property of being a dog. The general amendment then, is that when one possible meaning of a symbol is a property that is also included in a disjunction

[26] Though this reason is not one Fodor offers, it is consistent with the assumptions on which his position rests, namely A1 and A2.

constituting the other possibility so that the two[27] possibilities depend on each other symmetrically, the meaning of the symbol is the simple property; i.e. not the disjunction.

With this amendment to Fodor's conditions, we can readily see how the asymmetric dependence condition is meant to solve the disjunction problem. Given laws X → #X# and (X ∨ Y) → #X# that depend on each other symmetrically, according to the amended conditions the meaning of #X# is the property of being an X, *provided* the Y → #X# law and (X ∨ Y) → #X# law are not also symmetrically dependent. In that case there would be nothing to choose between (the property of being an X) and (the property of being a Y) for the meaning of X. The amended conditions would pick out *both* (the property of being an X) and (the property of being a Y) but not the disjunctive property (the property of being an X or Y) as the meaning of #X#. However, the Y → #X# law and (X ∨ Y) → #X# law will not be symmetrically dependent if the X → #X# law and Y → #X# law are not symmetrically dependent. Since we are supposing that breaking the X → #X# law breaks the (X ∨ Y) → #X# law, X → #X# and Y → #X# will not be symmetrically dependent when the Y → #X# law depends asymmetrically on the X → #X# law. That is, the amended conditions provide a solution to the disjunction problem if the asymmetric dependence condition holds; i.e., when C1 is true and C2 is false.

---

[27] The amendment can be extended to more than two possibilities in the obvious way. I have only considered two possibilities in my presentation for the sake of clarity.

Fodor's argument that C2 fails would have to go something like the following.[28] The counterfactual assumption is that every nomic relation in the actual world between the properties in virtue of which non-Xs cause #X#s and the property of being a cause of #X#s ceases to obtain. It is required to show that worlds in which the X → #X# law holds are nearer to the actual world than worlds in which the X → #X# law does not hold. However, the counterfactual assumption does not entail the failure of every nomic relation between properties in virtue of which Xs cause #X#s and the property of being a cause of #X#s. Since non-X → #X# laws arise from the fact that some non-Xs can be mistaken for Xs, any property in virtue of which some non-X causes #X#s must be one of the properties in virtue of which Xs cause #X#s, or arbitrarily close to such a property. Now if we break the nomic relations between non-X properties and the property of being a cause of #X# tokenings, there remain other properties that Xs possess in virtue of which they can cause #X# tokenings (assumption A1). Thus, making only the changes to the actual world that are minimally required by the counterfactual assumption results in a world in which the X → #X# law obtains. In order to break the X → #X# connection, we would have to eliminate all of the other nomic relations under which Xs reliably cause #X#s. This would involve more changes to the actual world than minimally required, however, making the worlds without an X → #X# law farther from the actual world, according to the C-metric, than the worlds in which the X → #X# law obtains, as required.

---

[28] Fodor does not actually present an argument that C2 fails since the Block objection focuses on C1. What follows is my own construction of what I think Fodor would say, given his other arguments and the overall position he is defending.

The above argument that C2 fails depends on assumption A1, that Xs possess properties in virtue of which they cause #X#s that non-Xs do not possess nor do non-Xs have properties that approximate them arbitrarily closely. To analyze the argument, it is important to be very clear as to what sorts of properties of Xs are nomically related to the property of being a cause of #X# tokenings. Obviously, if the property of being an X is one of the properties in virtue of which Xs cause #X# tokenings, then A1 is trivially true. All and only Xs possess the property of being an X. In defending his position against various objections, it seems that Fodor *does* think that the property of being an X is one of the properties in virtue of which Xs cause #X#s. For example, "it must be the property of *being a horse* and not the property of *being a small horse* that is connected with the property of *being a cause of "horse" tokens*" (Fodor 1990, p.102, emphasis in original). And, "[t]he semantics of the word "virtuous," for example, is determined by the nomic relation between the property of being a cause of tokens of that word and the property of being virtuous" (Fodor 1990, p.111). However, Fodor's appeal to the property of being an X in these examples is shorthand for a cluster of nomic relations that must be specified in one or other of the special sciences, nomic relations that might be beyond the current state of the art of the appropriate special science to articulate. If appeals to the property of being an X were not shorthand, it is hard to imagine that Fodor would need to dedicate an entire chapter of a book to the defense of his position. The defense could be done in one sentence. Since all and only Xs possess the property of being an X, which is a property in virtue of which #X#s are tokened, every non-X → #X# connection depends asymmetrically on the X → #X#

connection. That Fodor did write a chapter defending the asymmetric dependence condition suggests the issues are somewhat more subtle. No one doubts that the content of the symbol #X# is the property of being an X.[29] The point of Fodor's project is to provide a naturalized semantics that shows this fact, and simply restating it will not suffice. The very legitimacy of intentional explanation is what is at issue.[30] Fodor must account for the connection between the property of being an X and the property of being a cause of #X# tokenings within a naturalized semantics; hence, he cannot appeal to the property of being an X.

So, what are the properties in virtue of which Xs cause #X# tokenings to which Fodor can appeal? Since intentional explanation cannot be used in a naturalized semantics, the nomic relations available to Fodor are those in virtue of which Xs cause #X# tokenings stated in some physicalistic vocabulary. Given Fodor's token physicalism--any tokening of the symbol type #X# is identical to some tokening of a neural type--the appropriate special science for articulating the relevant nomic relations in virtue of which Xs cause #X# tokenings is neuroscience. And, since laws can only be stated within the discourse of a particular science[31], the *properties* of Xs to which Fodor can appeal are those that can appear in laws of neuroscience, and these are *detectable*

---

[29] Of course, some people doubt whether there is a symbol #X#, but that is a separate issue.

[30] Note that I am not suggesting that Fodor has got anything wrong by using this shorthand. My point is simply to make explicit what "the property of being an X" is shorthand for, from which I will generate an objection to Fodor's conditions.

[31] See "Special Sciences" in Fodor 1981.

properties.[32] We can see more directly now why the disjunction problem manifests itself for causal theories of content and why it is such a problem. In general, every detectable property of Xs is a property some non-X can have, or can be approximated arbitrarily closely by a property some non-X can have.[33] Since detectable properties are the only properties in virtue of which Xs cause #X#s to which causal theories can appeal, excluding non-Xs from the extension of #X# is a non-trivial matter.

2.11 First Objection

Possessing any specified detectable property of Xs is either sufficient for being an X or it is not. If possessing the property is not sufficient for being an X, then clearly some non-X can possess it, and token #X#s in virtue of that property. If possessing the property *is* sufficient for being an X, then clearly no non-X can possess it. Nonetheless, provided a non-X can possess a property that approximates the X property arbitrarily closely, the non-X can still reliably cause #X# tokenings. Now the detectable properties of Xs are those *physical* properties to which our sense organs are sensitive. But physical properties our sense organs can detect are measurable, and hence quantifiable. The X property can be specified by some numerical value resulting from an appropriate

---

[32] Fodor himself makes the point that micro-level properties are not semantically relevant (Fodor 1990, p.117). The point is that micro-level nomic relations must involve micro-level properties; so no micro-level property is nomically related to the property of being a cause of #X#s, a macro-level property. Of course, a micro-level law might provide the mechanism for the macro-level X → #X# law, but it does not thereby connect micro-level properties with the property of being a cause of #X# tokenings.

[33] There might be some Xs that have detectable properties sufficient for causing #X# tokenings that no non-X can possess and that cannot be approximated arbitrarily closely by a property some non-X can have, though I haven't been able to think of any. The point is that this is not the case in general so that Fodor's conditions do not provide a general theory of content. The argument for this claim is presented below.

measurement. Provided the measurement has a continuous range, the X property can be approximated arbitrarily closely by a property some non-X can have. But in general, macro-level detectable properties *do* have a continuous range[34]. These are properties that can appear in laws of neuroscience, so they must be the properties to which are sensory organs are sensitive; for example, frequency, intensity, and amplitude of light and sound waves. We are also capable of detecting various shapes through sight and touch, but shapes can be approximated arbitrarily closely by other shapes. For example, suppose that Xs are triangles. Possessing the property 'triangularity' is clearly sufficient for being a triangle. However, since degenerate quadrilaterals are triangles, quadrilaterals can approximate triangles arbitrarily closely, and so reliably token #t^r^i^a^n^g^l^e#. The point is that, in general, non-Xs can possess every property in virtue of which Xs cause #X#s or possess properties that approximate them arbitrarily closely. Exceptions will be Xs that possess macro-level detectable properties sufficient for being an X, such that measurements of that property have a discrete range.[35] However, in general, Xs do not possess such properties because of the nature of our sensory organs; i.e. what our sensory organs are sensitive to can be measured over a

---

[34] Micro-level measurements often have a discrete range because of the discreteness of the quantum numbers, but the properties with which we are concerned are exclusively macro-level properties.

[35] The property of a polygon having a certain number of sides may seem to be discrete, but being three-sided as opposed to the property of being four-sided means having three sides of non-zero length. A quadrilateral can have one of its sides of length arbitrarily close to zero and so token #t^r^i^a^n^g^l^e#.

continuous range.[36]

The conclusion that, in general, non-Xs can possess every property in virtue of which Xs cause #X#s or possess properties that approximate them arbitrarily closely is contrary to assumption A1, which is used in arguing that C2 fails. It follows that in certain counterfactual situations C2 does not fail. Suppose that #d^o^g# is reliably tokened by dogs, foxes at night, and cardboard dogs. In order to solve the disjunction problem, Fodor's conditions must show that #d^o^g# means the property of being a dog and not the property of being a dog or the property of being a fox at night or the property of being a cardboard dog. This can be done by showing that the nomic relations involving properties in the disjunction other than the property of being a dog depend asymmetrically on the nomic relation involving the property of being a dog. However, the asymmetry holds only if there is no collection of non-dogs that possess or have properties approximating arbitrarily closely every property in virtue of which dogs cause #d^o^g#s. But in general, every property in virtue of which Xs cause #X#s *is* a property that some non-X can possess or approximate arbitrarily closely. So for some X, for each property in virtue of which Xs cause #X#s, we can choose a non-X that possesses that property or one arbitrarily close to it. Let Y be the disjunction of those non-Xs. By the choice of Y, breaking the Y → #X# connection will also break the X → #X# connection, contrary to the failure of C2 required by Fodor, since breaking the Y

[36] Notice that this need not be the case for every X. The point is simply that, *in general*, non-Xs can possess (or approximate) the properties in virtue of which Xs cause #X#s to which Fodor's theory can appeal. A few exceptional cases does not thereby provide Fodor with a general theory of content, not even given his agenda of providing a possibility proof, because the exceptional cases cannot generalize.

→ #X# connection breaks every nomic relation between properties in virtue of which Xs cause #X#s, or properties arbitrarily close to them, and the property of being a cause of #X# tokenings. The X → #X# law and the Y → #X# law depend on each other symmetrically, so Fodor's conditions cannot exclude the property of being a Y from the meaning of #X#. Thus, in general, Fodor's conditions do not solve the disjunction problem.

In order to block the argument I have just presented, Fodor needs a way of precluding the property of being a Y, some disjunction of non-Xs, from the meaning of #X#. The amended conditions presented in the previous section will not do so, because they determine that the meaning of a symbol is not a disjunctive property only in cases of asymmetric dependence; yet the Y → #X# law and the X → #X# law depend on each other symmetrically. Nonetheless, there is still a difference in these cases that Fodor could try to exploit. Typically, individual Xs possess all of the properties in virtue of which Xs cause #X#s. This is not true of individual Ys. What Fodor requires then, is a condition like the following. Ceteris paribus, anything in the extension of #X# has *all* of the properties that are nomically related to the property of being a cause of #X# tokenings. The ceteris paribus clause is required because not all Xs possess every property in virtue of which Xs cause #X#s. It might be possible for Fodor to produce such a condition, but the constraint on him doing so is that the ceteris paribus clause must be stated in naturalistic terms. It is not clear that this task is significantly easier than the original task of producing a naturalized theory of content because it requires Fodor to state naturalistically what, "all else being equal", it is for something to be an

X.

The argument I have just presented requires constructing a disjunction of non-Xs that have, or arbitrarily closely approximate, every property in virtue of which Xs cause #X#s. Consider a similar case in which J_1s and J_2s reliably cause #J# tokenings, the J_1 → #J# law depends symmetrically on the J_2 → #J# law, and neither J_1s nor J_2s are disjunctions of non-Js. Furthermore, suppose that #J_1#s and #J_2#s have never been tokened because no property distinguishing J_1s and J_2s has ever been detected, though such a property exists. What is the meaning of #J# in this case? Clearly, #J# does not mean either the property of being a J_1 or the property of being a J_2, since both J_1s and J_2s are in the extension of #J#. But #J# does not mean the disjunctive property of being a J_1 or J_2 either, since #J#'s expressing a disjunctive concept presupposes the concepts of the disjuncts, which by supposition nothing has. In fact, it is not possible to express what #J# means in terms of the properties of being a J_1 or J_2. #J# means the property of being a J, where Js include J_1s and J_2s, but not disjunctively. J_1s and J_2s are not discriminated; they are lumped together in a single class, Js. So, Js are more appropriately a union of J_1s and J_2s.[37]

What is interesting about this example is what we say about the meaning of #J# when properties distinguishing J_1s from J_2s *are* detected, thereby tokening #J_1# and #J_2#.[38] Again, #J# does not mean either the property of being a J_1 or the property of

---

[37] Fodor makes this point in response to Baker's objection (Fodor 1990, pp.103-6).

[38] Of course, this is not a purely hypothetical example. Properties were detected that distinguished jadeite from nephrite, both of which previously had been thought to be the one substance jade.

being a J_2, since J_1s and J_2s both have all of the properties in virtue of which #J# was tokened. The $J_1 \rightarrow$ #J# law and the $J_2 \rightarrow$ #J# law depend on each other symmetrically. Nor does #J# mean the disjunctive property of being a J_1 or J_2, but not because nothing has the concepts of the disjuncts. Once #J_1# and #J_2# are tokened, organisms have the concepts of the disjuncts. The problem is that the properties in virtue of which the symbol meaning the disjunctive property, #$J_1 \vee J_2$#, is tokened are a proper superset of the properties in virtue of which #J# was tokened. In particular, whatever property it is that distinguishes J_1s from J_2s is a property in virtue of which #$J_1 \vee J_2$# is tokened, and this property is not one in virtue of which #J# could be tokened. Furthermore, #J# cannot mean the property of being a J, where Js are a kind of union of J_1s and J_2s, since there is no such property from the point of view of an organism that can distinguish J_1s from J_2s (Fodor 1990, p.105). So #J# is meaningless for an organism that can distinguish J_1s from J_2s. But in virtue of what does #J# become meaningless when a property distinguishing J_1s from J_2s is detected? What has changed is that in virtue of an additional property, J_1s and J_2s cause new tokenings of neural types previously not tokened. Thus, the nomic relations between properties of J_1s and J_2s in virtue of which #J#s were tokened and the property of being a cause of #J# tokenings depend on what other symbols, qua syntactic items, are tokened. So, the meaning of #J# depends at least in part on what other symbols are tokened, which is not an atomistic account of content, and hence not a naturalized account.

Another serious problem that the jade example suggests is that semantic notions are required for determining the content of symbols in the language of thought. The

reason is that what nomic relations properties of objects enter into seems to depend on the content of symbols in the language of thought. Fodor has taken the nomic relations that properties of objects enter into as fixed and considered relations between these nomic relations. However, that there are such nomic relations at all seems to be semantically determined. Certain properties of J_1 and J_2 cease to be nomically related to the property of being a cause of #J# tokenings once a nomic relation is established between some other properties of J_1 and J_2 and the property of being a cause of $\#J_1\#$ and $\#J_2\#$ tokenings, respectively. "By her present lights, ..., *there is no such property* [as #J#]" (Fodor 1990, p.105, emphasis in original). However, the expression, "By her present lights," is inherently a part of the intentional vocabulary. Fodor requires a naturalistic account of the determination of nomic relations between properties of objects and the property of being a cause of a tokening of some symbol in the language of thought, and given this example it is far from clear that he can provide one. However, without such an account Fodor is left on the horns of a dilemma. He could introduce semantic notions, thereby violating the naturalistic constraint, or give up on semantics in favour of eliminativism. I take it that Fodor would not readily endorse either option. Nonetheless, from the jade example it is not clear that Fodor's account can satisfy the naturalistic constraint, which motivates my second objection.

2.12 Second Objection

I now argue that the counterfactual condition C1 does not hold because assumption A2 on which it is based is false. Recall that C1 states that if the nomic relations subsumed by the X → #X# law did not obtain, then there would not be a non-X

- #X# law either; the argument for this position rests on assumption A2, which states that non-Xs do not possess properties in virtue of which they cause #X#s that Xs do not possess or have properties that approximate them arbitrarily closely. Block's objection challenges C1 but does not do so successfully because it considers possible nomic relations that properties of non-Xs could enter into rather than the actual nomic relations those properties are in. By introducing a ceteris paribus clause into the asymmetric dependence condition, Fodor can successfully argue that the only relevant nomic relations are the actual ones. What Block does not challenge is Fodor's intuition that non-Xs cause #X#s by being mistaken for Xs. This is one way that non-Xs token #X#, and it leads to the disjunction problem. Such tokenings are false tokenings of #X#. But not all tokenings of #X# by non-Xs are false tokenings, and the properties in virtue of which non-Xs *non-falsely* token #X# do not have to be properties in virtue of which Xs cause #X#s or arbitrarily close to properties in virtue of which Xs cause #X#s. The reason is that there are semantic relations between symbols in the language of thought.[39]

Semantic relations between symbols in the language of thought abound. The particular relations I am concerned with are those that hold between symbols because items in the extension of one symbol can reliably token another symbol. Now as Chomsky (1959) points out, in the right context any symbol can token any other symbol, but most of these are not reliable tokenings. The relations I am concerned with

---

[39] Loewer and Rey make this point, though they do not develop it (Loewer and Rey, eds. 1991, p.xxxvi, endnote 55).

are those that obtain because of how the world is. That is, because items in the extensions of certain symbols happen to be reliably associated with each other in the world, the symbols are associated. I will refer to these as semantic associations. So for example, there are semantic associations between "leash" and "dog", "chair" and "table", and "fish" and "water".[40] The importance of these associations is that things in the extension of one symbol can reliably token a semantically associated symbol, *non-falsely*. For example, a leash can cause "dog" tokenings, not in virtue of any property dogs have or a property arbitrarily close to a property dogs have, but in virtue of properties that leashes have. Leashes do not cause tokenings of "dog" by being mistaken for dogs; they do so by being correctly identified as leashes, which, being items in the extension of "leash", which is semantically associated with "dog", can result in dog-thoughts.[41] "[A]ssociation is ... supposed reliably to preserve semantic domains: *Jack*-thoughts cause *Jill*-thoughts, *salt*-thoughts cause *pepper*-thoughts, *red*-thoughts cause *green*-thoughts" (Fodor 1998, p.10, emphasis in original). In particular, semantic associations figure in rational explanations. The real world association of items in the extensions of semantically related symbols, ground the rationality of the tokening of one symbol resulting from the tokening of the other. The reliable association of leashes and dogs in the world, grounds the rationality of leash-thoughts causing dog-thoughts. Now, of course, there is some dependence of leashes' tokenings

---

[40] I am using the double quote notation to emphasize that these are symbols with semantic content.

[41] Adams and Aizawa (1994) make essentially the same point in discussing pathological cases of Fodor's theory, though they do not see the problem as being quite as pervasive as I do.

of "dog" on dogs' tokenings of "dog", but it is not the asymmetrical dependence that Fodor presents. Clearly if "dog" did not mean the property of being a dog, then leashes would not token it. However, Fodor's asymmetric dependence condition is supposed to *show* that #d^o^g# means the property of being a dog by showing that all non-dogs that reliably token #d^o^g# depend asymmetrically on dogs' tokenings of #d^o^g#. Break the dog → #d^o^g# connection and all non-dog → #d^o^g# connections are supposed to be broken, according to the asymmetric dependence condition. This simply is not the case. Imagine a counterfactual situation in which a deadly virus swiftly wiped out the canine population. Once people learned that all of the dogs were dead, properties in virtue of which dogs cause #d^o^g# tokenings would cease to do so; i.e. something else with some of those same properties, such as a fox, would not reliably be mistaken for a dog. Nonetheless, leashes would still cause #d^o^g# tokenings, perhaps by reminding people of the lost species. Break the dog → #d^o^g# connection and the leash → #d^o^g# connection remains intact, because the properties in virtue of which leashes cause #d^o^g#s are not those in virtue of which dogs cause #d^o^g#s or arbitrarily close to those in virtue of which dogs cause #d^o^g#s. The relation between leashes and the symbol #d^o^g# rests on a semantic association between "leash" and "dog". You only break the leash → #d^o^g# connection by supposing that #d^o^g# does not mean the property of being a dog; and supposing that in virtue of breaking the dog → #d^o^g# connection, #d^o^g# does not mean the property of being a dog is assuming the result Fodor needs to demonstrate.

A response Fodor might try to offer in defense of his position is that, while it is

true that leashes cause #d^o^g# tokenings, they do not do so reliably; i.e., there is no nomic relation connecting properties of leashes to the property of being a cause of #d^o^g# tokenings. However, this is a response that *Fodor*, in particular, cannot offer. Fodor's main project is to defend intentional explanation. Clearly, under certain conditions, all else being equal, *thoughts* about leashes do reliably cause *thoughts* about dogs; e.g. "What kind of animal do people regularly walk on leashes in many parks in North America?" Fodor simply cannot deny that there are intentional regularities between tokenings of symbols in the language of thought, if he wants to maintain that psychology is a special science. That semantic associations can figure in rational explanations depends on such regularities. But then, since there are properties of leashes in virtue of which they cause #l^e^a^s^h# tokenings, and there are some circumstances under which tokenings of the type #l^e^a^s^h# reliably cause tokenings of the type #d^o^g#, there are nomic relations between properties of leashes and the property of being a cause of #d^o^g# tokenings. Moreover, the properties of leashes that are nomically related to the property of being a cause of #d^o^g# tokenings are properties in virtue of which leashes cause #l^e^a^s^h# tokenings; in general, such properties are not, and do not, approximate properties in virtue of which dogs cause #d^o^g# tokenings, so they do not depend asymmetrically on the dog → #d^o^g# connection.

There was nothing particular to the relation between "leash" and "dog" that allowed me to construct my argument that the leash → #d^o^g# connection does not depend asymmetrically on the dog → #d^o^g# connection, except that "leash" and "dog" are semantically associated. The same argument will work for any pair of semantically

associated symbols in the language of thought, such as "fish and "water", and "chair" and "table". Because of the abundance of semantic associations between symbols in the language of thought, it will not, in general, be true that non-Xs do not possess properties in virtue of which they cause #X#s that Xs do not possess or have properties that approximate them arbitrarily closely; i.e. assumption A2 is false. It follows that the counterfactual condition C1, that if the nomic relations subsumed by the X → #X# law did not obtain, then there would not be a non-X → #X# law either, is false. Since the asymmetric dependence condition holds only if C1 is true, the asymmetric dependence condition, on which rests Fodor's solution to the disjunction problem, is not satisfied by any symbol in the language of thought that is semantically associated to any other symbol in the language of thought. Fodor's goal of defending intentional explanation suggests that *very many* of the symbols in the language of thought *are* semantically associated with some other symbol in the language of thought. Thus, Fodor has not provided a solution to the disjunction problem. Furthermore, the relevance of semantic associations between the symbols in the language of thought in undermining Fodor's attempt to provide a naturalistic semantics has important implications for Fodor's anti-bootstrapping argument that the language of thought is not a natural language.

CHAPTER 3

INFORMATION AND ASYMMETRIC DEPENDENCE

3.1 Introduction

In the previous chapter, I argued that Fodor's theory of content is unable to solve the disjunction problem, by focusing on specific details of Fodor's theory. Those considerations led me to the conclusion that Fodor's theory fails if there are semantic associations between symbols, something Fodor's project of trying to secure the place of intentional explanation commits him to. In this chapter, I explore the failure of Fodor's asymmetric dependence condition because of the semantic associations between symbols, in order to draw out two general points:

(i) The content that we can ascribe to our internal representational types solely in virtue of their syntax does not correspond to the content of semantic types that figure in intentional explanations;

(ii) The relations that obtain between internal representations solely in virtue of their syntax do not correspond to the semantic associations between symbols that figure in rational explanations.

I begin by presenting Dretske's notion of information. I then consider Fodor's distinction between information and meaning. Fodor's discussion of information is somewhat confused because Fodor takes causation to be sufficient for the transmission of information, which Dretske explicitly denies. Because of this confusion there are several possible readings of Fodor's statement of the asymmetric dependence condition in terms of information. I consider these possibilities given that causation is not sufficient for the transmission of information, from which I offer a reading of the

asymmetric dependence condition in terms of information that is consistent with Fodor's other presentations of the asymmetric dependence condition. I then show why, in terms of information, the condition fails. Also by considering asymmetric dependence in terms of information, I argue that Fodor's position results in a phenomenalist semantics. (i) follows from showing in terms of information how the asymmetric dependence condition fails. I then use (i) to establish (ii), which I use in my response to Fodor's anti-bootstrapping argument that the language of thought is not a natural language, presented in the next chapter.

3.2 Information and Meaning

"Errors raise the disjunction problem, but the disjunction problem isn't really, deep down, a problem about error. What the disjunction problem is really about deep down is the difference between *meaning* and *information*" (Fodor 1990, p.90, emphasis in original). Dretske states that "one is free to think about information (though not meaning) as an objective commodity, something whose generation, transmission, and reception do not require or in any way presuppose interpretive processes" (Dretske 1981, preface, p.vii). As an "objective commodity", information is always veridical; there is no mis-information. "What information a signal carries[1] is what it is capable of "telling" us, telling us *truly*, about another state of affairs. Roughly speaking, information is that commodity capable of yielding knowledge, and what information a signal carries is what we can learn from it" (Dretske 1981, p.44, emphasis in original).

[1] Signals are said to have a meaning and to carry information (Dretske 1981, p.44).

Information is carried by a signal from a source to a receiver. Being an "objective commodity", the information that a source can generate is information about the properties of that source. Communication theory quantifies how much information is carried by a signal.[2] Not all of the information generated at a source is transmitted in a given signal. For example, information about the temperature of an object is not transmitted by the visible light reflected[3] from it. The (amount of) information that is not transmitted is called the equivocation. Thus, the amount of information received from a source is the amount of information generated at the source minus the equivocation. The amount of information received from a source can also be conceived of as a portion of the total information available at the receiver. The total information available at the receiver consists of the information carried by a signal from the source in question and other information carried by signals from other sources. The (amount of) information from other sources is called noise. So, the amount of information received from a given source is the total information available at the receiver minus the noise. Since the amount of information generated at the source need not equal the amount of information available at the receiver, the equivocation need not, and in general will not, equal the noise.

Now it is important to distinguish between causation and the transmission of

---

[2] The precise mathematical formulations for quantifying the amount of information carried by a signal are not needed for my considerations, so I do not present them here. They are presented in Dretske 1981, with references to the mathematical literature on which they are based.

[3] If the object is hot enough, then the light it radiates carries information about its temperature, but this is not reflected light.

information. Very often information is transmitted from a source to a receiver because of a causal relation between them. For example, someone can receive information about a dog when she hears barking because the barking causes disturbances in the air which reach her. This need not be the case, however. Two televisions tuned to the same channel carry information about each other despite there being no causal interaction between them. So causation is not necessary for the transmission of information. Perhaps more surprisingly, causation is not sufficient for the transmission of information either. If an event at a receiver can have more than one cause, then the information carried by the event is not information about the particular cause. So if someone near a harbour that is frequented by seals hears barking that could have been produced by either a dog or a seal, she does not receive information about the source being a dog, *even if it was, in fact, a dog that caused the sound*. Notice that though the receiver does not receive information about the source being a dog, she does receive information about a sound; i.e. that there is a barking. But even this need not be the case. If the receiver is highly medicated and very likely to hallucinate, then even a genuine sensory experience will not carry information. When there is ambiguity as to the nature of the source, the equivocation is non-zero. "One can have full information without causality, and one can have no information with causality. And there is every shade of gray between these two extremes" (Dretske 1981, p.33).

Now in his discussion of meaning and information, Fodor does not clearly distinguish between causation and the transmission of information. In particular, Fodor takes reliable causal covariance to be sufficient for a token of a syntactic type to carry

information about a source. As Fodor puts it, "all you need for information is reliable causal covariance, whereas for meaning you need (at least) asymmetric dependence too" (Fodor 1990, p.93). But, as we have just seen, reliable causal covariance is not sufficient for the transmission of information. "It is sometimes carelessly assumed, for example, because the reflection of light from an object *causes* certain events to occur on the periphery of our visual receptors, and these events (the proximal stimulus) in turn cause other events to occur in the central nervous system, ultimately yielding some response from the organism, that therefore the subject has, in some vague sense of "information," received information about the distal stimulus (the object from which the light was reflected). This simply does not follow" (Dretske 1981, p.33). Dretske's point is that the actual causal situation is not enough to determine what information is transmitted, if any. Other possible causes are also relevant in determining what information is transmitted.

In supposing that "all you need for information is reliable causal covariance, whereas for meaning you need (at least) asymmetric dependence too" (Fodor 1990, p.93), Fodor has simply failed to respect the difference between causation and the transmission of information. Fodor is supposing that if there is reliable causal covariance between a source and tokenings of a syntactic type, then tokens of that syntactic type carry information about the source. Now it is possible that tokens of the syntactic type do carry information about the source if there is reliable causal covariance between them, but they need not carry that information. It depends on what else causes their tokenings. And in the situations that lead to the disjunction problem,

tokenings of a syntactic type can be caused by different kinds of things, which is precisely the problem. So in the situations that lead to the disjunction problem, the tokens do not carry information about what kind of thing the source is. If a situation is such that a tokening of a syntactic type could have been caused by either a dog or a fox, then the token does not carry information about the source being a dog, even if the source is a dog. What the source does carry information about is some shared property of dogs and foxes. We can see now in terms of information just why the disjunction problem arises. Items that are not in the extension of a symbol type can token the symbol because they can generate information that could have been, and sometimes is, generated by items that are in the extension of the symbol. For example, a fox can generate information about an animal having a certain shape, just as a dog can, in virtue of which the fox can token "dog". The "dog" token carries information about an animal having a certain shape, information that could have been generated by a dog, in some conditions, because of the similarities between the shapes of dogs and foxes.

Because symbol tokens can carry information about properties of items that are not in the extension of the symbol, the meaning of a symbol is not given by the information its tokens carry. Meaning also requires asymmetric dependence, according to Fodor; however, Fodor's own presentation of the asymmetric dependence condition in terms of information not only conflates causation with the transmission of information, it does not respect the distinction that he draws between meaning and information. By clearing up this confusion it is possible to state the asymmetric dependence condition in terms of information in a way faithful to Fodor's intentions,

but this statement lays bare the reason that the asymmetric dependence condition fails if there are semantic associations between internal representations.

3.3 The Asymmetric Dependence Condition in Terms of Information

Fodor distinguishes information from meaning as follows:

> Information is tied to etiology in a way that meaning isn't. If the tokens of a symbol have two kinds of etiologies, it follows that there are two kinds of information that tokens of that symbol carry. (If some "cow" tokens are caused by cows and some "cow" tokens aren't, then it follows that some "cow" tokens carry information about cows and some "cow" tokens don't). By contrast, *the meaning of a symbol is one of the things that all of its tokens have in common, however, they may happen to be caused. All* "cow" tokens mean *cow*; if they didn't, they wouldn't be "cow" tokens (Fodor 1990, p.90, emphasis in original).

Notice here again that Fodor fails to distinguish information and causation. If the tokens of a symbol have two kinds of etiologies, what follows is not that "there are two kinds of information that tokens of that symbol carry", but rather that tokens of the symbol carry information about some shared property of the different causes. Cow caused "cow" tokens[4] do not carry information about cows in situations where the disjunction problem arises; they carry information about some shared property of cows and the other possible causes of the "cow" token. Similarly, for non-cow caused "cow" tokens. But despite conflating causation and the transmission of information,[5] Fodor's basic

---

[4] For clarity of exposition, I am following Fodor's terminology that tokens are caused and carry information. Of course, if causation is taken as a relation between events, then technically tokenings, and not tokens, are caused. However, since it is tokens that have meaning, I take it that Fodor uses this terminology to avoid such locutions as "the token whose tokening was cow caused...". Dretske seems to allow that both states and events carry information.

[5] It is not clear that Fodor has corrected this error. See Fodor 1998, p.12.

distinction between meaning and information is correct. Meaning is etiologically robust in a way that information *about the properties that symbols mean* is not. All "cow" tokens mean the property of being a cow, regardless of how they are caused; but no non-cow caused "cow" tokens carry information about the property of being a cow, and many cow caused "cow" tokens do not carry information about the property of being a cow either. However, only one page after stating the difference between meaning and information, Fodor states the asymmetric dependence condition in terms of information thus: "'Cow' means *cow* because *but that 'cow' tokens carry information about cows, they wouldn't carry information about anything*" (Fodor 1990, p.91, emphasis in the original). In the statement of the asymmetric dependence condition, Fodor seems to be implying that every "cow" token carries information about cows, in direct contrast with the conclusion reached in distinguishing meaning from information that not all "cow" tokens carry information about cows.[6] Fodor's use of the word "wouldn't" in stating asymmetric dependence suggests that it is only in a counterfactual situation that "cow" tokens do not carry information about cows. It is not that *when* "cow" tokens do not carry information about cows they do not carry any information, according to Fodor; it is that *if* "cow" tokens did not carry information about cows they "*wouldn't*" carry any information. The implication is that in the actual world all "cow" tokens do carry information: information about cows, which is false.

---

[6] Fodor seems to be confusing the objective notion of information developed by Dretske with a subjective notion of "information for the system". In the subjective sense a "cow" token carries information about cows "for the system" because any behaviour that "cow" tokens elicit will be intentionally characterizable as some sort of "cow" behaviour.

Perhaps Fodor was just speaking loosely in stating the asymmetric dependence condition. Suppose that Fodor really meant "don't" rather than "wouldn't" in stating asymmetric dependence, so that "cow" tokens carry information about cows or they do not carry any information. But again, this claim is just false. "Cow" tokens do carry information when they do not carry information about cows. In the situations in which the disjunction problem arises, "cow" tokens carry information about some shared property of cows and other possible causes of the "cow" tokens, though they do not carry information about cows.

A third, more charitable reading of what Fodor intends in the asymmetric dependence condition is that if some "cow" tokens did not carry information about cows, no "cow" tokens would carry information about anything. This reading of asymmetric dependence respects the robustness necessary for meaning; not all "cow" tokens carry information about cows, because some are reliably caused by non-cows; nonetheless, all "cow" tokens *mean* the property of being a cow. Furthermore, this reading is consistent with the statement of asymmetric dependence we considered in the previous chapter. If a Y → #X# law depends asymmetrically on the X → #X# law, then it is only in virtue of properties that Xs possess or properties arbitrarily close to properties Xs possess that Ys cause #X#s. But then, if a Y causes an #X# tokening, any information that #X# token carries is information about a property that Xs possess or about a property arbitrarily close to a property Xs possess. In either case, if some #X# tokens did not carry information about Xs, the #X# token would not carry the information generated by Y. So if some #X# tokens did not carry information about Xs,

no #X# tokens would carry information about anything. Thus, this reading of asymmetric dependence in terms of information renders a consistent account of what Fodor is offering.

On this reading of asymmetric dependence, we see clearly that false tokenings of a symbol occur when the source of a signal is not in the extension of the symbol, but the source generates information that could have been, and sometimes is, generated by something in the extension of that symbol, information about a shared property of the source and items in the extension of the symbol. If "cow" tokens could carry information that could not be generated by cows, then "cow" tokens could carry information even though no "cow" tokens carried information about cows, contrary to the asymmetric dependence claim. So it follows on this reading of the asymmetric dependence condition that any information that a "cow" token can carry is information that could be generated by a cow.[7] Thus, a general necessary condition for asymmetric dependence on this reading is that #X# tokens carry only information that could be generated by Xs. However, by considering this condition, we can see just why the asymmetric dependence condition fails if there are semantic associations between

---

[7] Notice that on this reading Fodor is committed to a distinction between types of situations for determining meaning, a position he criticizes in Fodor 1990, chapter 3. The special type of situations on this reading of asymmetric dependence are those in which "cow" tokens carry information about cows. It is in virtue of these situations that "cow" means the property of being a cow. However, these situations only occur when the disjunction problem does not arise, so to specify them Fodor must already be able to specify when the disjunction problem does not arise. But if he could do so, he could use that account as a solution to the disjunction problem, making asymmetric dependence redundant. Furthermore, the burden is on Fodor to specify the relevant situations naturalistically, which is a large burden given that what information is transmitted depends on the state and circumstances of the receiver.

symbols.

3.4 The Failure of the Asymmetric Dependence Condition in Terms of Information

A necessary condition for asymmetric dependence is that #X# tokens carry only information that could be generated by Xs; however, as I will now argue, this condition fails to be satisfied, from which it follows that the asymmetric dependence condition also fails. Any item in the extension of a symbol #Y#, such that #Y# is semantically associated with #X# can reliably cause #X# tokenings. The information that the #X# token carries in such a case need not be, and in general will not be, information about a shared property between Xs and Ys. Ys do not falsely token #X# by being mistaken for an X, they veridically token #X# in virtue of a semantic association between #X# and #Y#. Just what information an #X# token carries when it is caused by a Y such that #Y# is semantically associated with #X# depends on the situation, of course, and there will be some situations in which it carries no information at all. But there will also be some situations in which #X# carries information about some property of Ys that Xs do not possess or approximate arbitrarily closely.[8] So some #X# tokens will carry information that could not have been generated by Xs, contrary to the necessary condition for asymmetric dependence. For example, when a leash causes a "dog" token, the token can carry information about the leash, some property of the leash, or perhaps no information. The point is simply that there are some cases in which a "dog" token carries information that could not have been generated by a dog, contrary to the

[8] Note that I am not assuming that #X# carries information about Ys, though it might, only that it carries information about some property of Ys because it also might be possible in the circumstances for non-Ys having those properties to cause an #X# token.

necessary condition for asymmetric dependence, from which it follows that the asymmetric dependence condition fails.

3.5 Phenomenalist Semantics[9]

As we saw above (3.2), causation is neither necessary nor sufficient for the transmission of information. What information is transmitted depends on what else could have reliably covaried with the current state of a receiver. Let us suppose that dogs cause "dog" tokens. Furthermore, let us suppose we have a situation in which a "dog" token would not reliably causally covary with anything but dogs, so that the token does indeed carry information about dogs. Nonetheless, dogs do not cause "dog"s directly. They do so in virtue of their detectable properties. So at least one detectable property of dogs covaries with "dog" in any situation in which dogs covary with "dog"s, though, of course, not the same detectable property in every situation. So every "dog" token that carries information about dogs must also carry information about some detectable property of dogs. Furthermore, since it is in virtue of their detectable properties that dogs cause "dog"s, if it were not the case that "dog" tokens carried information about the detectable properties of dogs, "dog" tokens would not carry information about anything. In particular, if "dog" tokens did not carry information about detectable properties of dogs, they would not carry information about dogs. But then on Fodor's account of meaning, the meaning of "dog" is the detectable properties

---

[9] My thanks to Rob Stainton for helpful discussions of this issue.

of dogs and not the property of being a dog, which is a phenomenalist semantics.[10]

3.6 The Syntax of Internal Representations

Because of the naturalistic constraint, Fodor must regard internal representations purely syntactically, as we saw in Block's objection. By definition, syntactic types are individuated according to syntactic features; that is, tokens of a syntactic type must all share some syntactic feature. The failure of the asymmetric dependence condition occurs because tokens of a single syntactic type can carry information about items in the extensions of distinct intentional types. In particular, it can be the case that information generated by an item in the extension of one intentional type cannot be generated by an item in the extension of another intentional type, yet tokens of a single syntactic type can carry information about items in the extensions of both intentional types. Thus, the information that can be carried by tokens of a syntactic type could not be generated by items in the extension of a single intentional type. So for a syntactic type #X#, there cannot be a law 'Xs cause #X#s' on which any law 'Ys cause #X#s', where Ys are non-Xs, asymmetrically depends, such that the Xs are in the extension of a *single* intentional type. Hence, the content that we can ascribe to our internal representational types in virtue of their syntax is too coarse-grained to correspond to the content of semantic types that figure in intentional explanations. For example, tokens of a syntactic type #X# can carry information about properties of leashes and properties of dogs. There is

---

[10] I take it that Sellars (1956/1997) showed that this position is untenable. His reasoning is that meanings cannot be constructed out of sense data as primitive given elements, for sense data themselves get their content because of their place in a semantic structure, a structure that could be created during (natural) language acquisition.

no one intentional type such that the items in its extension can generate all of the information carried by #X# tokens, so the content of #X# tokens, which on Fodor's view is determined by the information they carry, cannot correspond to the content of our intuitive semantic types.

It is worth emphasizing that it is because the syntax of internal representations cannot discriminate between the information generated by items in the extensions of distinct intentional types that the content we can ascribe to internal representational types in virtue of their syntax is too coarse-grained to correspond to the content of semantic types. Now it might be argued that there *is* a syntactic difference between, say, a leash-caused #d^o^g# tokening, and a dog-caused #d^o^g# tokening. Leashes cause #d^o^g# tokenings by tokening #l^e^a^s^h#, which then causes the tokening of #d^o^g#; tokenings of a syntactic type corresponding to a semantic type mediate the tokening of #d^o^g# in a leash-caused #d^o^g# tokening. Dog-caused #d^o^g# tokenings, on the other hand, are not mediated by tokenings of any syntactic type corresponding to a semantic type. Perhaps Fodor's theory of content can be salvaged by stipulating as one of the conditions for meaning that the only semantically relevant reliable causal covariances are unmediated. This would exclude items in the extensions of symbols semantically associated with a given syntactic type from being included in the extension of that syntactic type; and asymmetric dependence could deal with the error cases.

The difficulty with this suggestion is that it presupposes a syntactic difference between an innate stock of syntactic types that correspond to semantic types and other

syntactic types. When a dog causes a #d^o^g# tokening, it does so in virtue of information transmitted via a signal from the dog to a receiver; and part of the transmission of that signal occurs over tokens of neural types that causally covary with dogs. But since the covariance of tokenings of neural types with dogs is in virtue of syntactic features of those neural types, those syntactic features determine syntactic types that mediate the dog's causing of #d^o^g# tokenings. *In the same way*, #l^e^a^s^h# tokenings mediate leashes' causings of #d^o^g# tokenings. Furthermore, it cannot be supposed that some brute neurophysiological fact distinguishes a syntactic type whose tokenings mediate a dog's causing of #d^o^g# tokens from the syntactic type #l^e^a^s^h#, a fact such as tokens of the former occur only in the peripheral sensory organs and tokens of the later occur only in the central nervous system.[11] The reason is that often the equivocation--the amount of information generated at a source that is not transmitted in a signal--is quite high, such as when a dog is viewed in failing light conditions. In such cases, considerable activity in the central nervous system might mediate the tokening of #d^o^g#, and again, tokenings of some of those more central neural states will reliably covary with dogs in virtue of their syntax, thereby determining a syntactic type. From a purely syntactic point of view, there is no distinction between the mediation of a dog caused #d^o^g# by tokenings of some syntactic type, and the mediation of a leash caused #d^o^g# by tokenings of the syntactic type #l^e^a^s^h#. Only by presupposing that the syntactic type #l^e^a^s^h# is

---

[11] Of course, I am assuming token physicalism in this discussion.

also a semantic type can a distinction be drawn; but nothing grounds that presupposition.

Another possibility is that we have ignored some of the syntactic structure of internal representations in arguing that their content, given by the information they carry, does not correspond to the content of semantic types that figure in intentional explanations. In particular, we have only considered the syntactic structure involved in the causings of tokenings of a syntactic type, and not the structure involved in what tokenings of that type can cause. Yet, "it's plausible that at least some mental objects are distinguished by the kinds of mental processes that they *cause*; i.e. they are functionally distinguished" (Fodor 1998, p.19, my emphasis). But then Fodor is back on the horns of a familiar dilemma. How are we to characterize what it is that tokenings of a syntactic type cause? By referring to what is caused as "mental processes", Fodor seems to imply that we should use an intentional characterization;[12] indeed, by doing so we can ascribe content to syntactic types in correspondence with the content of our semantic types; but we violate the naturalistic constraint. However, by using a non-intentional characterization, the content that can be ascribed does not correspond to the content of our semantic types. The reason is that the only things that a tokening of a syntactic type #X# can cause by which content could be ascribed to #X# are other syntactic types to which content is ascribed, or overt bodily movements. In the case that a tokening of #X# causes a tokening of some syntactic type #Y# to which content can

[12] Notice that even if we take mental processes to be computational processes, what distinguishes a computational process from a "brute incursion from the physiological level" is that computations are performed on symbol tokens and have a semantic interpretation.

be ascribed, since the content that can be ascribed to #Y# is as coarse-grained as the content that can be ascribed to #X#, it cannot help determine a more fine-grained ascription of the content of #X#. In the other case, that #X# tokenings cause some overt bodily movement, #X# tokenings can cause *any* overt bodily movement.[13] The tokening of any symbol can cause any overt bodily movement in the right context. For example, for any proposition P, if you believe that P then raise your right hand, or shake your left leg, or point to something yellow, or shout "yahoo", or recite the first line of your favourite poem, or hum your favourite tune, or stand on your head, etc.. Unless the bodily movement is intentionally type-individuated, #X# tokenings' causings of that movement do not aid in ascribing content to #X#; but, if the bodily movement is intentionally type individuated, the ascription of content is not naturalistic.

In summary, because tokens of a single syntactic type are also tokens of symbols in the language of thought, according to Fodor, they can carry information generated by items in the extensions of distinct intentional types[14]. It follows that the content that can be ascribed to our internal representations, in virtue of the syntax that Fodor takes as determining meaning, does not correspond to the content of our intuitive semantic types. By considering associations between syntactic types in addition to those that

---

[13] Notice that it is not open to Fodor to appeal to reliable causal covariance between #X# tokenings and overt bodily movements because that would be to endorse some version of behaviourism.

[14] In particular, tokens of a single syntactic type can carry information generated by items in the extensions of distinct intentional types, such that the items in the extension of one intentional type could not generate information generated by items in the extension of the other intentional type.

Fodor takes as determining meaning, we discovered that we can only ascribe more fine-grained content to syntactic types by appealing to semantic notions. Hence, the content that we can ascribe to our internal representational types in virtue of their syntax does not correspond to the content of semantic types that figure in intentional explanations.

3.7 The Semantics of Internal Representations

It now follows immediately that the semantic associations between symbols that figure in rational explanation cannot be ascribed in virtue of the syntax of the internal representations. The semantic associations between symbols are determined in part by the semantic contents of symbols; semantic associations occur because of the way the world is, but which symbols are associated depends on the semantic content of those symbols. Therefore, what semantic associations obtain between internal representations qua syntactic types will be a function of the content that can be ascribed to those syntactic types, *solely in virtue of their syntax*. Yet we have just seen that the content that we can ascribe to our internal representational types in virtue of their syntax does not correspond to the content of the semantic types that figure in intentional explanation. Hence, the relations between internal representations that obtain solely in virtue of their syntax cannot correspond to the semantic associations between symbols.[15]

Recall that rational thought for Fodor is a sequence of transformations of tokens of symbols, sensitive solely to the syntax of those tokens that, nonetheless, has an

---

[15] It should not come as a great surprise that under the constraints Fodor imposes on his theory of content, relations between internal representations that correspond to semantic associations between symbols do not obtain, because the theory was designed to allow the possibility of punctate minds (Fodor 1990, p.51). A consequence of Fodor's atomism is that a mind that has many internal representations can be nothing more than a union of punctate minds.

intentional interpretation as steps in an inference. That the transformations constitute rational thinking at all depends essentially on their having an intentional interpretation as steps in an inference. But an intentional interpretation of the transformations depends in part on the semantic associations between the symbols being transformed. Without semantic associations between symbols, some tokenings of symbols by transforming other symbols would simply be "brute incursions from the physiological level" (Fodor 1975, p.200). Now we have just seen that the relations between internal representations that obtain solely in virtue of the syntax of those internal representations do not correspond to the semantic associations between symbols. Thus, it cannot be solely in virtue of the syntax of internal representations that the transformations of tokens of symbols have an intentional interpretation on Fodor's view of content. Hence, Fodor does not have an account of how thought is rational solely in virtue of the syntactic structure of our internal representational system.

CHAPTER 4

HOW THE LANGUAGE OF THOUGHT CAN BE A NATURAL LANGUAGE

4.1 Introduction

As we saw in chapter 1, Fodor has three arguments for why the language of thought cannot be a natural language. Briefly, they are:

(1) *Animals Think*: human infants and some animals think, though they are not natural language-users;

(2) *Innateness*: to learn the first predicates of any language requires being able to use predicates in some other representational system, coextensive with the predicates of the language being learned; hence there must be at least one representational system that we can use without learning, i.e. it must be innate; since all natural languages are learned, the representational system in terms of which natural language predicates are learned must be distinct from any natural language;

(3) *Anti-Bootstrapping*:[1] "What the child cannot do, in short, is use the fragment of the language that he knows to increase the expressive power of the concepts at his disposal" (Fodor 1975, p.84); it is not possible to use some part of a natural language in order to learn a more expressive part of the language, by bootstrapping into it, because the structure of truth rules requires that the predicates, in terms of which a new predicate is learned, are used; it follows that no predicate in the language being learned *essentially* mediates the learning of any new predicate; i.e., every predicate that is learned must be

[1] (3) really has the form of a reply to a counter-suggestion; however, since the counter-suggestion is quite a natural one to make, I present (3) as a separate argument.

coextensive with a predicate in the internal representational system; hence, the internal representational system is at least as expressive as any natural language.

In chapter 1, I considered the first two arguments in some detail and argued that it is compatible with these arguments that we possess an internal representational system that we use in early language acquisition, but once we have learned a portion of a natural language, we are able to use that learned portion to learn the rest of the natural language, a more expressive portion. Clearly the anti-bootstrapping argument is intended to block exactly the kind of move I am trying to make. The aim of this chapter is to show that the language of thought can be a natural language, by showing that the anti-bootstrapping argument is unsound.

First, I show how the reasoning for the anti-bootstrapping argument, if sound, would entail that the internal representational system is the medium of our thinking, the language of thought; that is, the transformations on internal representations have an intentional interpretation that grounds the rationality of thought independent of natural language acquisition. Indeed, Fodor's position is that the intentionality of natural languages is derived from the intentionality of the internal representational system, and for this reason the internal representational system is the medium of thought. I then argue that the anti-bootstrapping argument *cannot* be sound. I argue by reductio ad absurdum that the conclusion of the anti-bootstrapping argument contradicts the conclusion from the previous chapter (3.7), that the relations that obtain in virtue of the syntax of the internal representations do not correspond to the semantic associations between symbols that figure in rational explanation. In explaining the contradiction, I

suggest that it might be possible to ascribe content to the internal representations that corresponds to the content of natural language terms, provided there are no syntactic relations between symbols in the internal representational system corresponding to semantic associations between terms of natural languages; but in that case the semantic associations between terms of a natural language would not correspond to any relations between internal representations, hence the rationality of certain transformations in a natural language could not be derived from the intentionality of the internal representational system. Having established that the anti-bootstrapping argument cannot be sound, I consider why it is not sound. I argue that the anti-bootstrapping argument requires the assumption that all of a natural language must be learnable in terms of the internal representational system, which is false. Fodor's argument, that every predicate of a natural language must be learnable in terms of the internal representational system, presupposes that natural language predicates cannot be constructed by combining known predicates, using parts of a natural language not learned in terms of the internal representational system. However, the logical connectives of a natural language *need not* be learned in terms of the internal system; Fodor himself proposes a use-theory of meaning for the logical connectives (Fodor 1990, pp.110-111). It follows that natural languages can be more expressive than the internal representational system, if the internal system need not possess anything corresponding to the logical connectives. And if natural languages can be more expressive than the internal representational system then the intentionality of natural languages need not be derived from the intentionality of the internal representational system; in particular, the intentionality of a portion of a

natural language that the internal system could not express could not be derived from the intentionality of the internal system. So the internal representational system need not be the medium of thinking. Furthermore, the internal system need not be productive or systematic[2], i.e. it need not be linguistic, since Fodor's only argument that the internal system is linguistic follows directly from the conclusion of the anti-bootstrapping argument that the internal representational system must be at least as expressive as any natural language.

4.2 What the Anti-bootstrapping Argument Says

It is important to be clear exactly how the anti-bootstrapping argument is supposed to show that a natural language cannot be the language of thought. In itself, what the anti-bootstrapping argument claims is that the internal representational system is at least as expressive as any natural language. The anti-bootstrapping argument simply says that we cannot learn a predicate that is not expressible in terms of representations we know, which, *together with* the innateness argument, shows that prior to acquiring a natural language, we possess an internal representational system at least as powerful as any natural language. Since learning is one kind of rational thinking, and languages are learned in terms of the internal representational system, the internal system is the medium of our thinking. That is, the transformations of internal representations have an intentional interpretation, independent of natural language

---

[2] Notice that I am not claiming that without logical connectives the internal representational system could not be linguistic, only that it need not be. In particular, in chapter 6, I argue that in lacking anything corresponding to the logical connectives, we need not suppose that the internal representational systems of animals are systematic to explain their behaviour.

acquisition. Now since natural languages are learned in terms of the internal representational system, Fodor is assuming that the intentionality of natural languages is derived from the structure of the internal representational system.[3] "[It is not clear] what could make *language itself* systematic if not the systematicity of the thoughts that it is used to express... On balance, I think we had better take it for granted, and as part of what is not negotiable, that systematicity and productivity are grounded in the 'architecture' of mental representation and not in the vagaries of experience" (Fodor 1998, pp.26-7, emphasis in original). Thus, a natural language expresses our thoughts, but cannot be the medium of thinking, the language of thought. Notice that the internal representational system is the medium of thought, even if by learning a natural language we come to perform some, or indeed all, of the transformations constituting thinking over symbols in that natural language.[4] That transformations on symbols of a natural language *have* an intentional interpretation, and thereby constitute thinking, is derived from the intentional interpretation of transformations of internal representations

---

[3] Supposing the intentional structure of natural languages and the structure of the internal representational system are independent, yet in direct correspondence, seems tantamount to supposing a miracle.

[4] We might come to perform some or all of the transformations that constitute thinking over symbols in a natural language by using the natural language as a mnemonic device, in which natural language expressions abbreviate more complicated formulae of the internal representational system. Such a mnemonic device could be advantageous in two ways: "The most obvious possibility is [for a child] to use [the fragment of a natural language that she knows] for mnemonic purposes [to master the rest of that natural language]" (Fodor 1975, p.84). The other possibility is that natural language expressions could also allow us to think thoughts we could not otherwise think, because such thoughts would require processing internal formulae we cannot entertain. "If terms of the natural language can become incorporated into the computational system by something like a process of abbreviatory definition, then it is quite conceivable that learning a natural language may increase the complexity of the thoughts that we can think" (Fodor 1975, p.85).

coextensive with the natural language symbols being transformed.

It is worth noting how strongly Fodor is committed to the naturalistic constraint in this line of reasoning. Since it is compatible with Fodor's reasoning that all[5] of the transformations that we actually perform in thinking occur over terms of our natural languages, it is only in virtue of being the source from which the intentionality of natural languages is derived that the internal representational system constitutes the language of thought, i.e. the medium of our thinking. Thus, it is essential for Fodor that the intentionality of the internal representational system is not itself derived, but is "grounded in the 'architecture' of mental representation". That is, Fodor must offer a naturalistic account of how the intentionality of the internal representational system obtains, in order for the anti-bootstrapping argument to serve as an argument that the language of thought cannot be a natural language.

4.3 Argument that the Anti-bootstrapping Argument Cannot Be Sound

Let us assume that the reasoning in the anti-bootstrapping argument is sound, forcing the conclusion that the internal representational system is at least as expressive as any natural language. Now consider any natural language terms, such as "dog" and "leash" that are semantically associated. By supposition, the internal representational system contains terms corresponding to "dog" and "leash". Now the semantic association between the natural language terms means that there is a rational explanation for the use of one term in a response to some use of the other. These very same

---

[5] That is, in using natural language for mnemonic purposes nothing in what Fodor says precludes the possibility that all of the transformations we actually perform occur over terms of the natural language.

statements must be expressible in the internal representational system. But as we saw above (4.2), if the reasoning of the anti-bootstrapping argument is sound, the transformations of internal representations have an intentional interpretation that is not derived from any interpretation of transformations of natural language terms. So there must be some relation between the internal representations corresponding to "dog" and "leash", such that tokening one of the representations as a result of tokening the other has a rational explanation. Since there is nothing particular to the example of "dog" and "leash", there must be such relations between the internal representations corresponding to any semantically associated terms in natural languages. But given Fodor's token physicalism, the intentional relations that hold between the internal representations obtain solely in virtue of the syntax of the internal representations. Thus, some of the relations between internal representations that obtain solely in virtue of their syntax must correspond to the semantic associations that hold between symbols, which directly contradicts our result in the previous chapter that the relations between internal representations that obtain solely in virtue of their syntax do not correspond to the semantic associations between symbols. Thus, the anti-bootstrapping argument cannot be a sound argument. The internal representational system need not be as expressive as any natural language.

The contradiction just derived serves to suggest that the rationality of thought is not constituted solely of syntactic relations. However, a contradiction from the anti-bootstrapping argument can be produced much more directly. The conclusion of the anti-bootstrapping argument is that the internal representational system must be at least

as expressive as any natural language. Now, according to Fodor, the content of the internal representations is determined by the causal relations they enter into with things in the world, as determined solely by their syntax. But as we have seen above (3.6), the content that can be ascribed to internal representations solely in virtue of their syntax does not correspond to the content of our intuitive semantic types. It follows that the internal representational system cannot express any of the predicates corresponding to our intuitive semantic types, and so is not as expressive as natural language. Now the crucial point that emerges from this discussion is that carriers of only information that could be generated by items in the extension of some intuitive semantic type cannot be tokens of that symbol type. The tension arises because tokens must play a certain functional role relative to tokens of other types to preserve psychological explanation; but in playing a certain functional role, tokens of one type can be caused by tokens of a second type and so carry information generated by the items that token the second type, which need not be items in the extension of the first type. The carriers of only information that can be generated by items in the extension of some semantic type cannot play the functional role of tokens of that type. This tension cuts very deeply into Fodor's project. Fodor could respond to the argument that I have just presented by allowing that the theory of content is inadequate, but nonetheless, there is nothing wrong with the arguments for the language of thought. After all, the contradiction I presented assumed Fodor's theory of content. Since I have already provided independent grounds for rejecting it, invoking it against the language of thought is not of itself a serious blow to the language of thought hypothesis. But as we see now, all

that is required from Fodor's theory of content to generate the contradiction with the conclusion of the anti-bootstrapping argument is that the content of a symbol is determined by the information its tokens carry, assuming token physicalism. So while it is still possible to avoid the contradiction, it comes with a high price. Fodor can give up the idea that meaning is information, something he is not presently inclined to do (Fodor 1998, p.12). The reason Fodor wants to keep this assumption is that he sees it as the only possible way of satisfying the naturalistic constraint (Fodor 1990, chapter 3). The other possibility is to give up token physicalism, which can be done in two ways. Accept eliminativism thereby giving up the overall project of preserving psychological explanation; or, accept some kind of interpretationist position. As I said, this tension cuts deeply. Counterintuitively, if we suppose that *there are no semantic associations between the internal representations*[6], then something like Fodor's theory of content[7] might be able to ascribe content to the internal representations that corresponds to the content of natural language terms. The reason is that the carriers of information would not need to play a functional role that required them to carry more information than that by which content corresponding to our intuitive semantic types could be ascribed to them; I explore this possibility in chapter 5.

---

[6] Of course, this is not an assumption that anyone holding that the internal representational system is the language of thought can make. To make this assumption, one must be prepared to accept that the semantic associations between natural language terms are not derived from the structure of the internal representational system.

[7] Not exactly Fodor's theory if we are to avoid the problem of having to give naturalistic conditions for what it is for something to be an X, all else being equal. See (2.11).

4.4 Why the Anti-bootstrapping Argument is Unsound

We have just seen that the anti-bootstrapping argument is unsound, given the assumptions that meaning somehow reduces to information and token physicalism. Now since neither of these assumptions is obviously false, perhaps we can find some independent argument that the anti-bootstrapping argument is unsound. In this section, I show that the reason the anti-bootstrapping argument is unsound is that it is based on a false premise. Fodor argues that *every* predicate in a natural language must be coextensive with a predicate in the internal representational system. The reason he offers is that every predicate of a natural language must be learned, which requires (at least) confirming some truth rule. In any truth rule, the predicate on the right-hand side--the predicate in terms of which the natural language predicate is being learned--must be used. The very first predicates we learn must be learned in terms of the internal representational system, since we cannot use any other predicates coextensive with the ones we are trying to learn. Fodor falsely supposes that it follows that every predicate is coextensive with a predicate in the internal representational system, by supposing that even if we learn a natural language predicate Q in terms of another predicate of natural language P, P must be coextensive with an internal representation. What Fodor does not consider is that since P need not be an atomic predicate of the natural language, P might be constructed out of predicates in a natural language coextensive with internal representations, using some other components of the natural language that *do not* correspond to anything in the internal representational system. That is, natural languages might have compositional resources that are not capturable in the internal

representational system. For instance, in a way that is not learning in Fodor's sense of learning, i.e. via truth rules, we might acquire certain terms of natural language with which we can construct new predicates from those we already have. Acquiring these terms would increase the expressive power of the portion of a natural language that we know beyond the expressive power of the internal representational system. These terms would enable a language-learner to construct new predicates that she could not construct from the internal representations alone. Fodor is correct in supposing that *Q* cannot be learned in terms of a portion of a natural language that cannot express a predicate *P*, coextensive with *Q*; but, in virtue of there being components of natural languages that need not correspond to anything in the internal representation system, *P* need not correspond to anything in the internal representational system. *P* can be a predicate constructed from predicates that are coextensive with terms in the internal representational system, and so learned in Fodor's sense, by combining these predicates using a term that was acquired--i.e. *not* learned in Fodor's sense--which, being acquired, does not have anything corresponding to it in the internal representational system. In that case there need[8] not be anything coextensive with *P*, hence *Q*, in the internal representational system. Fodor's reasoning implicitly assumes that *all* of a natural language must be learned in terms of the internal representational system, for then a language-learner could not acquire components of her language that do not

[8] There still could be something in the internal representational corresponding to *P* because the compositional resources of the internal representational system need not correspond to those of natural languages. My point is simply that there need not be anything corresponding to *P* if natural languages have ways of producing predicates that the internal representational system does not have. Recall that I am not arguing that there is no mentalese, only that there need not be.

correspond to internal representations. However, this assumption is false; the logical connectives of a natural language are not predicates and need not be learned in terms of the internal representational system via a truth rule. It follows that the anti-bootstrapping argument is unsound.

4.5 The Logical Connectives

> *What about the Logical Vocabulary?* I don't know what about the logical vocabulary. ... I'm inclined to think that maybe there is *no* objection to the idea that "+", "and", "all" and the like have the meanings they do because they play a certain causal role in the mental lives of their users. This would, of course, be to accept a distinction in kind between the logical and the nonlogical vocabularies. (The semantics for the former would be a kind of 'use' theory, whereas the semantics for the latter would depend on nomic, specifically mind-world, relations.) Gilbert Harman somewhere suggests that to be a logical word *just is* to be the sort of word of which a use-theory of meaning is true. This proposal strikes me as plausible. ... I know of no principled reason why some such proposal shouldn't be endorsed (Fodor 1990, pp.110-111, emphasis in original).

Several points need to be made about what Fodor says concerning the logical connectives. The main point is, of course, that Fodor endorses a use-theory of meaning for the connectives. This is important because, as Fodor remarks, it makes "a distinction in kind between the logical and the nonlogical vocabularies." But the distinction runs deeper. The causal theory of content that Fodor offers is a theory about the content of internal representations; in considering the logical connectives, Fodor is considering the role that certain *words* play in *a natural language*. "[A] logical *word* just is to be the sort of *word* of which a use-theory of meaning is true" (my emphasis). Now talking about words as opposed to internal representations might just be

carelessness on Fodor's part, but it is very revealing. Because the meaning of the logical connectives is different in kind from the meaning of non-logical words, the logical connectives do not have to be learned via truth rules in terms of coextensive internal representations. In fact, the logical connectives need not be *learned* at all, at least not in the sense of knowing-that; all that's required is that we master their use, a kind of knowing-how, which we can do in terms of a natural language directly. In fact, we can restate the false premise of the anti-bootstrapping argument as the assumption that every word in a natural language must be learned in the sense of knowing that. If this were the case then it would follow that the internal representational system contained something corresponding to the logical connectives. The reason is that in learning the word "and", for example, one would learn a recursive axiom of the form: A sentence of the form $\ulcorner P_x$ and $Q_y \urcorner$ is true if and only if $\ulcorner P_x \urcorner$ is true and $\ulcorner Q_y \urcorner$ is true. Now this axiom would be expressed in the internal representational system and the right-hand side uses "and", so it would follow that the internal representational system contains something corresponding to the natural language term "and", and mutatis mutandis for the other connectives. However, since the content of the logical connectives is given by a use theory, they do not have to be learned in the sense of knowing-that, only knowing-how.[9] This difference is crucial. Knowing how to use a logical connective requires only that it is used correctly, though of course not infallibly, in the appropriate (truth) conditions. It does not also require one knows *that* it is used correctly in those

---

[9] My thanks to Rob Stainton for suggesting I make the distinction between learning predicates and learning the logical connectives in this way.

conditions, which would indeed require something corresponding to the logical connectives in the internal representational system.[10] Finally, since "there is no reason at all to suppose that the logico-syntactic vocabulary is itself interdefined with the *non*-logical vocabulary" (Fodor 1994, p.76, emphasis in original), *there need not be internal representations corresponding to the logical connectives of natural languages.*[11]

Logical connectives are used in a natural language to combine elements of that natural language, when certain truth conditions obtain. Thus, to master the use of a logical connective, a child must learn to combine elements of her natural language in certain conditions. This, of course, presupposes that she has already learned some of her natural language. But learning natural language predicates might *trigger* the ability to use the logical connectives. Having learned some of the predicates of her natural language, a child might explore and manipulate the space of predicates she has learned. The tools with which the child manipulates the natural language predicates could just be the natural language words for the logical connectives. In the same way that children can come to use objects in various ways just by manipulating those objects, such as stacking blocks, they might master the use of the logical connectives by forming

---

[10] In (6.6) below, in explaining how animals could learn concepts we would express using a logical connective, such as RED AND TRIANGULAR, without requiring that the internal representational system possess anything corresponding to the logical connectives, in this case conjunction, I give a model on which one could know how to use a connective without knowing that the relevant truth conditions obtain.

[11] Again , I am not claiming that learning the logical connectives could not be a kind of learning-that, only that it need not be. That is, we need not suppose that the internal representational system possesses anything corresponding to the logical connectives to explain how natural language connectives are acquired.

combinations of predicates of their natural language, using the logical connectives. Now, it could be that we cannot master the use of a logical connective until we have learned a certain number of natural language predicates, because we simply do not have a large enough sample to manipulate the predicates in the ways that correspond to the logical connectives. It might be just a matter of brute empirical fact that systematic uses of the logical connectives require manipulating enough predicates for that use to be systematic.

It is important to note that the use of the logical connectives is systematic and productive. Terms in a system that includes "and" can be systematically conjoined ad infinitum. The point is not that something corresponding to the logical connectives is required for a system to be systematic and productive, only that a system which has the connectives is productive and systematic in its use of the connectives.

4.6 Why the Internal Representational System Need Not Be the Language of Thought

Nothing in what Fodor has argued compels us to suppose that the combinations of natural language predicates that competent use of the logical connectives enables are even possible for predicates in the internal representational system. If it is not possible to combine predicates in the internal representational system in a way corresponding to the way that predicates of a natural language are combined by logical connectives, it follows that combinations of natural language predicates formed by the logical connectives need not be coextensive with anything in the internal representational system. But then the combinatorial power of natural languages could exceed that of the internal representational system. Recall, however, that in (1.3), (1.9), we saw that

Fodor's only argument that the internal representational system is a linguistic system depended on the anti-bootstrapping argument. Only by assuming that the internal representational system is at least as expressive as any natural language could it be concluded that the internal system is productive and systematic. But, since the anti-bootstrapping argument is unsound, the internal representational system need not be as expressive as any natural language, hence it need not be linguistic.

Many of the semantic relations between elements of a linguistic system are determined by the compositionality of that linguistic system, because the meanings of complex expressions are determined by the meanings of the components of the expressions and the connectives by which those components are combined. Such semantic relations between elements of a linguistic system are essential for ascribing an intentional interpretation to transformations of symbols in that linguistic system. Those intentional interpretations underpin the rationality of thought. A crucial feature of the intentional interpretation of transformations of symbols is that the transformations are truth preserving. Thus, it is quite plausible that the *logical* connectives are essential for constructing the semantic structure comprising rationality. Also, another way transformations can be interpreted as truth preserving is by reflecting associations that obtain in the world; i.e. by transforming tokens of one symbol into tokens of semantically related symbols. So two features of a linguistic system by which the transformations of representations can have a rational interpretation are having something corresponding to the logical connectives, and having representations that are semantically associated. Now since the internal representational system need not possess

anything corresponding to the logical connectives, and since it is not clear how the internal representations could be semantically associated on Fodor's view, transformations on the internal representations need not have an intentional interpretation; the operations of the internal representation system need not be rational. But as we saw in (4.2), it is only in virtue of being the source from which the intentionality of natural languages is derived that the internal representational system must be the language of thought, i.e. the medium of thinking. It follows that if the transformations on internal representations need not have an intentional interpretation, the internal system need not be the language of thought.

4.7 How a Natural Language Could Be the Language of Thought

Since it is compatible with Fodor's reasoning that the actual transformations that constitute thinking occur over tokens of natural language terms, and since an intentional interpretation of those transformations need not be derived from an intentional interpretation of the internal system, it can be the case that once we use a natural language for thinking it is the language of thought. Specifically, the transformations over tokens of our natural language terms could constitute rational thought once we have sufficient mastery of the logical connectives, and semantic associations between the terms, so that those transformations could have an intentional interpretation. However, the rationality of thought depends on one of the logical connectives in particular. As we saw in (1.4), in order for a creature to be rational it must represent its options in a given situation, evaluate those options, and choose what it determines to be the most preferred option. "It was, for example, implicit in the model that the organism

has available means for representing not only its behavioural options but also: the probable consequence of acting on those options, a preference ordering defined over those consequences and, of course, the original situation in which it finds itself" (Fodor 1975, p.31). But to explicitly represent its hypothetical options, it must have mastery of the *conditional*. Now recall that *all* rational thought processes are processes of hypothesis formation and confirmation, according to Fodor. "[M]odels of perception have the same general structure as models of concept learning [and considered action]: One needs a canonical form for the representation of the data, one needs a source of hypotheses for the extrapolation of the data, and one needs a confirmation metric to select among the hypotheses" (Fodor 1975, p.42). Thus, it seems a creature can only perform rational cognitive processes if it has mastery of the conditional. A natural language can be the language of thought partly because mastery of a natural language entails mastery of the conditional, presupposed by (rational) thought.

4.8 But What About...?

The internal representational system need not be the medium of thinking, because it need not possess anything corresponding to the logical connectives, and it is not clear that the representations can be semantically associated; in particular, it need not possess anything corresponding to the conditional. In that case, mastery of the conditional would seem to require mastery of a natural language. But then if rational thought requires mastery of the conditional, rational thought would require mastery of a natural language. So how could it be that some animals and preverbal children perform the rational cognitive processes of considered action, perceptual integration, and

concept learning? And how is early language acquisition to be explained? Without an alternative explanation for these phenomena, Fodor could simply respond to the arguments I have presented by pointing out that the reason the language of thought cannot be a natural language is that "[r]emotely plausible theories are better than no theories at all" (Fodor 1975, p.27). My argument that the language of thought could be a natural language depends on the claim that there need not be anything in the internal representational system corresponding to the logical connectives. My position, in direct response to the anti-bootstrapping argument, is that we could master the use of the logical connectives, even if there is nothing in the internal representational system corresponding to the logical connectives. However, the competencies of preverbal children and some animals, together with the result that rational thought seems to require mastery of the conditional, suggests that the internal representational system *must* possess something corresponding to the conditional (and the other logical connectives). In order to complete my argument that the language of thought could be a natural language, I must offer an explanation of animal and preverbal human behaviour that does not require the internal representational system to possess anything corresponding to the logical connectives. Whether or not animals think is an open question, one that Fodor begs in supposing that animals do think. What we do know is that animals demonstrate behaviour homogeneous to our own, and it is this behaviour that I must explain. That explanation comprises chapter 6. In the next chapter, I present an alternative approach to content on which my explanation of animal and preverbal human behaviour is based. Also, I have used the fact that there are semantic

associations between terms in a natural language to argue that Fodor's theory of content is unsatisfactory. Since it is unclear how internal representations with content corresponding to our intuitive semantic types could be semantically associated, it seems that any rational explanation of human verbal behaviour based on semantic associations between terms of a natural language is not derived from the intentional structure of the internal representational system, in which case the internal system need not be the medium of thinking. Thus, the fact that there are semantic associations between terms is working quite hard in my argument. It would be satisfying, therefore, if my account of content gave some indication as to how semantic associations are formed between terms of a natural language. I suggest how this might occur at the end of chapter 6.

CHAPTER 5

AN ALTERNATIVE APPROACH TO CONTENT

5.1 Introduction

As I concluded in the previous chapter, to complete my argument that the internal representational system need not be linguistic, I must offer an explanation of animal and preverbal human behaviour that does not require the internal representational system to possess anything corresponding to the logical connectives. The required explanation of animal and preverbal human behaviour that I present in chapter 6 is based on a general approach to giving a theory of content, what I call the embodied approach to content, different from the kind of theory Fodor offers. Part of the motivation for adopting this approach comes from seeing where Fodor's theory fails. That the semantic associations between symbols undermine Fodor's attempt to provide a naturalized semantics supports the standard assumptions concerning intentionality: "(*i*) the intentional/semantical predicates form a closed circle and (*ii*) intentional states are intrinsically holistic" (Fodor 1990, p.51), which Fodor rejects. It seems that securing intentional explanation does not lie in there being a naturalized semantics. My analysis of Fodor's theory of content, in the previous three chapters, continually reveals that the source of the failure of Fodor's theory is his attempt to impose the naturalistic constraint on a defense of intentional explanation. But Fodor's dilemma need not be ours. We can simply abandon the naturalistic constraint and accept the standard assumptions concerning intentionality. Thus, I take these standard assumptions concerning intentionality to be assumptions of the general approach to

content that I develop in this chapter.

The embodied approach to content that I am developing is motivated by the notion of embodied cognition, developed by Andy Clark (1997).[1] The general approach is to give a dispositional account of content, so it falls in the tradition beginning with Gilbert Ryle, through David Armstrong, Robert Stalnaker, and Daniel Dennett. Since I am only making the case that the internal representational system *need not* be linguistic and *need not* be the medium of thought, all I require of the embodied approach to content that I am developing is that it be a viable philosophical position. There may be independent philosophical worries, such as mental causation, that bear on the details of which fully developed alternative to Fodor's theory can be sustained, but those details are incidental to my argument.

I begin this chapter with a presentation of the notion of embodied cognition. I then argue that the competencies that creatures having concepts display require that they possess neural structures--not necessarily type individuated by neuroscience--that covary with the environment in a way similar to that suggested by Fodor for symbols in the language of thought qua syntactic structures, though these neural structures need not have content. Given these neural structures, I consider how concepts could be learned on the model of embodied cognition, and why behavioural responses to stimuli in the extension of a concept are generally appropriate when such stimuli are present. I then

---

[1] I am not suggesting that this position is entirely due to Clark, simply that I am basing my discussion on his presentation. Embodied cognition is a general approach in robotics, which is an alternative to more traditional centralized processing approaches to artificial intelligence. Clark (1997) presents many systems having this basic design. Rodney Brooks of the *M.I.T. Artificial Intelligence Laboratory*, for example, uses this approach (Brooks 1991).

characterize what it is to have a concept, on the embodied approach to content, as having a behavioural disposition that can figure in an intentional explanation of the creature, such that the explanation appeals to the concept. I evaluate this account of concepts according to five conditions that Fodor deems an acceptable theory of concepts would have to meet, and show that the embodied approach to content meets these conditions. I conclude by considering what the structure of the internal representation system is, according to the embodied approach to content.

5.2 Embodied Cognition

"The notion of internal representation still plays a key role, but the image of such representations is undergoing some fundamental alterations... [The] combination of decentralization, recurrence, ecological sensitivity, and distributed multidimensional representation constitutes an image of the representing brain that is far, far removed from the old idea of a single, symbolic inner code (or "Language of Thought")" (Clark, 1997, pp.141-2).

The intuition on which embodied cognition is based is that a primary function of a brain is to move the body of the organism it inhabits about the environment it finds itself in, in a way that promotes the survival and fitness of the organism. Solutions to problems are rough and dirty, exploiting any aspects of the local environment and features of the creature's body that can reduce computations a brain needs to perform. The key notion of embodied cognition is that of decentralization. Unlike traditional approaches to cognition that were used in early artificial intelligence research, embodied cognition does not suppose that there is a central processing system required

to produce intelligent behaviour.[2] Instead, intelligent behaviour emerges from the interactions of a network of simple mechanistic procedures that are influenced only by local factors, i.e. proximal stimuli. The idea is simply that ordered structure can be created without anything in the process having a representation of the structure being created. An example Clark[3] offers to illustrate the difference between traditional approaches to cognition and embodied cognition is the following. The task is to decide on the optimum placement of footpaths between newly constructed buildings, on a new university campus. One approach is for a single individual to consider the entire layout of the campus, what each building is for, how many people are likely to use each building, and any other number of factors, to determine some optimal pattern. This approach is that of having a central processor represent the entire situation in order to determine a solution. Another solution is to put grass between all of the buildings and open the university. Individuals will be forced to determine their routes based on their specific needs. Over time, paths will emerge in the grass, and since people have a tendency to follow emerging paths, the required network will be established. No one person is required to produce a global representation of the situation in order for a solution to emerge. Individuals make decisions based on their immediate needs. Notice also that the solution depends on the local environment, the scheduling of classes, the layout of the buildings, etc., yet none of these things needs to be globally represented in

---

[2] Intelligent behaviour is behaviour that can figure in intentional explanations, i.e. explanations in terms of beliefs and desires.

[3] He attributes the example to Aaron Sloman.

producing a solution; no one has the big picture, and no resources are expending in producing the big picture (Clark 1997, p.79). The position of embodied cognition is that brains function in the latter way in this example to solve real-world problems for organisms in specific environments.

One advantage of an embodied cognition design is that the system is more robust to local damage than a central processing unit. Very local damage to a central processing unit can disable the entire system. However, a system with an embodied cognition design does not have a central processing area. Functions are determined locally, so that other parts of the system can adapt their activities in accordance with any damage that has occurred, by simply responding to proximal stimuli. Most of the system is still active when local damage occurs and can compensate for the damaged area to produce an appropriate behaviour under the circumstances. Clark presents an example due to Pattie Maes[4] of the M.I.T. Media Laboratory that nicely demonstrates how local damage is not as critical in an embodied cognition design, as well as making clear the difference between a classical central processing design and an embodied cognition design. The task is to design a system that can schedule, in the most efficient way, several different machines to perform a variety of jobs as they arise. The difficulty in the task is that new jobs are always arising and the load on any one machine is constantly changing. A centralized approach to this problem would require one system to represent the states of all of the machines and the job load. It would have

[4] Maes 1994, pp.145-6.

to frequently update this information, which it would use to assign jobs as they arose. Notice that if the central system were damaged it would disable the entire process. On the decentralized approach suggested by Maes, each machine controls its own workload. When a machine creates a job, it requests bids to do the job from all of the other machines. Each machine's bid is simply its estimate of when it could accomplish the task, given its current workload. The machine that can complete the job mostly quickly is assigned that job. Thus, machines with relatively small workloads typically take on the new jobs. If one machine is disabled, the remaining machines can continue the task. Scheduling is an emergent property that does not require a representation of the entire situation (Clark 1997, pp.43-44).

Another advantage of an embodied cognition design is that it avoids the bottleneck of centralized processing. There are real-world time constraints on creatures' problem solving procedures. Often a creature does not have the luxury of forming a complete representation of its situation in order to determine its action. For example, if someone throws a stick at us, we typically do not have the time to say to ourselves, "Someone just threw a stick at me. If I remain where I am it will hit me. If it hits me it will hurt. I do not wish to be hurt. I should move." If we do form all of those explicit representations, by saying the sentences in our head, for example, we will be hit by the stick. Embodied cognition rejects the notion that we passively receive information from the world, which is transmitted to a central processor that represents and integrates all

incoming information, from which an action is determined.[5] "[R]eal-time, real-world success is no respecter of this neat tripartite division of labour. Instead, perception is itself tangled up with specific possibilities of action--so tangled up, in fact, that the job of central cognition often ceases to exist. The internal representations the mind uses to guide actions may thus be best understood as action-and-context-specific control structures rather than as passive recapitulations of external reality" (Clark 1997, p.51). Rather than filtering all of the information received by the sensory organs through a single processing unit, local neural structures interact with each other to initiate actions. Nothing in a creature's nervous system need correspond to an integrated representation of the information it receives in order for an appropriate behaviour--i.e. advantageous to the creature--to emerge from the local interactions of neural structures.

It is important to emphasize that an embodied cognition design exploits the physical attributes of a creature and the environment in which it is situated, not by representing them and determining how best to use them, but by their very interactions. By changing its position or altering its environment a creature can simplify its task. For example, an animal that is foraging might push plants aside, or move rocks to give itself easier access to its food. On the model of embodied cognition, a creature does not require a central processor with a representation of the entire environment in order for the creature to perform these actions. We do not represent the detailed shape of a piece of a jigsaw puzzle to determine where it fits. We pick the piece up, turn it and try

---

[5] Notice that the conception of a central processor is very much like the notion of a Cartesian Theater that Dennett attacked (Dennett 1991, chapter 5).

placing it until we put it where it does fit.[6] It can be in the same way that a bear rearranges its environment by breaking open a log to eat the termites inside. "[The] world can provide an arena in which special classes of external operations systematically transform the problems posed to individual brains" (Clark 1997, p.66). Of course, the physical attributes of the creature determine in what way it can interact with its environment. Aardvarks do not forage in the same way as bears, but they do so to the same effect. The brain, body, and world form an interactive system, such that a body's transformations of the world reduce the computational load on the brain during problem solving. However, features of the environment also dictate what actions creatures with particular physical attributes can take, thereby constraining the behavioural options of a creature, which in effect, partially determines behaviour. The path a human takes up a steep hill might be different from the path a mountain goat takes, in part because of the physical terrain of the hill. A person who would generally follow a route of switchbacks might avoid the extra distance and go up directly where there are protruding branches she can use to pull herself up. The fortuitous presence of branches provides an opportunity for a novel solution as to how to traverse the terrain. "[Systems whose control structures are decentralized] create actions from an 'equal partners' approach in which the local environment plays a large role in selecting behaviours. In situations where a more classical, inner-model-driven solution would break down as a result of the model's incapacity to reflect some novel environmental

---

[6] This example is from Clark 1997, pp.36, 63-4.

change, 'equal partners' solutions often are able to cope because the environment itself helps to orchestrate the behavior" (Clark 1997, p.43).

5.3 Neural Structures Implicated in Concept Learning

The general approach to content that I am developing assumes the model of embodied cognition. I now argue that there are certain neural structures implicated in concept learning. In this section I characterize those neural structures, and in the next section I describe their role in concept learning, in order to lay the foundation for providing an alternative approach to content based on the model of embodied cognition. Fodor describes concept learning in an experimental situation as follows:

> In the typical experimental situation, the subject (human or infrahuman) is faced with the task of determining the environmental conditions under which a designated response is appropriate, and learning is manifested by *S*'s increasing tendency, over time or trials, to produce the designated response when, and only when, those conditions obtain. The logic of the experimental paradigm requires, first, that there be an 'error signal' (e.g., reinforcement or punishment or both) which indicates whether the designated response has been appropriately performed and, second, that there be some 'critical property' of the experimentally manipulated stimuli such that the character of the error signal is a function of the occurrence of the designated response together with the presence or absence of that property (Fodor 1975, p.35).

Notice that, though a creature manifests that it has learned a concept by an increasing tendency to produce a designated response "when, and only when" certain conditions obtain, this does not entail that a creature will reliably produce the response when, and only when, the relevant conditions obtain, *in general*; i.e. outside of the experimental situation. Systematic errors are always a possibility, but in an experimental situation, the conditions can be adequately controlled so that a creature *can* reliably discriminate a

particular set of conditions from the range of conditions with which it is presented, in the experiment--assuming it can learn the concept. It is also worth reiterating that on Fodor's own account, a creature learns just when to produce a designated response by being given an error signal, the character of which (reinforcement or punishment) is determined by which stimuli do, or do not, prompt the response. I highlight this point to emphasize that holding a position in which learning a concept is achieved through some reward and punishment mechanism that results in a response behaviour being produced reliably when, and only when, certain stimuli are present does not commit one to behaviourism. The alternative approach to content I am developing is based on Fodor's characterization of concept learning, but it does not deny that internal states are relevant for determining content, and so is not behaviouristic.

Take Fodor's account of concept learning as given. Learning a concept is manifested by a creature's increasing tendency to produce a designated response when, and only when, certain environmental conditions obtain.[7] Now in order for a creature to reliably produce a response when, and only when, certain conditions obtain, it must be sensitive to those conditions. It must have some neural structure(s)[8] that reliably

---

[7] Even in an non-experimental situation, a creature manifests having a concept by being able to reliably respond to stimuli in the extension of the concept, though in a non-experimental situation the response is not designated. Of course, not any reliable response to items in the extension of some concept manifests having a concept, in a non-experimental situation; otherwise, sneezing in the presence of ragweed would manifest having the *concept* RAGWEED. I address this issue below (5.5).

[8] Since this is a discussion about animal and human cognition, I will refer only to neural structures, however, if robotics projects prove successful, the discussion could be generalized by talking about internal structures of systems.

causally covaries with those conditions, i.e. a structure that is reliably activated when, and only when, those conditions obtain.[9] Now since by supposition, nothing else in the situation reliably covaries with the neural structures, these neural structures carry information about, and only about, the conditions they reliably causally covary with. But in situations where having a concept is manifested, those conditions just are the presence of stimuli that are in the extension of the concept. Hence, from the nature of concept learning, we can conclude that for any concept a creature can learn, it must possess some neural structure that can carry only information that could have been generated by items in the extension of that concept.[10] Notice that the claim is that the information could be generated by *all* items in the extension of the concept[11], since in manifesting a concept, creatures tend to reliably respond to any item in the extension of the concept, regardless of whether they have encountered it previously. What this entails is that items in the extension of a concept have some detectable property or properties--Fodor's "critical property"--to which creatures are sensitive in virtue of having neural structures which are activated by signals carrying information about the

---

[9] This neural structure need not be a well-defined neural type in that it might be impossible to individuate the relevant neural activity. The point is simply that there is something about the structure of the nervous system that allows it to respond discriminately to environmental conditions. For this to be possible, whatever is reliably caused to occur in a creature's nervous system by distinct conditions, which the creature can discriminate, must be different. I express this by saying that the creature has a neural structure that causally covaries with some environmental condition.

[10] Millikan (1998) makes much the same point.

[11] Technically, *ceteris paribus*, the information could be generated by all items in the extension of the concept. The CP clause is required for degenerate cases, such as severely mutated specimens, which, because of the mutation, do not have detectable properties that other items in the extension of the concept share.

property or properties. For example, dogs have a certain odour, and a characteristic shape, particularly the shape of their muzzles, to which creatures can be sensitive.

Now the reader will recall that in (3.3) I presented a necessary condition for the asymmetric dependence condition in terms of information very similar to the claim above. Fodor takes the activation of certain neural structures to be tokenings of symbols in the language of thought. The necessary condition for asymmetric dependence in terms of information is that #X# tokens carry only information that could be generated by Xs. It would follow from the asymmetric dependence condition that there are neural structures that can carry only information that could be generated by items in the extension of some symbol, as I am asserting. However, the reason the asymmetric dependence condition does not hold is that the semantic associations between symbols in the language of thought require that the neural structures would have to carry more than just information that could have been generated by items in the extension of a particular symbol. That is, because the activation of a certain neural structure is a tokening of a symbol in the language of thought, on Fodor's view, and because symbols in the language of thought are semantically associated, that neural structure cannot carry only information that could be generated by items in the extension of a particular symbol. We can avoid the fatal problem for the asymmetric dependence condition that neural structures must carry more information than the theory allows them to carry, if we do not suppose that the causal interactions between neural structures correspond to the semantic associations between symbols in the language of thought. But, it only follows that the causal interactions of neural structures do correspond to the semantic

associations between symbols in the language of thought, if the activations of certain neural structures are tokenings of symbols in the language of thought; and, unlike Fodor, we need not suppose that the activations of certain neural structures are tokenings of symbols in the language of thought. Henceforth, I take it as an assumption of the approach to content that I am developing that the neural structures implicated in concept learning are not related in a way that corresponds to the semantic associations between symbols in a language of thought.

The assumption that the neural structures implicated in concept learning are not related in a way that corresponds to the semantic associations between symbols in a language of thought is compatible with two general positions: either the neural structures have no intentional content, in which case they are not in any semantic associations; or, they can be ascribed content, but are not related to each other in virtue of their content. For both of these general positions, it follows from the foregoing assumption that, though two concepts are semantically associated, neural structures that can carry only information generated by items in the extensions of those concepts are not causally related. For example, a neural structure that can carry only information that could be generated by dogs does not reliably activate a neural structure that can carry only information that could be generated by leashes, despite the semantic association between DOG and LEASH. I discuss this consequence below (5.6) and (6.8). For my purpose of developing an alternative approach to content, it is not necessary to choose between the positions that neural structures have no content or that neural structures have content but are not related to each other in virtue of that content;

thus, I propose to leave both possibilities open. Issues concerning mental causation will figure largely in determining between these positions; below (5.5) I mention some of these issues, but a full discussion of mental causation is beyond the scope of this project.

Notice that supposing a neural structure can carry only information that could have been generated by Xs does not beg any questions on this account. Even if there were no semantic associations between symbols, Fodor would still have a disjunction problem in supposing that activating a neural structure that can carry only information that could have been generated by Xs is tokening a symbol "X", because of the error cases. Recall from (2.11) that to solve this problem, Fodor requires a *naturalistic* account of what it is for something to be an X, all else being equal. All the alternative embodied approach requires is that, as a matter of fact, some neural structure can carry only information--an objective commodity--that could have been generated by an X. Of course, there are many ways of characterizing this information, but since the embodied approach to content is not subject to the naturalistic constraint, we can characterize the information *intentionally* as information that could have been generated by Xs, where Xs constitute the extension of some concept "X", without begging any questions.

5.4 Concept Learning

Concept learning is manifested by an increasing tendency to produce some response when, and only when, a stimulus is in the extension of the concept being

learned.[12] Now if a creature is truly in a concept learning situation, it cannot have another behaviour that it generally produces when, and only when, it encounters a stimulus in the extension of the concept to be learned.[13] If it could perform such a behaviour, than it would already have the concept, as evidenced by its reliable tendency to produce that behaviour when, and only when, stimuli in the extension of the concept are present. In that case, learning some new designated response to stimuli in the extension of the concept would be learning a new behaviour relative to the concept and not learning a new concept.[14] So a creature enters a concept learning situation without being able to reliably produce any behaviour in the presence of stimuli in the extension of the concept being learned. But, as we saw above (5.3), for any concept that a creature can learn, it must possess a neural structure that can carry only information that could be generated by items in the extension of that concept. Hence, at the beginning of a concept learning situation *no* behaviour is reliably initiated by a neural structure that can carry only information that could be generated by items in the extension of the concept being learned.

In an experimental learning situation, once a concept C has been learned, a

---

[12] Again, not just any response is sufficient for manifesting a concept, as I discuss in the next section (5.5).

[13] I will use the terms "behaviour" and "response" interchangeably, and, following Dretske, take them to be processes that end in bodily motions (Dretske 1988). My arguments could be recast taking behaviours as inner causes of bodily motions, as Hornsby does, without substantially changing them, but my view requires that behaviour not be just bodily motions, for it must be possible to initiate a behaviour that is not manifested.

[14] Fodor claims that on the radical behaviourist account, concept learning just is learning to produce the designated response (Fodor 1975, p.35, footnote 6).

creature can reliably respond to stimuli in the extension of C. That is, it can reliably produce a response when, and only when, presented with stimuli in the extension of C. In non-experimental situations creatures do not always reliably produce a response when, and only when, they are presented with stimuli in the extension of some concept. When the conditions are not controlled, creatures are always subject to making systematic errors, since there are certain conditions in which the information they receive from a source is information that could have been generated by something in the extension of some particular concept, even though the source is not itself in the extension of that concept. Nonetheless, in supposing that a creature has a concept, it is supposed that a creature can distinguish items in the extension of that concept from other items, *in at least some conditions*, even though it cannot make such a discrimination in all conditions. Thus, even in the non-experimental situation a creature *can* reliably produce a response when, and only when, presented with stimuli in the extension of a concept it has learned. Now since it produces the response discriminately, it must be in virtue of being sensitive to properties of the stimuli. But a creature is sensitive to properties of stimuli in the extension of a concept C because it has a neural structure that can carry only information about those properties. The presence of stimuli in the extension of C activates the neural structure. Hence, in the causal chain from stimulus to response the activation of the neural structure must initiate the creature's behaviour. However, at the beginning of the concept learning situation, the neural structure does not initiate any behaviour. Thus, the first phase of a concept learning process is to associate a behavioural response with the neural structure,

so that the activation of the neural structure will initiate that behavioural response.

Associations between a neural structure and a behavioural response can be made by rewarding the behaviour when it is performed in the presence of stimuli in the extension of a concept C, and not rewarding it, or punishing it, otherwise--i.e. by providing "an error signal". Actually getting the creature to perform the response when stimuli in the extension of C are present can be achieved by presenting the stimuli when some independent factor that produces the response is also present. The creature initially performs the response for independent reasons, but soon also associates it with stimuli in the extension of C.[15] The result is that creatures come to produce the response when stimuli in the extension of C are present without the independent factor. This does not yet constitute concept learning, however, because it might be that several neural structures are associated with the behaviour, because several neural structures could be activated by some initial set of stimuli presented.[16] Thus, to complete learning the concept, a creature must disassociate all neural structures with the behaviour, except for the structure which can carry only information that could be generated by items in the extension of the concept. The error signals--reward and punishment--given over an increasingly large number of trials effect the disconfirmation, at which point the concept is learned.

We can now see why, in general, behavioural responses are appropriate to

---

[15] In fact, the association may not require a reward, since non-punishment and repetition can serve as reinforcement, though a reward might accelerate the formation of the association.

[16] This must be the case if the concept learning situation is one of hypothesis formation and confirmation.

stimuli in the extension of some concept. In an experimental situation, the response can be any natural behaviour of the learner, but in the context of the experiment it is an appropriate behaviour given the stimuli, because it is reinforced by the error signal rewarding the creature when, and only when, those stimuli are in the extension of the concept. In a natural (non-experimental) concept learning situation, a behavioural response can be associated with the neural structures activated by a stimulus by imitating the behaviour of a conspecific, usually a parent. In this case, the behaviour of the parent is an independent factor that produces a response in the offspring because it imitates the parent. But the parent's behaviour in the presence of some stimulus will be some behaviour it has learned that is rewarded in the presence of that stimulus--the parent has already learned the concept, in the sense that it can reliably produce a behaviour when, and only when, it is presented with stimuli in the extension of the concept.[17] The offspring, in imitating the parent is also rewarded, and since it has a neural structure activated by the stimulus that is in the extension of the concept being learned, the appropriate behaviour is associated with that neural structure. When learning is complete, the offspring possesses a neural structure that initiates[18] a behaviour appropriate to the stimuli that activate it--stimuli in the extension of some concept. Another way in a natural concept learning situation that a behavioural response

---

[17] Again, this is not a claim of infallibility, only that there are conditions in which it can discriminate the items in the extension of the concept.

[18] The activation of the neural structure initiates a behaviour, but as we will see below and in the next chapter, on the model of embodied cognition, the action need not be manifested, so creatures are not stimulus/response systems.

can be associated with the neural structures activated by a stimulus is simply by a creature interacting with the stimulus in a way that results in something beneficial to it--the behaviour must be reinforced to establish an association. When a creature interacts with some stimulus, it can receive considerable information about the stimulus, so that the concept in whose extension the stimulus is might be learned very quickly, perhaps even a single instance. Whatever the means by which a particular neural structure comes to initiate some behavioural response in learning some concept C, because learning depends on the behaviour being reinforced, and because the neural structure reliably causally covaries with items in the extension of C, the behaviour is generally[19] appropriate in the presence of items in the extension of C: it is a C behaviour.

Now creatures are clearly not foolproof; sometimes they produce a behaviour that is not appropriate in the circumstances; sometimes non-dogs elicit DOG behaviour, for example. These are the error cases that lead to the disjunction problem for Fodor. However, they raise no difficulty for the embodied approach to content, because the nature of the error is evident. When a behaviour produced by a creature is inappropriate given the stimulus to which the behaviour is a response, information generated by the stimulus could have been generated, under the circumstances, by something in whose presence the behaviour would be appropriate. In conditions of poor light, for example,

---

[19] The reason the behaviour is "generally" appropriate to the stimulus and not always appropriate, is that contingent associations are always possible, such as when a creature repeats a behaviour in the presence of some stimulus purely accidently, but nonetheless comes to associate the behaviour with stimuli in the extension of some concept. Not all behaviours produced by such associations will manifest having a concept, however. I return to this issue in discussing what responses manifest having a concept below (5.5).

a fox can generate information that could have been generated by a dog, thereby activating some neural structure that can carry only information that could have been generated by dogs. The neural structure initiates a DOG behaviour, which is inappropriate because the stimulus is a fox.

5.5 What it is to Have a Concept

It seems untendentious that learning a concept is manifested by a creature's increasing tendency to produce a certain kind of response--to be made precise presently--when, and only when, certain environmental conditions obtain. On the embodied approach to content, however, having a concept is not just manifested by reliably producing some behaviour in the presence of stimuli in the extension of the concept, but *consists in* reliably initiating a behaviour in the presence of stimuli in the extension of the concept. The reason is that on the model of embodied cognition, perception is not passive reception of information from the environment, "[instead] perception is itself tangled up with specific possibilities of action" (Clark 1997, p.51). Because there is no central control structure on the model of embodied cognition, the activation of a neural structure that can carry only information that could be generated by items in the extension of some concept C must be what initiates a behavioural response. Thus, in circumstances in which a behaviour is not manifested, even though the neural structure is activated, it must be that other neural activity inhibits the production of the behaviour. To perceive something as an item in the extension of some concept C requires the activation of a neural structure that can carry only information that could be generated by items in the extension of C, ***which thereby initiates a behavioural***

response that is generally appropriate in the presence of items in the extension of C. So from the embodied approach to content, a creature has a concept if and only if it has a neural structure that can carry only information that could be generated by items in the extension of C, which when activated initiates a behaviour generally appropriate to the presence of items in the extension of C.[20] Notice that on this account of having a concept, we are not committed to supposing that Aristotle had the concept of an airplane, just because he had neural structures that would have reliably causally covaried with airplanes had there been any in his environment. Aristotle had no dispositions to behave given airplanes as a stimulus, because his environment was such that no connection had been made between his neural structures that would have covaried with airplanes and any behaviours. Aristotle had no AIRPLANE behaviours, and so no concept of an airplane.[21]

Clearly the embodied approach to content that I am developing is a dispositional account of content. In order to have a concept, a creature must be disposed to produce a behaviour when it receives information that could have been generated by an item in the extension of the concept. Now as we saw above (5.3) and (5.4), just being disposed to

---

[20] Notice that to have an innate concept on this account requires not only being endowed by evolution with neural structures that covary with certain types in the environment, but also having some behavioural response when confronted by stimuli in the extensions of those types. Since evolution only allows creatures whose behaviours are generally rewarded, and not too severely punished, to survive, the behaviour will generally be appropriate to the stimuli, but if not appropriate at least relatively neutral, just as in the case for learned concepts.

[21] This does not imply, of course, that had Aristotle encountered an airplane, he would not have produced some behaviour, only that he had no dispositions to behave in the presence of airplanes. Aristotle could not produce a response behaviour when presented with airplanes, and only airplanes, though he could have learned to do so.

produce a behaviour upon receiving information that could have been generated by items in the extension of a concept cannot be sufficient for having the concept; otherwise, reliably sneezing in the presence of ragweed would entail having the concept RAGWEED. Furthermore, at least some dispositional accounts of content, such as Ryle's[22], have proven unsuccessful. Indeed, Fodor begins *The Language of Thought* with a critique of Ryle's position. Turning back the clock fifty years is not the way to respond to Fodor. Thus, the embodied approach to content must be such that it precludes cases like reliably sneezing in the presence of ragweed from being a manifestation of the concept RAGWEED, and it must not be subject to the sorts of criticisms to which earlier dispositional accounts of content are subject.

Fodor's criticism of Ryle's position is directed towards Ryle's logical behaviourism. In particular, Fodor cogently argues that nothing in Ryle's conceptual analysis of behaviour precludes an explanation of behaviour in terms of underlying internal mechanisms. Fodor offers the example that there are two kinds of answers to the question, 'What makes Wheaties the breakfast of champions?', neither of which precludes the other. One answer is that Wheaties are eaten for breakfast by non-negligible numbers of champions. This is the sort of answer that Ryle's position requires. Fodor's point is simply that the answer that Ryle's position requires in no way precludes some causal story about the mechanisms by which Wheaties affect a person's body to make her a champion. It is the latter sort of answer that Fodor is interested in

---

[22] Ryle 1949.

(Fodor 1975, pp.2-9). All of this is perfectly sound, of course; behaviourism did not work.[23] But Armstrong had the insight that a dispositional account of content need not be behaviouristic. "Ryle's dispositional account of belief was developed as part of, and in order to support, a Behaviourist or Behaviourist-oriented theory of mind. It is important, therefore, to appreciate that there is nothing in the mere *dispositional view of belief* which entails the manifestations or expressions of a man's belief (if they occur at all) are all pieces of outward bodily behaviour" (Armstrong 1973, p.8). Indeed, Fodor's own recent view of content is dispositional; "an informational semantics... takes *the content of one's concepts to be constituted by one's dispositions to apply them.* And informational semantics is being assumed for the purposes of this discussion" (Fodor 1994, p.31, my emphasis).[24] The embodied approach to content that I am developing is dispositional, but not behaviouristic, as I now make clear.

The embodied approach to content that I am developing is very similar to views advanced by Stalnaker and Dennett. "Belief and desire... are correlative dispositional states of a potentially rational agent... [i]n ascribing beliefs and desires to a person, we not only make conditional predictions about how the person will behave; we also commit ourselves to claims about the kind of mechanisms which explain why a person behaves the way he does" (Stalnaker 1984, pp.15-17). Stalnaker makes two critical points, both of which are features of the embodied approach to content. First, mental

---

[23] I take it that this was established beyond question by Chomsky (1959). However, see Dennett (1987, pp.44-5) for an account of what can be preserved from Ryle's conceptual analysis.

[24] Fodor (1998) expresses the view that what makes a concept token a token of a specific type are the dispositional properties of the token.

ascription does not deny the relevance of internal states in determining content. Far from attempting to preclude an explanation of behaviour in terms of internal mechanisms, the embodied approach to content is grounded on there being such mechanisms. In order to have a concept, a creature must possess some neural structure that disposes it to produce a certain behaviour in the presence of items in the extension of that concept. On the embodied approach to content, no overt bodily movement need be produced, even when a neural structure is activated. All that is required by the embodied cognition model is that some behaviour is initiated when the neural structure is activated. Furthermore, there is no particular behaviour that is either necessary or sufficient for having a concept. What is required is that a creature can reliably respond to items in the extension of the concept, how it so responds is irrelevant. But the nature of the response is crucial, and this connects with Stalnaker's second point. The response must be one that can figure in an intentional explanation, which is why Stalnaker supposes the potential *rationality* of the agent. And it is for this reason that being disposed to sneeze in the presence of ragweed is not a manifestation of possessing the concept RAGWEED; sneezing is not a behaviour that figures in intentional explanation, at least not as an allergic reaction of the kind we are supposing.

Intentional explanation assumes the rationality of the creature whose behaviour is being explained. In particular, a creature is assumed to have beliefs and desires; the creature's rationality consists in the creature behaving so as to satisfy its desires given

its beliefs. "A system's[25] behaviour will consist of those acts that *it would be rational* for an agent with those beliefs and desires to perform" (Dennett 1987, p.49, italics in original). Now recall (5.1) that one of the assumptions of the embodied approach to content is that mental states are intrinsically holistic. The assumption of holism is tied directly to the nature of intentional explanation. To have a concept, a creature must possess a neural structure that can carry only information that could be generated by items in the extension of the concept, such that when the neural structure is activated it initiates a behaviour that can figure in intentional explanation. That is, it initiates a behaviour which is rational given that a creature has certain beliefs and desires. Since the behaviour must be able to figure in intentional explanation--i.e. it must license the ascription of beliefs and desires to the creature--creatures do not possess concepts in isolation. The embodied approach to content is similar to the views of Dennett and Stalnaker in that every individual ascription of a concept to a creature is made with regard to the background of beliefs, desires, and other concepts[26] that a creature has.[27] Any element in the background is subject to modification during an individual ascription, though significant modifications would be warranted only in the case of near

---

[25] Dennett's discussion includes all intentional systems, not just creatures.

[26] Again, beliefs and desires are dispositions that can figure in intentional explanations (Stalnaker 1984, p.15) and (Dennett 1978, pp.3-22 and 1987, p.49).

[27] Davidson's view is also holistic. "There is no assigning beliefs to a person one by one on the basis of his verbal behaviour, his choices, or other local signs no matter how plain and evident, for we make sense of particular beliefs only as they cohere with other beliefs, with preferences, with intentions, hopes, fears, expectations, and the rest" (Davidson 1970, p.221). I have restricted my presentation in the main body of the text to discussing Stalnaker and Dennett, not to exclude Davidson's view from my general approach, but because of their explicit focus on the dispositional nature of intentionality.

total explanatory failure.[28] Of course, this position is circular in that ascriptions are interdependent, a consequence of the assumption that the intentional predicates form a closed circle (5.1), but it is not viciously circular because intentional explanation is successful in explaining and predicting creatures' behaviours.[29] What is offered "is a whole system of interlocking attributions, which is saved from vacuity by yielding independently testable predictions" (Dennett 1987, p.50).

Dennett presents his position of what is involved in the ascription of mental states and properties through the idea of a "notional world" (Dennett 1987, pp.151-173). A notional world is a fictitious construct of how a creature takes the world to be at a given moment, as determined by its dispositions at that moment. "The notional world we describe by extrapolation from current state is ... the apparent world of the creature, the world apparent *to* the creature as manifested in the creature's current total dispositional state" (Dennett 1987, p.157). A notional world captures what it is about a creature that contributes to the content of its mental states.[30] In order to determine the notional world of a creature, one must consider its total dispositional state, since content is ascribed in virtue of the functional relations of dispositions between each other and the world--mental ascription is holistic. Furthermore, the dispositions must be

---

[28] Another conclusion that can be drawn from near total explanatory failure is that the creature is not rational.

[29] Stalnaker accepts the naturalistic constraint, that is he rejects the assumption that the intentional predicates form a closed circle (Stalnaker 1984, pp.15-6), because he is worried about the threat of vicious circularity, but he overlooks the success of intentional explanation as a response to the threat.

[30] A notional world delimits the domain of narrow psychology.

those that figure in intentional explanation, since the notional world of a creature is the "apparent" world of the creature, how it takes the world to be. That a creature is disposed to sneeze in the presence of ragweed does not reveal anything about how the creature takes the world to be. A creature's notional world is not unique, however. Worlds that differ in respects to which a creature is not sensitive, in that it has no dispositions to behave differently in the worlds, are not discriminated in the creature's notional world. Thus, a creature's notional world is indifferent between Earth and Twin Earth.[31] The actual relations between a creature and its environment are required to exclude XYZ from being in the extension of its WATER concept.

Given that the relevant dispositions for having concepts are those that figure in intentional explanation, we can now see more clearly the sense in which a behaviour which manifests that a concept has been learned is generally appropriate in the presence of items in the extension of that concept. Because the behaviour must be one that can figure in intentional explanation, it must be a behaviour through which the creature attempts to satisfy its desires given its beliefs. Now in the process of concept learning the behaviour is reinforced, so that upon learning a concept a creature comes to believe that it can obtain something it desires by producing the behaviour in the presence of items in the extension of that concept. The behaviour is appropriate in the presence of items in the extension of the concept because the creature is viewed rationally as trying to satisfy its desires given its beliefs. Notice that "the logic of the experimental

---

[31] "The Meaning of 'Meaning' ", in Putnam 1975, especially pp.223-227.

paradigm", as Fodor puts it, smuggles in this sense of the appropriateness of a behaviour to the stimulus through the notion of reinforcement, since reinforcement satisfies some of a creature's desires.

5.6 The Embodied Approach to Content

I have now presented all of the features of the embodied approach to content. In this section I present them together to articulate just what the position is. I then consider some of the implications of the approach, and I evaluate the notion of concepts presented using Fodor's conditions on what concepts have to be (Fodor 1998, pp.23-39). The embodied approach to content has four assumptions:

E1. The intentional/semantical predicates form a closed circle;

E2. Intentional states are intrinsically holistic;

E3. Causal relations between neural structures implicated in concept learning and concept possession do not correspond to the semantic associations between symbols in a language of thought;

E4. Creatures' nervous systems have an embodied cognition design.

On the embodied approach to content, having a concept consists in (E4) possessing a neural structure that can carry only information generated by items in the extension of the concept (E3), which when activated initiates a behaviour that can figure in an intentional explanation of the creature[32]; that is, the creature is disposed to produce a behaviour that can figure in an intentional explanation in the presence of items in the extension of the concept (E2 and E1). It is open on this approach as to whether content

[32] I am assuming that intentional explanations appeal to the actual causal histories of creatures to deal with Putnam's Twin Earth cases (Putnam 1975, pp.223-227).

is ascribed to neural structures, though in any case concepts are ascribed to creatures as a whole; "[t]he *subject* of all the intentional attributions is the whole system (the person, the animal ...) rather than any of its parts" (Dennett 1987, p.58, emphasis in original). Jennifer Hornsby also endorses this view; "an account of what the attitudes are is a personal-level account" (Hornsby 1997, p.169). Even if neural structures are ascribed content, concepts are not identical with or reducible to those neural structures. Having a concept requires having a disposition to produce a behaviour that can figure in an intentional explanation. A neural structure is simply the mechanism realizing the disposition, but is not itself the disposition. In virtue of having neural structures with certain functional roles, *creatures* have dispositions that can figure in intentional explanations, and hence concepts. Seeing the relation between neural structures and concepts, we can distinguish between having a concept and using the concept. To have the concept, a creature must have a neural structure, which, when activated, initiates a response appropriate in the presence of items in the extension of the concept. To use the concept in determining behaviour, the neural structure must be activated.[33] If content is ascribed to the neural structure in virtue of its functional role, then the activation of that neural structure is a tokening of the concept. If content is not ascribed to the neural structure, then the activation of the neural structure is using the concept in determining behaviour because activating the neural structure initiates a behaviour that figures in an

---

[33] Clark points out that once a creature (or system) has some appropriate disposition to behave it has a concept, and provided it maintains that disposition, the neural structure responsible for it need not remain constant, as analogous structures are not constant in connectionist systems (Clark 1995, p.350).

intentional explanation of the creature which ascribes the concept to the creature; in that case, the activation of the neural structure is not a tokening of the concept. A final point about the embodied approach to content is that it does not entail any particular theory of content; it includes, not exclusively, the positions of Davidson (anomalous monism)[34], Dennett (the intentional stance)[35], Putnam (internal realism)[36], Clark[37], Burge[38], Hornsby[39], and Pietroski[40]. Now many of these writers are not explicitly committed to all of the assumptions of the embodied approach to content, particularly E3 and E4, but their positions are compatible with these assumptions. As I indicated above (5.1), issues such as mental causation may favour a particular position, but those issues are tangential to my project.

Since the embodied approach to content is being offered as an alternative to Fodor's theory of content with the aim of refuting his argument that the internal representational system must be linguistic and must be the medium of thought, I now evaluate the embodied approach to content according to Fodor's "five not negotiable conditions on a theory of concepts" (Fodor 1998, p.23). Fodor's five conditions are:

---

[34] Davidson 1970, pp.207-225.

[35] Dennett 1978, pp.3-22, Dennett 1987.

[36] Putnam 1988.

[37] Clark 1995, pp.347-350.

[38] Burge 1986, pp.3-46, Burge 1993, pp.97-120.

[39] Hornsby 1997.

[40] Pietroski forthcoming.

1. Concepts are mental particulars; specifically, they satisfy whatever ontological conditions have to be met by things that function as mental causes and effects.

2. Concepts are categories and are routinely employed as such.

3. Concepts are the constituents of thoughts and, in indefinitely many cases, of one another.

4. Quite a lot of concepts must turn out to be learned.

5. Concepts are *public*; they're the sorts of things that lots of people can, and do, *share* (Fodor 1998, pp.23-28, emphasis in original).

Condition 1 requires some remarks about mental causation, which I present below. Condition 2 simply says that things in the world are in the extensions of concepts and it is clearly satisfied by the embodied approach to content. Condition 3 is about systematicity and productivity. I argue in the next chapter that language-using humans have these capacities, in virtue of language, but animals do so only to a limited degree. Clearing up this issue is essentially what remains of my project, so I let the discussion of it unfold in my explanation of animal and preverbal human behaviour. Condition 4 is certainly satisfied by the embodied approach to content, and is particularly true of humans, because we aren't born with very many behavioural responses.[41] Finally, condition 5 is also clearly satisfied by the embodied approach to content, because anything that has a neural structure that can carry only information generated by items in the extension of some concept, and which can associate some behaviour with that

[41] My account of early language acquisition (6.7) also suggests that virtually all human concepts are learned.

neural structure, can learn that very concept.

A full response to condition 1 is a thesis (at least) unto itself. The condition is central to Fodor's project, because it entails his approach to trying to preserve psychological explanation in the face of eliminativism. Since a psychological explanation is just an explanation of behaviour in psychological terms, if there really are psychological explanations, it must be because there are mental causes and effects. The challenge of this claim to the embodied approach to content comes from cashing out the "ontological conditions [that] have to be met by things that function as mental causes and effects". Many philosophers share a "metaphysical prejudice" that Fodor admits to having. "I'd better 'fess up to a metaphysical prejudice... I don't believe that contents per se determine causal roles. In consequence, it's got to be possible to tell the whole story about mental causation (the whole story about the implementation of the generalizations that belief/desire psychologies articulate) *without referring to the intentional properties of the mental states that such generalizations subsume*" (Fodor 1987, p.139, emphasis in original). If one is moved by this "metaphysical prejudice", concepts must be identified (at least token-wise) with physical structures in order to function as mental causes and effects, since their content does not determine their causal role. This "metaphysical prejudice" is unproblematic for the embodied approach if one holds that content is ascribed to neural structures. The activations of certain neural structures *are* tokenings of concepts on this view, hence concepts satisfy the ontological conditions that Fodor thinks have to be met for them to function as mental causes and effects. In fact, on the embodied approach to content under the assumption that neural

structures are ascribed content, concepts satisfy the same ontological conditions that they do in Fodor's own token physicalism. On the other hand, if one holds the "metaphysical prejudice", then concepts that are ascribed to an entire creature, in virtue of the functional role of neural structures that are *not* ascribed content, do not "satisfy whatever ontological conditions have to be met by things that function as mental causes and effects". So if one is moved by this "metaphysical prejudice" the embodied approach to content cannot satisfy condition 1 if neural structures are not ascribed content; and concepts that play no causal role are ripe for elimination. Interestingly enough, however, this "metaphysical prejudice" is the same one that leads Fodor to adopt the naturalistic constraint (2.2), which leads to problems for Fodor's theory of content. Perhaps in rejecting the naturalistic constraint, we can also reject the "metaphysical prejudice" that motivates it.

The challenge for the embodied approach to content is to satisfy Fodor's condition 1, when it is assumed that the neural structures, in virtue of whose functional role concepts are ascribed to creatures, do not themselves have any content. In particular, if neural structures do not have content, then concepts cannot be mental causes and effects if one assumes that "it's got to be possible to tell the whole story about mental causation (the whole story about the implementation of the generalizations that belief/desire psychologies articulate) *without referring to the intentional properties of the mental states that such generalizations subsume*". Thus, this version of the embodied approach to content can only satisfy condition 1 if the "metaphysical prejudice" is rejected. However, since the "metaphysical prejudice" is widely held, it

cannot be rejected without an argument. Burge (1986, 1993), Hornsby (1997), and Pietroski (forthcoming) are engaged in a research project that offers just such arguments. Their general move is to reflect on explanatory practice. In particular, they focus on the nature of causal explanations. According to Pietroski, we understand instances of causation in terms of paradigmatic cases of causation, which thereby constrains what kind of notion causation is. Now one paradigmatic case of causation is *mental* causation, as borne out by the success of intentional explanation; i.e. the success of explaining and predicting creatures' behaviours in terms of their beliefs and desires within the context of their conceptual frameworks. Thus, on this view there is simply no room to doubt mental causation. Questions about how reasons could be causal betray a fundamental misconception of the nature of mental causation (Pietroski forthcoming). According to Burge, the reason for this misconception of the nature of mental causation is that materialist metaphysics has been given too much importance, and too little attention has been given to explanatory practice. While Burge offers several arguments that vary in detail, his main point is simply that in giving a causal explanation "we consider an entity's causal powers relative to the kind in terms of which the entity is specified" (Burge 1993, p.101). Demanding an underlying mechanism for mental causation requires an explanation in terms of properties that appear in the physical sciences, so it is not surprising that this way of conceiving of mental causation leads to epiphenomenalism. "One cannot understand mentalistic causation... and mental causal powers by concentrating on properties characterized in the physical sciences. Our understanding of mental causation derives primarily from our understanding of

mentalistic explanation, independently of our knowledge--or better, despite our ignorance--of the underlying processes" (Burge 1993, p.103). Systematic, informative, explanatory schemes indicate causal relevance, and the causally efficacious properties are those that enter into explanations under such schemes. Intentional explanations *are* systematic and informative, demonstrating the causal relevance of mental properties. "The probity of mentalistic causal explanation is deeper than the metaphysical considerations that call it into question (Burge 1993, pp.117-8).

Hornsby's analysis of intentional explanation focuses on the notion of agency. Intentional explanation is a rationalizing explanation of an agent's actions, i.e. an explanation of what agents do as rational. "What we rely on is only a network of intelligible dependencies between the facts about what an agent thinks, what she wants, and what she does.... And the dependence is of a causal sort, of course" (Hornsby 1997, p.135). Hornsby argues that agency just is a causal notion. "Our conception of a person as an agent is a conception of something with a causal power" (Hornsby 1997, pp.131-2). The events in which agents cause something are actions; that is, actions are what an agent contributes to an event. Thus actions are ascribed to agents in a rational explanation of what the agent did, and hence are intentionally characterized. Now the key step in Hornsby's argument is a rejection of token physicalism. Actions cannot be identified with anything in a physical (neural) causal chain of events because by their very nature, actions are what agents do. Any physical event that is a candidate for being identified with an action is preceded by some other physical event, which caused it. But the notion of agency disappears if physical events cause actions, so none of the physical

events that are candidates for identification with an action can be so identified. Now it is only "once this further idea is in place [the idea that actions are token identical with physical events inside a person], it can come to seem that the explanatory value of *belief* and *desire* is quite unconnected with the value of those concepts in causal understanding--as if the particular contents of a particular person's beliefs and desires had nothing to do with her tendencies to do one thing rather than another" (Hornsby 1997, p.135). By rejecting token physicalism, rational explanation just is causal explanation. The success of rational explanation leaves no room to doubt mental causation.

Finally, even Fodor, when he is not being "metaphysically prejudiced", makes a case for mental causation that does not ground mental causation on the physical properties of mental events. "I'm not really convinced that it matters very much whether the mental is physical; still less that it matters very much whether we can prove that it is. Whereas, if it isn't literally true that my wanting is causally responsible for my reaching, and my itching is causally responsible for my scratching, and my believing is causally responsible for my saying..., if none of that is literally true, then practically everything I believe about anything is false and it's the end of the world" (Fodor 1990, p.156). Fodor's first point in arguing for mental causation is that the arguments for epiphenomenalism are sufficiently general so as to apply to all of the special sciences[42], from which he makes the case that causally responsible properties are

[42] Burge also makes the point that arguments for epiphenomenalism "would [leave] no room for causal efficacy in the special sciences, even in natural sciences like chemistry and physiology" (Burge 1993, p.102).

"those in virtue of which individuals are subsumed by causal laws... then intentional properties are causally responsible in case there are intentional causal laws" (Fodor 1990, p.143). The tension that Fodor must deal with is that causal laws are strict deterministic laws, of which the only instances are physical laws; laws of the special sciences, psychological laws in particular, are ceteris paribus laws. Fodor's proposal for resolving this tension is to relax the condition that causal laws are strict deterministic laws. The point of the strictness requirement is that a causal law must be such that an instantiation of an antecedent property must be sufficient for the instantiation of a consequent property. Fodor's insight is that special science laws do necessitate their consequents, when the ceteris paribus conditions are met, so that they can serve as the covering laws in virtue of which intentional properties are causally responsible. That is, mental properties are causally responsible because intentional states are nomologically sufficient for producing behavioural outcomes, when the ceteris paribus clauses of laws subsuming intentional states and behavioural outcomes are discharged (Fodor 1990, p.152).[43]

The general conclusion that I want to draw from this discussion is that even if neural structures are not ascribed content on the embodied approach to content, arguments by Burge, Hornsby, Pietroski, and Fodor himself provide reasonable

---

[43] It is worth noting that though Fodor argues that mental properties are causally responsible, he also argues for token physicalism. Physicalism follows because every non-basic law, including every law of the special sciences, has an implementing mechanism, which is plausibly physical, according to Fodor. Fodor takes the connection of mental causes to their effects via some physical mechanism to entail token physicalism. Nonetheless, mental properties are not causally inert on this view.

grounds for thinking that the embodied approach to content can satisfy Fodor's condition 1. The nature and success of intentional explanation suggest that concepts as construed in the embodied approach to content "satisfy whatever ontological conditions have to be met by things that function as mental causes and effects". Thus, pending some issues about condition 3 to be addressed in the next chapter, the embodied approach to content can satisfy all of Fodor's non-negotiable conditions on a theory of concepts.

5.7 The Internal Representational System

Recall that the reason we supposed there is an internal representational system is that animals and preverbal humans exhibit some of the cognitive capacities that language-using humans demonstrate, and models of those capacities are computational. I am not challenging that there is an internal representational system, though clearly from the embodied approach to content I am endorsing, it does not have the same characteristics that Fodor takes it to have. Fodor argues that the internal representational system is innate. What this entails is that "if a concept belongs to the primitive basis from which complex mental representations are constructed, it must ipso facto be *un*learned" (Fodor 1998, pp.27, emphasis in original). Thus, we have an innate primitive basis of concepts and some ability to combine them to form new concepts. A particular mental representation is a tokening of some concept type. From the alternative I have presented, what is innate is neural structures that can carry only

information that could be generated by items in the extension of some concept.[44] These neural structures may or may not be associated with behaviours that can figure in an intentional explanation such that they initiate the behaviours when activated, in which case a concept may or may not be innate. Similar to Fodor's view, the types of mental representations are determined by the concepts a creature has. But since having a concept on the embodied approach to content is quite different from Fodor's view, so is having a representation. The internal representations just are a creature's concepts, and a creature is ascribed a particular representation of something when an intentional explanation of the creature's behaviour includes the concept of that thing. Now since concepts are dispositions to behave in ways that figure in intentional explanation, the internal representational system of a creature just is its total disposition relevant to intentional explanation.

One merit of the embodied approach to content is that the content of concepts is unproblematically the content we use in psychological explanation. Certain neural structures carry only information that could have been generated by the items in the

---

[44] Fodor (1998, chapter 6, especially p.143) seems to suggest that perhaps only neurological states are innate, as I have argued. I have had the text too short a time to assess this claim and its implications, but a few brief remarks are appropriate. First, if neurological states do not have content, then Fodor owes us an account of what it is to have a representation and a concept, and my guess is that he will have considerable difficulty in avoiding the kind of account I have developed. Furthermore, in the interests of giving an atomistic account of content, Fodor is suggesting that concept learning is a kind of *non-inductive* "locking on" procedure, in which neurological states get "locked" to properties in the world. If the procedure truly is non-inductive, then none of Fodor's arguments for why the language of thought is distinct from natural language hold. However, I suspect that when the "locking on" procedure is cashed out it will be an inductive procedure, since there are indefinitely many features of the environment onto which we could become "locked". Determining which feature is correct will likely require a model of hypothesis formation and testing. My account of concept learning, (5.3) and (5.4), can be seen as an attempt to cash out the "locking on" procedure.

extension of our intuitive semantic types. It is not miraculous that they should do so because creatures co-evolve with the items in those semantic types and their survival depends on being able to reliably discriminate those items from items of another type. Of course, it might be that some animal does not have the perceptual capacity to distinguish items belonging to what we consider two distinct types--it has a neural structure that can carry only information that could have been generated by items in the extensions of both types but not either type individually. In this case, the animal has a more coarse-grained concept than we do. Furthermore, the error cases that lead to the disjunction problem for Fodor's theory of content pose no problem for ascribing the content used in psychological explanations on the embodied approach. The reason is that the embodied approach to content rejects the naturalistic constraint, so ascriptions of content are licensed by *intentional* explanations of creatures' behaviours. But the content appealed to in intentional explanations just is the content of our intuitive semantic types used in psychological explanation.

Because humans and animals can both have neural structures that can carry only information that could have been generated by items in the extension of some concept, they can have the same concept, provided they can associate some behaviour that can figure in intentional explanation with the neural structure. But in one important respect, internal representations as I have characterized them are different from our intuitive semantic types. The difference is that there are no semantic associations between the internal representations. The internal representational system is not a semantic net; the only representations that creatures can have are of items they encounter, and only when

they encounter them. Fodor might take this to be a reductio of my position, but I argue in the next chapter that this structure is sufficient to explain the cognitive capacities of animals and preverbal children. Furthermore, the embodied approach to content is no worse off than any theory currently available in not having a complete story of how we can think of things that either are not present or that we have never encountered. Nonetheless, I think this approach offers some insights as to how to proceed to develop an account of abstract thought, which I present below (6.8). I turn now to explaining animal and preverbal human behaviour using the embodied approach to content.

CHAPTER 6

EMBODIED COGNITION: AN ALTERNATIVE TO MENTALESE

6.1 Introduction

I now return to the outstanding issue, raised in (4.8), of explaining the cognitive achievements of preverbal children and some animals without requiring that the internal representational system possess anything corresponding to the logical connectives. From the embodied approach to content, I argue that it is possible to explain animal behaviour and early language acquisition in terms of hypothesis formation and confirmation, without requiring that the internal representational system possess anything corresponding to the conditional. I then consider how it could be that animals learn concepts that we express by using a logical connective in natural language, which I refer to as concepts having a logical form, without possessing anything corresponding to the logical connectives. From this account it follows that natural languages could be more expressive than the internal representational system; hence the language of thought could be a natural language and the internal representational system need not be linguistic. I conclude by considering some merits of my position over Fodor's view.

6.2 Animal Behaviour on the Model of Embodied Cognition

Recall the animal behaviour that must be explained: "All three of the processes that we examined [earlier]--considered action, concept learning, and perceptual integration--are familiar achievements of infrahuman organisms and preverbal children. ... Computational models of such processes are the only ones we've got" (Fodor 1975, p.56). Specifically, we require an account of how it could be that animals form and

confirm hypotheses in considered action, perceptual integration, and concept learning, without requiring that anything in the internal representational system correspond to the conditional (or any of the other logical connectives).

Dennett presents a multiple drafts model of cognition, in which creatures determine several possible behaviours that might be appropriate in virtue of the information a creature receives (Dennett 1991, pp.101-138). The possible behaviours then compete amongst themselves for control of the actual behaviour a creature manifests.[1] Now while it does not matter for our purposes just how the internal structure is organized, a decentralized model of cognition does entail something along the lines of a multiple drafts model. On the model of embodied cognition, the information a creature discriminates partially determines the creature's behaviour. The reason is that to discriminate information, some neural structure of the creature is activated in virtue of covarying with things that generate that information, and the activation of that neural structure initiates a behaviour. However, the behaviour that a particular neural structure initiates need not be manifested. Creatures often obtain different information from different senses, which can activate distinct neural structures, thereby initiating distinct behaviours simultaneously. On the embodied cognition model, the neural activity initiating distinct behaviours spreads by local interactions until it becomes coordinated with respect to what other parts of the nervous system are doing, from which a single behaviour emerges. The coordination of activity

---

[1] Classifier systems are designed with exactly this structure (Holland 1975, Holland 1992, Holland et. al. 1987).

itself occurs at a local level, with no overall representation of the different behaviours initiated. Like the multiple drafts model, one behaviour emerges from multiple competing possibilities. It is important to note that while the specific nature of the mechanisms by which a behaviour emerges need not concern us, there are constraints on the mechanisms. For example, the process is not a random selection from the initiated behaviours. The emergent behaviour is determined in some way as a function of the past successes of the behaviours initiated; that is, in virtue of the reinforcement the behaviours initiated have received.[2] Furthermore, in instances of cognition, the behaviours initiated can figure in intentional explanation. Thus according to the embodied approach to content, the behaviour that emerges is the one most likely to satisfy the creature's desires, given its beliefs, which is the sense in which the creature is rational.

In fact, information need not be received by different senses in order for several behaviours to be simultaneously initiated. For example, an animal viewed in dim light might generate information that could have been generated by a fox or a dog. A neural structure that can carry only information that could be generated by a dog, and a neural structure that can carry only information that could be generated by a fox might both be activated,[3] initiating a DOG behaviour and a FOX behaviour, from which the local

---

[2] Behaviours that are reinforced dispose the creature to produce those behaviours more often. Classifiers systems function in this way.

[3] Of course, it is possible that the creature could have a neural structure that can carry information generated by a dog or a fox, which initiates a behaviour in such a situation. The embodied approach to content does not preclude such structures. It merely requires that in addition there be a structure that can carry only information generated by a dog and a structure that can

neural processes interact with each other to determine a single behaviour. A special case of this might be when the information received initiates distinct behaviours that are appropriate in the presence of the same stimulus. When a creature learns a concept it does so by associating some behaviour with a neural structure that can carry only information that could be generated by items in the extension of the concept. But nothing precludes a creature that has already learned a concept from associating another behaviour with that neural structure, in the same way that the first behaviour was associated with the neural structure (5.4).[4] If the stimulus is a dog, then the creature has two DOG behaviours, which are both initiated in the presence of dogs and compete to be manifested. Thus, there are several different ways that, on the model of embodied cognition, creatures determine possible behaviours in parallel.

Now from the embodied approach to content, each possible behaviour that a creature initiates licenses the ascription of a hypothesis to the creature. However, the form of these hypotheses is not hypothetical and so does not require that the internal representational system possess something corresponding to the conditional. Through the activation of certain neural structures, the creature *actually initiates* each of its

---

carry only information generated by fox, both of which initiate behaviours that can figure in intentional explanations, if the creature has the concepts DOG and FOX rather than some more coarse-grained concept.

[4] Or what amounts to the same thing, that a creature has distinct neural structures that can carry only information that could be generated by items in the extension of some symbol. A second behaviour can be associated with a second neural structure. So, the information can be conceived as activating a single neural structure that initiates distinct behaviours, or initiating distinct neural structures that each initiate a single behaviour. How this gets worked out depends on how neural structures get individuated, but nothing in my discussion depends on how the individuation goes.

possible behaviours, and were it not for interactions between the neural activity initiating each behaviour, those behaviours would be manifested. Thus, what we ascribe to a creature from the embodied approach to content does not have a hypothetical form; it is only a hypothesis because of the context in which it is ascribed.[5] To see this more clearly we must consider the circumstances in which creatures engage in considered action, perceptual integration, and concept learning, and determine the nature of the hypotheses ascribed in each case.

6.3 Considered Action

An animal engages in considered action just when there are distinct possible behaviours that can figure in an intentional explanation, which the creature can manifest in its circumstances. In order for there to be distinct possible behaviours it can manifest, it must be the case that distinct neural structures are activated that can initiate distinct behaviours. However, on the model of embodied cognition, the activation of neural structures *actually initiates* behaviours. The behaviours are possible and not actual only in the sense that they might not be manifested, because other neural activity might interact with the neural processes initiating one behaviour, such that a different behaviour emerges. In considering what action to take, an animal tries to determine what behaviour is the most appropriate in its circumstances--what behaviour will be most strongly reinforced or least severely punished. Since neural structures activated by the circumstances actually initiate behaviours, on the embodied approach to content, an

---

[5] Very often it is the context of a representation and not its form that makes it a hypothesis. Mathematical conjectures are stated as assertions, for example.

animal determining what action to take is not ascribed an internal representation of the hypothetical form, "If I do behaviour B, it will be most appropriate in these circumstances". Rather, it is ascribed an internal representation of the assertive form, "Doing B now is most appropriate". The assertion serves as a hypothesis only because in the context it is ascribed, other assertions about what behaviour is most appropriate are also ascribed. Furthermore, the animal does not make a comparative judgment between behaviours. That an animal initiates a behaviour in some circumstance *is* its expression that the behaviour is the most appropriate in those circumstances, for which the hypothesis is ascribed. The actual behaviour that emerges is deemed by the animal to be the most appropriate in the circumstances, in virtue of that behaviour being manifested.[6] How the environment responds to the behaviour manifested determines whether the hypothesis is confirmed or disconfirmed. Thus, on the embodied cognition model, considered action is a process of hypothesis formation and confirmation, in which none of the hypotheses is a hypothetical statement. The internal representational system need not possess anything corresponding to the conditional. For example, a hare might receive information that could have been generated by a fox. Suppose that appropriate behaviours for hares in the presence of foxes are to run or to lie still. Given the information the hare receives, both behaviours are possible. That is, early processing of the information activates a neural structure[7] that could produce each of the

---

[6] Recall (6.2) that what behaviour emerges is a function of the extent to which the behaviours initiated have been reinforced in the past.

[7] Or structures; again this depends on how the individuation of neural structures goes.

behaviours, though, of course, only one of them will actually emerge as the hare's behaviour. But neither behaviour is only hypothetically entertained, since both are initiated, and it is only the intervention of other neural processing that prevents one of the behaviours from being manifested. Some of the processing, if left uninterrupted will result in the hare running, other processing will result in the hare lying still. From the embodied approach to content the hare is ascribed two hypotheses about what behaviour to perform: "Running now is most appropriate" and "Lying still now is most appropriate". Clearly, nothing in these hypotheses presupposes anything corresponding to the conditional.

Now it might be wondered how it is that from the embodied approach to content a hypothesis that asserts some behaviour B is the most appropriate in the circumstances is ascribed, when B is not the behaviour manifested. This is a serious challenge from a Fodorian perspective, because on the embodied approach to content, an animal is only entertaining a hypothesis if it can be ascribed the hypothesis. Content is typically ascribed in virtue of the actual behaviour an animal manifests in its circumstances. But, many hypotheses must be ascribed in virtue of non-actual, possible behaviours. However, a non-actual, possible behaviour is non-actual only in the sense that it is not manifested. Because the activation of some neural structure *actually initiates* a behaviour B, if a stimulus activates the neural structure, we can ascribe the hypothesis, "Doing B now is most appropriate", to the animal, even if B is not manifested. If the activation of that neural structure initiates more than one behaviour, we ascribe more than one hypothesis to the animal. One way we might *tell* that some stimulus generates

information that activates a neural structure in an animal, which thereby initiates a behaviour that is not manifested, is by presenting the animal with various stimuli in controlled conditions, or simply observing the animal in natural conditions, and determining what it responds to and how. Any response, B, a creature produces to some stimulus must be initiated by a neural structure that the stimulus activated. Hence, even when the behaviour B is not manifested in the presence of the stimulus, the neural structure is activated nonetheless, which initiates B.[8] Thus, we can ascribe the assertion, "Doing B now is most appropriate", as a hypothesis, given the context of ascription, that the animal entertains in considering what action to take. Furthermore, considered action is a cognitive achievement of animals according to Fodor, so there must be some evidence that an animal is considering what action to take.[9] That evidence can be used to ascribe hypotheses to the animal asserting the appropriateness of behaviours that are not manifested. If there is no evidence, there is no reason to accept that the animal is considering what action to take. Since it is possible to ascribe hypotheses asserting the appropriateness of behaviours that are not manifested to animals considering what action to take, considered action is a process of hypothesis formation and confirmation, on the embodied approach to content. The same reasoning shows that hypotheses concerning behaviours that are not manifested can be ascribed to animals during perceptual integration and concept learning, and to preverbal humans in early language

---

[8] There are worries about whether the animal is attending to the stimulus present, so this should be hedged with a ceteris paribus clause; but surely Fodor would not object to the use of a ceteris paribus clause, especially since it need not be cashed out naturalistically.

[9] Fodor offers no examples of considered action by animals.

acquisition.

6.4 Perceptual Integration

Recall the problem of perceptual integration is "that of choosing the best hypothesis about the distal source of proximal stimulations" (Fodor 1975, p.50). Sensory mechanisms are sensitive to physical properties, though perceptual categories are not captured by the vocabulary of physics. Thus, in perceptual integration hypotheses must be formed about what the source of given information is. Before giving an account of perceptual integration on the model of embodied cognition, it is important to consider in what circumstances this cognitive process is performed. Curiously, Fodor offers no examples of the circumstances under which animals perform the process of perceptual integration. Furthermore, Fodor concedes that "there is no reason to believe that organisms are usually conscious of the sensory analyses that they impose" (Fodor 1975, p.48). Thus, it is not clear that our conscious deliberations about what we experience are a very good model for animal perceptual integration. The first problem is that "an indefinite number of perceptual analyses will, in principle, be compatible with any given specification of a sensory input" (Fodor 1975, p.50), but animals with finite capacities cannot entertain all of these possibilities as hypotheses. Something in the way animals are designed must constrain the possibilities, so that the process of hypothesis formation and confirmation is one that animals can actually perform. Now an animal can only categorize a distal source of proximal stimulations as being in the extension of a concept that it has, and we saw in (5.3-5.5) that an animal cannot have a concept unless it has a neural structure that can carry only information

that could be generated by items in the extension of that concept. Since an animal has only finitely many neural structures, it can entertain only finitely many hypotheses about a distal source of proximal stimulations. The question remaining then is in what circumstances must a creature entertain more than one hypothesis; i.e. in what circumstances is perception really hypothesis formation and confirmation? An animal can categorize a stimulus as being in the extension of some concept C, only if the stimulus generates information that could have been generated by items in the extension of C, thereby activating a neural structure that can carry only information that could be generated by items in the extension of C. So perceptual integration occurs only in cases where the information a creature receives activates distinct neural structures; that is, when the information a creature receives could have been generated by an X or a Y, thereby activating neural structures in the animal that can carry only information that could have been generated by an X or a Y.[10] For example, a hare might receive information that could have been generated by a fox or a dog, posing it with a problem of perceptual integration.[11] Where there is no perceptual ambiguity, only one hypothesis is formed.

Recall that the discrimination of information, on the model of embodied cognition, is essentially tied to the production of behaviour. When neural structures that reliably causally covary with items in the extension of a concept C are activated, they

---

[10] If a neural structure is activated that can carry information about both Xs and Ys, such that the creature's behaviour is determined by the more coarse-grained discrimination of the information, then the creature is not engaged in perceptual integration.

[11] Notice that these are the very cases that give rise to the disjunction problem.

initiate some (one or more) behavioural response(s). When an animal is in the position of perceptual uncertainty, distinct neural structures that can carry only information that could have been generated by an X or a Y are activated by the same information. So, on the model of embodied cognition, the problem of perceptual integration is a problem of how to behave upon receiving underdetermined information. An animal's initial response is to *initiate* behaviours appropriate to the presence of both an X and a Y, because distinct neural structures that can carry only information that could have been generated by an X or a Y are *activated* by receiving the information. In determining a behaviour, the animal makes a determination of the source of the information it received. What an animal does with the information it receives determines what it takes the source of that information to be. Consider the example of a hare that receives information that could have been generated by a fox or a dog. Suppose that an appropriate behaviour for a hare in the presence of foxes is to run, whereas an appropriate behaviour for a hare in the presence of dogs is to lie still. Given the underdetermined information the hare receives, both behaviours are possible. If the interactions of neural structures result in the hare running, it takes the source of information to be a fox, lying still and it takes the source to be a dog.

Now from the embodied approach to content, an animal that initiates behaviours appropriate to the presence of an X and a Y is ascribed two hypotheses: "An X is present", and "a Y is present".[12] These hypotheses do not have the hypothetical form,

---

[12] Notice again that dispositions determine the content of hypotheses that are ascribed to the creature. Also recall (5.4) that the notion of appropriate behaviour is cashed out in terms of how the behaviour figures in intentional explanations, as Stalnaker and Dennett have argued.

"It is possible that an X (Y) is present", because behaviours appropriate to the presence of an X and a Y are actually initiated. As in the case of considered action, it is the context in which the assertion "An X is present" is ascribed that makes it a hypothesis. Mutually exclusive assertions are ascribed concerning the source of the information, in which context each assertion is a hypothesis about the source. Again on this account, the hypotheses formulated do not require that the internal representational system possess anything corresponding to the conditional, since the hypotheses do not have a hypothetical form. It could be the case, however, that a stimulus is in the extension of a concept that we would express in a natural language using a logical connective, such as RED AND TRIANGULAR. In order to correctly categorize the stimulus, the animal must behave in a way that licenses the ascription of a hypothesis that captures the logical form of the concept. Since this entails having a concept with that logical form, I defer discussion of the issue until I present an account of how animals can have concepts with a logical form.

6.5 Concept Learning

In one respect the problem facing an animal in the concept learning situation is the same as the problem of perceptual integration. The animal must categorize a stimulus based on what information the animal receives from the stimulus, and hence which of its neural structures are activated in the presence of the stimulus. The difference in the situations is that in perceptual integration a stimulus is categorized according to concepts the animal already has, whereas in concept learning it must

determine a categorization for a stimulus.[13] Now since the animal can only categorize a stimulus based on what information it receives from the stimulus, it must categorize it as a generator of certain information. A given stimulus can be categorized as a generator of information that can be carried by a particular neural structure, for each neural structure it activates. In a sense each such categorization is correct because the stimulus does, in fact, activate several neural structures. However, on the model of embodied cognition, having a concept entails more than just having a neural structure activated, in addition it requires being disposed to reliably produce a behaviour that can figure in intentional explanation when, and only when[14], stimuli in the extension of that concept are present. Now as we saw in (5.4), in a concept learning situation an animal associates a single behaviour[15] with all of the neural structures that are activated by a stimulus. But when a behaviour is associated with several neural structures, the animal

---

[13] Fodor does not always present concept learning as determining a categorization. For example, Fodor presents the following as an instance of concept learning. "*S* might be asked to sort stimulus cards into piles, where the figures on the cards exhibit any combination of the properties red and black with square and circular, but where the only correct (e.g., rewarded) sorting is the one which groups red circles with black squares" (Fodor 1975, p.35). In this case, however, since the subject is a language-user she already has the concept RED CIRCLE OR BLACK SQUARE, because she can express the concept with her verbal behaviour when, and only when, presented with red circles or black squares. Thus, she is merely learning a new behaviour to express the concept and not learning the concept. Contrast this situation to someone learning the concept of a mathematical GROUP, who initially produces no behaviour when, and only when, she is presented with a group and so is genuinely *learning* the concept.

[14] Again this claim needs to be hedged with a CP clause, for the creature might not be attending to the stimulus and so not produce the behaviour. Also the claim that a creature produces the behaviour only when a stimulus in the extension of the concept is present is contextualized to exclude only the other stimuli the creature might take to be in the extension of the concept it is learning. Outside of that context the creature can produce the behaviour.

[15] Recall (5.4) that a behaviour is a process that ends with a bodily motion.

cannot produce the behaviour when, and only when, stimuli in the extension of some concept are present. It has not discriminated the salient features of items in the extension of the concept it is learning, and so does not have the concept. Learning consists in breaking the associations between all neural structures and the behaviour, except for the neural structure that can carry only information that could be generated by items in the extension of the concept being learned. Associations between neural structures and the behaviour are broken by producing the behaviour when the neural structure that can carry only information that could be generated by items in the extension of the concept being learned is not activated, and the behaviour is not rewarded (or punished). So the hypotheses that must be ascribed to an animal in a concept learning situation are of the form, "B is an X behaviour", where B is the response behaviour and Xs are items that generate information that can be carried by a particular neural structure. Thus, the force of the hypothesis is that B is a behaviour appropriate (because it will be rewarded) in the presence of items that generate information that can be carried by some particular neural structure. If B is not rewarded when initiated by some neural structure N, the association between B and N is broken; the hypothesis "B is an X behaviour", where Xs generate information that can be carried by N is disconfirmed.[16] When learning is complete, only one neural structure initiates B, the neural structure that can carry only information that could be generated by items in the extension of the concept C being learned. B is a C behaviour.

---

[16] Of course, disconfirmation might require several disconfirming instances.

It is important to notice that the hypotheses ascribed are of the form, "B is an X behaviour", and not of the form, "Do B when, and only when, an X is present". The reason the latter cannot serve as a hypothesis is that it cannot be confirmed or disconfirmed, because it is an imperative.[17] What is intended by the imperative is that doing B when an X is present, as opposed to say a Y, will result in reward. This could be captured by the assertion, "Doing B is appropriate when, and only when, an X is present", but this assertion has the same force as "B is an X behaviour". Now it might be wondered how, on the embodied approach to content, one of these assertions is ascribed over the other. The short answer is that one is not ascribed over the other. The hypotheses ascribed to animals need not be supposed to have any internal constituent structure, so assertions that have different forms in English, but the same force, all express the content of the hypothesis ascribed to the animal. This point is crucial because the hypotheses ascribed in concept learning do have a hypothetical force. However, if the hypotheses ascribed need not have an internal constituent structure, the internal representational system need not possess anything corresponding to the conditional. I deal with this case and the problem of how animals can learn concepts with a logical form in the next section.

One final time, it is worth showing why my account does not fall prey to the disjunction problem. The worry is that the animal seems to end up with a categorization that includes everything that can generate certain information. It deems a behaviour B

---

[17] For the same reason, in considered action an animal is ascribed a hypothesis, "Doing B now is most appropriate", and not an imperative "Do B now".

to be appropriate in the presence of anything that can generate that information. The animal is wrong, however. B is only appropriate in the presence of things in the extension of some concept C. This is essential in the animal coming to associate B with only one neural structure[18] in concept learning. If B is associated with more than one neural structure, the animal has not learned a concept. The things in the extension of the concept ascribed to the animal are all and only those things in whose presence B is appropriate, because it is through reinforcement that B gets associated with just one neural structure. But because the animal is wrong in deeming B to be appropriate in the presence of anything that can generate certain information, it makes systematic errors. This is not a problem for the position, however, because animals *do* make systematic errors. The point is simply that an animal's way of detecting items in the extensions of its concepts does not determine those extensions. Furthermore, my position does not entail that animals cannot produce a particular behaviour that is appropriate in the presence of things in the extensions of different concepts; only that unless they can produce some behaviour that is appropriate to only items in the extension of one of the concepts, they have a more coarse-grained concept than we do.

6.6 Learning Concepts with a Logical Form

In the case of concept learning, it is not a straightforward matter to dismiss the possibility that the internal representational system must possess something corresponding to at least some of the logical connectives, because the concept being

---

[18] Or again, many that carry only the same information depending on how the individuation goes.

learned might have a logical form.[19] It is important to recognize, however, that the hypotheses ascribed to creatures in concept learning (or considered action or perceptual integration) need not have constituent structure.[20] That is, a representation of the concept RED AND TRIANGULAR need not be a representation of the concept RED conjoined with the concept TRIANGULAR. The latter would require a representation corresponding to conjunction, but the former would not. An animal learns the concept RED AND TRIANGULAR by confirming the hypothesis "B is a red and triangular behaviour"; i.e. by learning to produce some response when, and only when, red triangles are present. Now in order to ascribe this hypothesis to the animal, the animal must possess a neural structure that can be activated by red triangles. It is in virtue of the activation of the entire neural structure, *as a unit* initiating some behaviour, that the hypothesis is ascribed; so the hypothesis itself can have the form of a unit. The hypothesis only seems to have constituent structure because it is stated in English, which does have constituent structure; it might more appropriately be expressed "B is a red-and-triangular behaviour". To suppose it must have constituent structure is simply to beg the question. The issue is whether or not the internal representational system must be linguistic.

---

[19] Fodor also cites evidence that logically equivalent hypotheses are not learned at the same rate, suggesting that the form of the hypothesis itself is essential to learning (Fodor 1975, pp.39-41). This does not count as evidence against my position, however, because the evidence involves language-using humans as subjects. I am not denying that many concepts are learned by explicitly forming hypotheses sensitive to logical form *in natural language* and testing them. My point is simply that animals do not learn concepts this way.

[20] Schiffer 1991, pp.183-184 considers a similar move.

Now it might very well be the case that the neural structure activated by red triangles[21] is just the mutual activation of a neural structure activated by triangles and one activated by red things. The neural structures that are activated by environmental stimuli can be combinations of other neural structures that also are activated by environmental stimuli. In fact, the activation of different neural structures that constitutes a combination of those structures into a new structure can even match the way logical connectives are used to combine terms in natural languages. The connective "and" can be used to combine "red" and "triangular" to form "red and triangular". Similarly, the mutual activation of a neural structure activated by triangles and a neural structure activated by red things can constitute the activation of a neural structure activated by red triangles. Nonetheless, nothing corresponding to the logical connective conjunction need be ascribed to the animal. The hypothesis, "B is a red-and-triangular behaviour" need not be ascribed in virtue of the animal having a representation for conjunction; rather, it can be ascribed in virtue of the animal discriminating information through having neural structures activated in a way corresponding to conjunction. In general, all that is required for an animal to learn a concept having a logical form is that it can have neural structures activated in a way corresponding to that logical form. The concept is learned when that specific activation pattern of the relevant neural structures and only that activation pattern initiates the response. In that case, the hypotheses ascribed to an animal need not have a constituent structure entailing the

---

[21] I am using "a neural structure activated by red triangles" as shorthand for "a neural structure that can carry only information that could be generated by red triangles", and a similar shorthand throughout.

explicit representation of a logical connective, so the internal representational system need not possess anything corresponding to the logical connectives.

Notice that if an animal did learn the concept RED AND TRIANGULAR by forming a hypothesis that represented the conjunction, so that the internal representational system possessed something corresponding to conjunction, then any animal that could learn RED AND TRIANGULAR could learn the conjunction of any of its concepts. The reason is that conjunction has a systematic use. Suppose, for example, that a wolf can learn the concept RED AND TRIANGULAR in that it can produce some response when, and only when, red triangles are present. If the wolf's learning this concept presupposes it has something corresponding to conjunction in its internal representational system, then if it has a concept of the moon and a concept of water, it can learn the concept MOON AND WATER, by conjoining its internal representations of the moon and of water using its internal representation of conjunction. That is, it could learn to produce a response when, and only when, the moon and water are present. Now it is counterintuitive at best that a wolf's ability to produce a response in the presence of red triangles should enable it to produce a response in the presence of the moon and water. Animal behaviour does not seem to exhibit this degree of systematicity. "A dog may be able to represent the fact that there is no water in the dish and the fact that there are no people in the room but not be able to represent the fact that there are no people in the dish" (Kaye 1995, p.105).[22] Of

[22] Dennett gives an example in which we would not want to claim that an animal's thoughts were systematic. We can suppose that a gazelle could think that a lion wants to eat it, but do we then have to suppose that the gazelle can think it wants to eat the lion (personal communication).

course, animals do exhibit some systematic behaviour, which we would expect. An animal that can make some discriminations must be built in such a way that it can make others. For example, an animal that can discriminate red triangles and yellow circles can surely learn to discriminate red circles.[23] On the embodied approach to content that I have developed, systematicity can occur to greater or lesser degrees in different animals' behaviours, but need not be ubiquitous. The lack of systematicity in animal behavioural responses suggests that concepts with logical form need not be represented with a constituent structure that includes a representation of a logical connective.[24]

To complete the argument that concept learning does not require the internal representational system to possess anything corresponding to the logical connectives, we require descriptions of neural activation corresponding to each of the logical connectives. Animals that initiate a behaviour by activating neural structures according to these descriptions can learn concepts having a logical form, without requiring that the internal representational system possess anything corresponding to the logical connectives. As we have seen in the example of the red triangles, conjunction requires that the mutual activation of neural structures initiates the response, but the activation of either neural structure independently does not. In addition, the individual neural

---

[23] Fodor and Pylyshyn argue for this kind of systematicity (Fodor and Pylyshyn 1988, p.41).

[24] It is compatible with what I have argued that creatures could possess an internal representation of a limited conjunction, such as one that conjoins only 'property' concepts. In such cases however, the internal representational system does not contain something corresponding to conjunction in natural language and so need not be as expressive as natural language. But more importantly, I am not arguing that there cannot be a language of thought distinct from any natural language, only that there need not be.

structures must each initiate a behaviour inappropriate to items that activate the other neural structure; the creature must be able to discriminate the conjuncts.

An activation pattern corresponding to disjunction is given when the independent activation of distinct neural structures initiates the same behavioural response.[25] Recall (5.4) that in the initial learning phase a creature connects the behavioural response to all of the neural structures activated by the stimulus. In learning a disjunctive concept, confirming evidence for one disjunct disconfirms the other, whereas in learning a conjunctive concept, confirming evidence for the conjunction necessarily confirms both conjuncts, since the conjuncts are conjoined in the conjunction, and hence co-occur. That is, disjuncts are reinforced separately in learning a disjunction, whereas conjuncts are reinforced only when they co-occur in learning a conjunction. Thus, we would expect disjunctive concepts to be harder to learn, as Fodor claims they are. "[H]uman subjects typically have more trouble mastering disjunctive concepts than they do with conjunctive or negative ones. ... Animals, too, typically find (what *we* take to be) disjunctive concepts hard to master" (Fodor 1975, p.57, emphasis in original). In order for the concept to be truly disjunctive, and not merely a more coarse-grained concept than ours, the distinct neural structures must also initiate distinct behaviours, such that the behaviour initiated by one neural structure is inappropriate in the presence of stimuli that activate the other neural structure; that is, the animal must be able to discriminate

[25] Whether this is inclusive or exclusive disjunction depends on whether the response is produced when both neural structures are activated.

the disjuncts.[26]

An activation pattern corresponding to negation is that the activation of any neural structure not activated by stimuli in the extension of some concept C, initiates *the* response. It is crucial to emphasize that the neural structures must all initiate the same response and that it must be appropriate to the presence of items not in the extension of the concept being negated. Since appropriateness is cashed out in terms of intentional explanation, the response must be such that an explanation of the creature's behaviour in terms of its beliefs and desires includes the negation of the concept C. This places a context on the ascription of hypotheses asserting that a behaviour is appropriate in the presence of all and only those items not in the extension of concept C, thereby avoiding pan-negationism--the phenomenon that virtually any behaviour licenses the ascription of the negation of virtually any concept. That B is not a RED behaviour does not entail that B is a NOT-RED behaviour.

Activation patterns corresponding to the conditional and biconditional can be constructed out of the activation patterns corresponding to conjunction, disjunction, and negation by the basic logical equivalences. On this account, animals whose neural structures can be activated according to these descriptions or any logically equivalent

---

[26] The activation patterns for both conjunction and disjunction must be such that the creature has concepts of the individual conjuncts and disjuncts in order to be ascribed a concept that we would express as conjunctive or disjunctive. This requires that the creature can produce a behaviour--a process resulting in a bodily movement--appropriate in the presence of items in the extension of one of the conjuncts or disjuncts but not the other. Now since the notion of appropriateness is cashed in terms of intentional explanation, the bodily movements produced will generally be different in order for the behaviours culminating in those bodily movements to yield distinct intentional explanations. But the bodily movements need not be different provided the context is enough to give intentional explanations that discriminate the behaviours.

description can learn concepts with a logical form, whereas animals whose neural structures cannot be so activated cannot.[27] And since it is possible to ascribe animals concepts with a logical form, it is possible to ascribe them hypotheses of the form, "An X is present" in perceptual integration, where Xs are expressed in natural languages by using a logical connective, without requiring that the internal representational system possess anything corresponding to the logical connectives.

Now notice that having neural structures that can be activated in a way corresponding to one of the logical connectives does not entail having neural structures that can be activated in a way corresponding to any other of the logical connectives. Nor does it follow that because some neural structures can be activated in a way corresponding to a logical connective, that all of the neural structures can be systematically activated in that way can. Moreover, neural activation alone is not enough for possessing a concept, since the neural structure activated must initiate some particular behavioural response. Even if neural structures can be activated in ways corresponding to the logical connectives, it does not follow that the animal can learn a concept having any particular logical form, because it might not be the case that a behaviour that can figure in an intentional explanation can be associated with just a particular activation pattern of certain neural structures. An animal's ability to learn

---

[27] If animals can have concepts with a logical form that we describe using the logical quantifiers, there must be a description of neural activation corresponding to the quantifiers. Neural activation that would correspond to universal quantification could be given as a conjunction, within some context. Existential quantification can be given in terms of universal quantification and negation. The complexity of these activations patterns suggests that animals rarely have concepts that we describe using the logical quantifiers.

concepts having a logical form need not be systematic.

The only case remaining is to show that an animal can be ascribed a hypothesis of the form, "B is an X behaviour", without requiring that the internal representational system possess anything corresponding to the conditional. The worry here is that the hypothesis has the same force as, "Doing B is appropriate when, and only when[28], an X is present", which clearly has a biconditional structure. The hypothesis is ascribed, however, because of the internal structure of the animal and not its internal representations. When an animal learns a concept, it comes to associate a behaviour with one neural structure. The behaviour is initiated when, and only when, the neural structure is activated. But the behaviour actually is initiated by the neural structure. Thus, the animal need not represent to itself that it would initiate the behaviour when, and only when, the neural structure is activated. It will, as a matter of fact, because of the way it is built, initiate the behaviour when, and only when, the neural structure is activated. We ascribe a hypothesis with a biconditional structure to capture the internal structure, but in so doing we need not ascribe a representation of the biconditional.[29]

On the model of embodied cognition, I have given an account of how animals

---

[28] Again, the point of saying "only when an X is present" is to rule out only the alternatives for the creature in the concept learning situation. Recall (5.4) that in learning a concept, a behaviour is associated with several neural structures, only one of which carries only information that can be generated by Xs. The hypothesis that B is an X behaviour is simply the hypothesis that the behaviour is appropriate in the presence of just Xs as opposed to any of the items that activate the other neural structures associated with B in the concept learning situation. Of course, outside of this context, B might be appropriate in the sense of being reinforced and hence figuring in intentional explanations in many other contexts.

[29] Note that my position here is just another way of stating the embodied cognition thesis. The cognitive capacities of an animal are determined by its constitution.

could perform the cognitive processes of considered action, perceptual integration, and concept learning qua processes of hypothesis formation and confirmation, without requiring that they possess anything corresponding to the logical connectives in their internal representational systems. In particular, the internal representational system need not possess anything corresponding to the conditional, because some of the hypotheses ascribed to animals by the embodied approach to content do not have a hypothetical form, and those that do have a hypothetical form are ascribed in virtue of the possible activation patterns of neural structures given an animal's internal constitution, and so do not require a representation for the conditional. The reason that it seemed from our earlier discussion (4.7) and (4.8) that these processes did require the internal representational system to possess something corresponding to the conditional is that we were supposing that a process of hypothesis formation and confirmation required the explicit representation of hypothetical options. This is the manner of hypothesis formation and confirmation familiar to us from introspection, but as we have just seen it is not the only model. Since animals can engage in hypothesis formation and confirmation without explicitly representing their hypothetical options as hypothetical, or by embodying the hypothetical form, animals' competencies of considering actions, integrating perceptions, and learning concepts do not require that the internal representational system possess anything corresponding to the conditional. Nonetheless, animals represent their options in a given situation, evaluate those options, and choose what they determine to be the most preferred option; i.e. they are rational.

6.7 Early Language Acquisition

It remains to give an account of how children learn the first words of a (first) natural language without requiring that the internal representational system possess anything corresponding to the logical connectives. My argument has two steps. First, children's learning their first words is a special case of concept learning.[30] Second, in learning that producing a word is an appropriate response in the presence of items in the extension of some concept, children learn truth rules (1.7). The first point to notice is that the first words of a first natural language are taught demonstratively. Parents point to objects in the environment and say a word. A child must produce a response behaviour when, and only when, stimuli of a certain type are present, which is exactly a concept learning situation. The only thing that distinguishes this case from any other concept learning situation, is that the behavioural response is a word. So in learning a concept, the child learns that a word is appropriate behaviour when, and only when, items in the extension of that concept are present. Now in learning a concept the child confirms a hypothesis of the form "B is an X behaviour". In the case of learning first words, B is just a behaviour producing a word. For example, in learning the word "dog" a child learns ""dog" is a DOG behaviour"; that is, she learns that saying "dog" is appropriate when, and only when, dogs are present. But saying "dog" in virtue of a dog being present is just predicating DOG of that thing. So in learning the concept

[30] This claim is slightly too strong. A child might already possess the concept and learn the word as a new behaviour appropriate in the presence of items in the extension of the concept. Since the procedures involved are the same, I treat early word acquisition exclusively as concept learning in the text to simplify the presentation.

DOG with a verbal response, a child learns to predicate DOG of all and only dogs. That is, she learns "dog" is true of all and only dogs, in just the sense that she will say "dog" of some designated object, say Rex, if and only if whatever is named by Rex is in the extension of DOG. But this has exactly the form of Fodor's truth rules "$\ulcorner P_y \urcorner$ is true iff x is G' (1.7). Now since learning the first predicates of a (first) language is just a special case of concept learning, and concept learning did not require that the internal representational system possess anything corresponding to the logical connectives, neither does early language acquisition.

This completes my argument that on the model of embodied cognition and the embodied approach to content, it is possible to account for animal and preverbal human behaviour without requiring that the internal representational system possess anything corresponding to the logical connectives. But as I argued in (4.6), if the internal system need not possess anything corresponding to the logical connectives, there need not be an intentional interpretation of the transformations between internal representations. As we saw in (4.2), since it is compatible with Fodor's reasoning that the actual transformations performed in thinking occur over terms in natural languages, the internal representational system is the medium of thinking only if the intentional interpretation of transformations on terms of natural languages is derived from the intentional interpretation of transformations of internal representations, abbreviated using terms in natural languages. It follows that the internal system need not be the medium of thinking; natural languages can serve as the language(s) of thought. Also, if the internal representational system need not possess anything corresponding to the

logical connectives, it need not be as expressive as natural languages, in which case the conclusion of the anti-bootstrapping argument is false. But that conclusion is used in the only argument Fodor offers for the productivity and systematicity of the internal representational system. Hence the internal system need not be productive or systematic. Furthermore, since the systematicity exhibited in animal behaviour is limited, and can be explained in terms of the internal structure of an animal, rather than its internal representations, we have independent reason to conclude that the internal representational system need not be linguistic; there need not be a mentalese.

To conclude, I consider two phenomena that Fodor's position is unable to explain, and sketch how one might make progress in understanding these phenomena from the embodied approach to content.

6.8 Abstraction

The account I have given of early language acquisition leaves us a good way off from the capacities of competent language-users. For one thing, we do not use words when, and only when, items in the extension of a concept expressed by a word are present. In explaining how it could be that we diversify our use of words, I will suggest how it might be that we could come to use words to represent things that are not present or that we have never encountered, or that are abstract. First note, as I mentioned above (6.6) that the claim that in concept learning a behaviour is reliably produced when, and only when, items in the extension of a concept are present is too strong. This claim refers to a controlled experimental situation. Animals often manifest the same behaviour in different circumstances. All that is required is that a behaviour which

licenses the ascription of a concept is reliably produced when, and only when, items in the extension of the concept are present, *in a certain context*. In the experimental situation the context is well-defined because the stimuli are discreetly presented until the behaviour is produced reliably when, and only when, items in the extension of the concept C, being learned, are presented. Outside of that context, a manifestation of the behaviour in circumstances other than when items in the extension of C are present does not indicate that the concept is not known. To have a concept it must be possible to discriminate items in its extension. But if a behaviour is manifested in the presence of something sufficiently unrelated--in the sense of the information it generates--to items in the extension of C, then the manifestation of that behaviour is not an indication of an inability to discriminate items in the extension of C, but a different use of the behaviour; i.e., in virtue of the context in which the behaviour is manifested an intentional explanation of the behaviour would not include the concept C. The context in a natural setting is less clear but still sufficiently well-defined to distinguish cases of an inability to discriminate items in the extension of some concept from different uses of a behaviour.

Now one of the interesting features of words is that they are both behavioural responses and stimuli. So a word can be reliably produced in the presence of items in the extension of the concept it expresses[31], which determines its meaning, and it can also be produced in response to other words. For example, "dog" express the concept

---

[31] Any behaviour that licenses the ascription of a concept can be said to express the concept, though the terminology is more natural when the behaviour is the production of a word.

DOG because it is appropriate in the presence of all and only dogs. "Dog" is also an appropriate behaviour in the presence of the word "leash", however. But since dogs and tokens of the word "leash" have few salient features in common, these uses of the word "dog" are simply different uses. Note that when "dog" is used as a response to "leash", the behaviour "dog" is initiated by a neural structure that can carry only information that can be generated by "leash"es and not leashes, that is, instances of the word, not the straps one uses to control a dog. Just what words can be used in response to other words depends on the concepts that the individual words express, the nature of the world, and the grammar of the language.[32] For example, "dog" is appropriate when "leash" is uttered because "dog" expresses the concept DOG, "leash" expresses the concept LEASH, and in our world, dogs and leashes go together. Similarly for any other semantically associated items, such as fish and water, or tables and chairs. So, in virtue of the fact that words are both behaviours and stimuli, and that a single behaviour can have more than one use, we can begin to see how associations of words expressing concepts allow us to use words when items in the extension of the concept expressed by a word are not present, or have never been encountered. These words are appropriate responses to other words. Word associations aid in the construction of the semantic net that characterizes our language. That is, semantic associations between words provide structure for the rational explanation of human verbal behaviour. Of course, given the main thrust of this dissertation, learning to use the logical

[32] I don't mean to presuppose any particular linguistic theory in this dissertation. For example, the embodied cognition design I have presented might require specific elaborations to accommodate a universal grammar. I leave these issues to the linguists.

connectives, from which words can be combined to express new concepts is also integral in the construction of the semantic structure of natural languages. As the semantic structure of natural language is constructed, words function as the medium by which new concepts are constructed out of old ones. And since the internal representational system need not possess anything corresponding to the logical connectives or have relations between the internal representations corresponding to semantic associations, the semantic structure of natural languages that grounds rational explanation need not be derived from the internal representational system. So natural languages can be the medium of thought, necessarily so for abstract thought. Recalling Fodor's condition 3 on a theory of concepts (5.6), "Concepts are the constituents of thoughts and, in indefinitely many cases, of one another", we see that on my account the condition is satisfied for language-users. Since the point of the condition is to account for the concepts that language-users express, I take this as sufficient to satisfy the condition.

From the embodied approach to content, learning a natural language might facilitate abstraction in two senses. First, it might allow us to express increasingly abstract relations, by learning words to express the concepts expressed by combinations of other words.[33] This learning might also restructure the brain producing neural structures that are sensitive to information about more abstract properties, in virtue of which we could learn more abstract concepts directly. Second, from the embodied

---

[33] This is Fodor's point of using language as a mnemonic device to restrict the length of any formulae we entertain, thereby allowing us to think thoughts we could not otherwise think (Fodor 1975, pp.84-5).

approach to content, learning a natural language might allow us to represent things that are not present or that we have never encountered. Because a word can be an appropriate behaviour for another word, it can be produced even though nothing in the extension of the concept expressed by the word is present. For example, "dog" can be produced in response to the word "leash", even though no dog is present. Being able to represent things that are not present might make it possible to master the use of the logical quantifiers, which entails the ability to represent items not encountered, hypothetical items etc..

Having a suggestion as to how we might give an account of abstraction is a merit of my position over Fodor's. Fodor has difficulty with abstraction because his theory of content is causal. The theory is plausible for those things with which we can be in a causal relation. But when it is not clear that we can be in a causal relation with something, it is not clear that Fodor has much of a story. What follows is Fodor's full account of abstraction.

> *What about Predicates that Express Abstractions (like "Virtuous")?* All predicates express properties and all properties are abstract. The semantics of the word "virtuous,"[34] for example, is determined by the nomic relations between the property of being a cause of tokens of the word and the property of being virtuous. It isn't interestingly different from the semantics of "horse" (Fodor 1990, p.111).

What is interestingly different about the semantics of the word "virtuous" from the semantics of the word "horse" is that nomic relations between properties have some

---

[34] Notice that Fodor is talking about the word in a natural language and not the symbol in the language of thought.

underlying mechanism. Horses generate information about their properties, which creatures receive and can learn to discriminate. More abstract entities do not of themselves generate information. They are entailed in the information generated by other things. There is no obvious reason why creatures that have evolved neural structures that reliably causally covary with items in their environment are also sensitive to abstract relations between those items, and without an argument to that effect, there is no reason to suppose they are. On my account we can begin to see how learning a language might allow us to discriminate these abstract relations. Language-users are sensitive to abstract patterns, because they can articulate them through semantic associations of words, and through successive iterations of combining words to express new concepts and associating a single word with the combination.

6.9 How Can Thought Be Truth Preserving?

Fodor's language of thought hypothesis offers an account of what constitutes rational thought. One of the crucial features of rational thought is that true thoughts generally result in other true thoughts. Without this feature thought could not be rational, for if a creature adopts its best guess of what its best behaviour is in a given situation, but the expected outcome is in no way tied to what will be the case, nothing in what the creature is entertaining is relevant in determining how it should behave.[35] Thus, an account of how thought can be truth preserving seems essential to understanding what thought consists in. If thought is just transformations of tokens of

[35] "Should" in the sense of what will be most strongly reinforced, thereby satisfying its desires, etc..

internal representations, as Fodor claims, we want to know what makes those transformations truth preserving. What makes it the case that constraints on certain of our physical state changes are semantically evaluable as *truth preserving* transformations of internal representations? This problem is particularly difficult for Fodor because on his view it is not the case that external factors confer intentionality on the internal system, since it would follow that the intentionality of the internal system would be derived from those factors, which Fodor explicitly denies. But then, it is hard to see how it could be the case that the *operations* of the internal system reflect the world. Why should it be the case that the syntactic features of symbols in virtue of which they reliably covary with items in the world, and hence get their content on Fodor's view, are also such that the possible syntactic transformations of tokens of these symbols are truth preserving? Even if atomic symbols have semantic content because of their (causal) connections to the world, it does not follow that syntactic operations on these symbols will yield semantic truths. The syntactic operations themselves must somehow relate to the world if the expressions they generate are to represent the world.

It is worth recalling that Fodor has no independent argument that transformations in the internal system of representations are truth preserving. Following his analogy between the language of thought and machine languages in computers, Fodor's claim is simply that creatures are built in such a way that some of the physical transformations that occur in their brains have a (naturalistic) semantic interpretation as transformations of symbols constituting rational thought. "[C]omputers... just *are* environments in which the causal role of a symbol token is made to parallel the

inferential role of the proposition that it expresses.... So *if* the mind is a sort of computer, we begin to see how you can have a theory of mental processes... which explains how there could regularly be nonarbitrary content relations among causally related thoughts" (Fodor 1990, p.22-3, emphasis in original). But *we* build computers so that their causal interactions have a semantic interpretation, so if the analogy is to work we need some analogous explanation of *our* design to explain the fortuitous parallel between inferences and our neural state changes. Fodor offers no such explanation.

On the embodied approach to content, concepts are learned by associating behaviours with neural structures, *as a result of interactions with the environment.* A behaviour is appropriate in the presence of just those items in the extension of the concept C being learned, because it is reinforced in the presence of those items, and only those items, in learning C. And because the neural structure associated with the behaviour can carry only information that could be generated by items in the extension of C, the activation of the neural structure generally results in an appropriate behaviour in the circumstances a creature finds itself. Thus the computations a brain performs, activated by environmental signals carrying information, initiate behaviours appropriate to the circumstances. The mechanism by which a single behaviour emerges need not concern us, but the emergent behaviour can be determined as a function of the past successes of the behaviours initiated. Thus a creature receives information from the world, an objective commodity, which causes neural activation initiating appropriate behaviours in the circumstances, from which, based on previous experience, the

behaviour most likely to lead to the highest reward emerges. This computational process has an intentional interpretation as a discrimination of the circumstances, from which possible behaviours are determined and the behaviour deemed most appropriate is produced. The transformations are generally truth preserving because the associations of neural structures sensitive to environmental information with behaviours are determined by the nature of the environment. What a creature takes to be an appropriate behaviour generally is appropriate because of how the environment determines the circumstances in which the behaviour is produced.

Because early language acquisition is a special case of concept learning, in learning a concept by producing a word, and hence learning a truth rule, children learn what a predicate is true of. How words are semantically associated is then a function of what concepts the words express and what associations exist *in the world* between the items in the extension of those concepts. It is truths about the world that determine what words are appropriate responses to other words. Furthermore, the logical connectives by their very nature are truth functional, so if components represent truths about the world, combinations of those components using the logical connectives also represent truths about the world. Now with competent use of the connectives we can explicitly represent our behavioural options hypothetically, and if the representations are true then any new representations that are produced as an appropriate response to our hypotheses will generally be true. Just like we build computers to have a particular semantic interpretation, in learning language, we construct ourselves to have a semantic interpretation. So from the embodied approach, we can see how it might be that thought

is truth preserving transformations of representations. If one holds that rational thought requires the explicit representation of options as hypothetical in determining an action, then animals do not think rationally; however, on the model I have presented we can see the homogeneities between human thought and animal thought, and how natural language gives humans vastly superior representational capacities.

BIBLIOGRAPHY

Adams, F. and Aizawa, K. 1994. "Fodorian Semantics" in Stich and Warfield 1994, pp.223-242.

Armstrong, D. 1973. *Belief, Truth and Knowledge*. Cambridge, MA: Cambridge University Press.

Brooks, R. 1991. "Intelligence without representation", *Artificial Intelligence 47*, pp.139-159.

Burge, T. 1986. "Individualism and Psychology", *Philosophical Review 95*, pp.3-46.

Burge, T. 1993. "Mind-Body Causation and Explanatory Practice" in Heil and Mele 1993, pp.97-120.

Chomsky, N. 1959. "A Review of B.F. Skinner's *Verbal Behaviour*" in *The Structure of Language Readings in the Philosophy of Language*, edited by Fodor, J. and Katz, J. 1964. Englewood Cliffs, NJ: Prentice-Hall Inc..

Clark, A. 1995. "Connectionist Minds" in Macdonald and Macdonald 1995, pp.339-356.

Clark, A. 1997. *Being There: Putting Brain, Body, and World Together Again*. Cambridge, MA: Bradford/The MIT Press.

Davidson, D. 1970. "Mental Events" in Davidson 1980, pp.207-227.

Davidson, D. 1980. *Essays in Actions and Events*. Oxford: Clarendon Press.

Davidson, D. 1984. *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press.

Davidson, D. 1993. "Thinking Causes" in Heil and Mele 1993, pp.3-18.

Dahlbom, B. 1993. *Dennett and His Critics*. Cambridge, MA: Blackwell.

Dennett, D. 1978. *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge, MA: Bradford/The MIT Press.

Dennett, D. 1987. *The Intentional Stance*. Cambridge, MA: Bradford/The MIT Press.

Dennett, D. 1991. *Consciousness Explained*. Boston: Little Brown and Company.

Dennett, D. 1995. *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. New York: Simon and Schuster.

Dennett, D. 1996. *Kinds of Minds: Towards an Understanding of Consciousness*. London: Weidenfeld and Nicolson.

Dretske, F. 1981. *Knowledge and the Flow of Information*. Cambridge, MA: Bradford/The MIT Press.

Dretske, F. 1988. *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: Bradford/The MIT Press.

Field, H. 1978. "Mental Representation", *Erkenntnis 13*, pp.9-61.

Fodor, J. 1975. *The Language of Thought*. Cambridge, MA: Harvard University Press.

Fodor, J. 1981. *Representations: Philosophical Essays on the Foundations of Cognitive Science*. Cambridge, MA: Bradford/The MIT Press.

Fodor, J. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: The MIT Press.

Fodor, J. 1990. *A Theory of Content and Other Essays*. Cambridge, MA: Bradford/The MIT Press.

Fodor, J. 1994. *The Elm and the Expert: Mentalese and Its Semantics*. Cambridge: Bradford/The MIT Press.

Fodor, J. 1998. *Concepts: Where Cognitive Science Went Wrong*. New York: Oxford/Clarendon Press.

Fodor, J., Garrett, M., and Brill, S. 1975. "Pe, ka, pu: the perception of speech sounds in prelinguistic infants", *MIT Quarterly Progress Report*, January 1975.

Fodor, J. and Lepore, E. 1993. "Is Intentional Ascription Intrinsically Normative?" in Dahlbom 1993.

Fodor, J. and Pylyshyn, Z. 1988. "Connectionism and Cognitive Architecture: A Critical Analysis", *Cognition 28*, pp.3-71.

Harman, G. 1973. *Thought*. Princeton: Princeton University Press.

Heil, J. and Mele, A. (eds.) 1993. *Mental Causation*. New York: Oxford/Clarendon Press.

Holland, J. 1975. *Adaptation in Natural and Artificial Systems*. Ann Arbor: The University of Michigan Press.

Holland, J. 1992. "Genetic Algorithms", *Scientific American, July 1992*, pp.66-72.

Holland, J., Holyoak, K., Nisbett, R., and Thagard, P. 1987. *Induction: Processes of Inference, Learning, and Discovery*. Cambridge, MA: The MIT Press.

Hornsby, J. 1993. "Agency and Causal Explanation" in Heil and Mele 1993, pp.161-188. Reprinted in Hornsby 1997, pp.129-153.

Hornsby, J. 1997. *Simple Mindedness: In Defense of Naive Naturalism in the Philosophy of Mind*. Cambridge, MA: Harvard University Press.

Kaye, L. 1995. "The Languages of Thought", *Philosophy of Science 62*, pp.92-110.

Loewer, B. and Rey, G. (eds.) 1991. *Meaning in Mind: Fodor and his Critics*. Oxford: Blackwell.

Macdonald, C. and Macdonald, G. (eds.) 1995. *Connectionism: Debates on Psychological Explanation*. Oxford: Blackwell.

Maes, P. 1994. "Modeling Adaptive Autonomous Agents", *Artificial Life 1*, pp.135-162.

Millikan, R. 1998. "A common structure for concepts of individuals, stuffs, and real kinds: More Mama, more milk, and more mouse", *Behavioral and Brain Sciences 21*, pp.55-100.

Pietroski, P. (forthcoming). *Causing Actions*. New York: Oxford University Press.

Putnam, H. 1975. *Mind, Language and Reality*. Philosophical Papers II. New York: Cambridge University Press.

Putnam, H. 1988. *Representation and Reality*. Cambridge, MA: The MIT Press.

Ryle, G. *1949. The Concept of Mind*. Chicago: The University of Chicago Press.

Schiffer, S. 1991. "Does Mentalese Have a Compositional Semantics?" in Loewer and Rey 1991, pp.181-199.

Sellars, W. 1956/1997. *Empiricism and the Philosophy of Mind*. Cambridge, MA: Harvard University Press.

Stainton, R. 1996. *Philosophical Perspectives on Language*. Peterborough: Broadview Press.

Stalnaker, R. 1984. *Inquiry*. Cambridge, MA: Bradford/The MIT Press.

Stich, S. and Warfield, T. 1994. *Mental Representation: A Reader*. Oxford: Blackwell.

Wason, P. and Johnson-Laird, P. 1972. *Psychology of Reasoning: Structure and Content*. Cambridge, MA: Harvard University Press.