# A NATURALISTIC ACCOUNT OF MENTAL REPRESENTATION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

TEVFİK AYTEKİN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
THE DEPARTMENT OF COGNITIVE SCIENCE

FEBRUARY 2007

Approval of the Graduate School of Informatics

_____

Assoc. Prof. Dr. Nazife Baykal
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Doctor of Philosophy.

_____

Prof. Dr. Deniz Zeyrek
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Doctor of Philosophy.

_____

Assoc. Prof. Dr. Erdinç Sayan
Supervisor

Examining Committee Members

| | | |
|---|---|---|
| Prof. Dr. Wolf König | (METU, COGS) | _____ |
| Assoc. Prof. Dr. Erdinç Sayan | (METU, PHIL) | _____ |
| Assoc. Prof. Dr. Cem Bozşahin | (METU, COGS) | _____ |
| Assist. Prof. Dr. Joshua D. Cowley | (BİLKENT, PHIL) | _____ |
| Assist. Prof. Dr. Bilge Say | (METU, COGS) | _____ |

**I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.**

**Name, Last name : Tevfik Aytekin**

**Signature : _____**

## ABSTRACT

A NATURALISTIC ACCOUNT OF MENTAL REPRESENTATION

Aytekin, Tevfik
Ph.D., Department of Cognitive Science
Supervisor: Assoc. Prof. Dr. Erdinç Sayan

February 2007, 102 pages

My thesis is an attempt to develop a naturalistic account of mental representation based on the notion of causation. The thesis consists of two main parts. The first part (chapters II and III) develops an understanding of naturalization. According to my proposal, naturalization is a two-step process: in the first step a set of conditions is specified which are thought to be the essential aspects of the notion under study and in the second step a naturalistic system is proposed which is claimed to satisfy these conditions. In accordance with this understanding of naturalization, the second part (chapters IV and V) of the thesis sets out the conditions which a successful naturalization of mental representation has to satisfy and then develops a new naturalistic account of mental representation based on the causal connections between environmental properties and the brain.



Keywords: Mental representation, intentionality, naturalization.

# ÖZ


ZİHİNSEL TEMSİLİN DOĞALCI AÇIKLAMASI

Aytekin, Tevfik
Doktora, Bilişsel Bilimler Bölümü
Tez Yöneticisi: Doç. Dr. Erdinç Sayan

Şubat 2007, 102 sayfa

Bu tez nedenselliğe dayalı doğalcı bir zihinsel temsil açıklaması geliştirme girişimidir. Tez iki ana kısımdan oluşmaktadır. İlk kısımda (bölüm II ve III) bir doğallaştırma anlayışı geliştirilmiştir. Buna göre doğallaştırma iki adımlı bir süreçtir: ilk adımda üstünde çalışılan kavramın başlıca özellikleri olduğu düşünülen şartlar belirlenir ve ikinci adımda bu şartları sağlayan doğalcı bir sistem önerilir. Doğallaştırmanın bu anlayışına uygun olarak tezin ikinci kısmında (bölüm IV ve V) zihinsel temsilin başarılı olarak doğallaştırılmasını gerçekleştirmek için gerekli şartlar belirlenmiş ve zihinsel temsilin beyin ve çevre arasındaki nedensel ilişkilerine dayanan doğalcı bir açıklaması verilmiştir.



Anahtar Kelimeler: Zihinsel temsil, yönelimsellik, doğallaştırma.

## ACKNOWLEDGMENTS

I would like to thank the following: Assoc. Prof. Dr. Erdinç Sayan, my supervisor, for his guidance, support, and expert advice; Assoc. Prof. Dr. Samet Bağçe, Assoc. Prof. Dr. Cem Bozşahin, Assist. Prof. Dr. Bilge Say, Prof. Dr. Wolf König, Prof. Dr. Deniz Zeyrek, Assist. Prof. Dr. David Davenport, and Assist. Prof. Dr. Joshua D. Cowley, my thesis committee members, for their helpful comments and criticisms; Ayşenur, my wife, for her patience and understanding; and finally my parents and my sister, for their support throughout my research.

## TABLE OF CONTENTS

# CHAPTER I

## INTRODUCTION

This work is primarily an attempt to develop a new naturalistic framework for the semantics of intentional states. Intentional states are such states of the mind which are denoted by the terms such as, "believe," "desire," "think," "fear," "hope," etc. It is generally accepted that such states have semantic (or intentional) properties very much like natural language expressions; they are about objects, events, states of affairs, etc., beyond themselves. To explain the nature of intentional states (especially their semantic properties) is one of the central areas of study in philosophy of mind.

One specific project about intentional states is to naturalize their semantic properties. Briefly, naturalization project aims to explain the nature of the semantics of intentional states by using only naturalistically respectable terms (and without using any semantic or intentional terms). To put this in other words, the naturalistic project seeks to find a place in nature for the semantic properties of intentional states or, as it is sometimes stated, to *reduce* semantic properties to naturalistic properties. A question here is what constitutes naturalistic properties. The shortest answer I can give at this point is that naturalistic properties are such properties as those that figure in the explanations of natural sciences such as physics, chemistry, or biology. Psychology, which invokes intentional terms, is not considered to be a branch of the natural sciences with respect to the aims of this project.

Although intentionality (that is, aboutness) is the most important property of intentional states, it does not constitute everything about intentional states. To see where exactly the naturalization problem occurs we need a theory of intentional states. In this thesis the framework within which I will discuss the naturalization problem will be Fodor's representational theory of mind. This theory construes intentional states as relations to mental representations. For example, the belief that the cat is on the mat is construed as having a mental representation which *represents* that the cat is on the mat and which plays the functional role of a belief state. The mental representation which represents that the cat is on the mat is further construed as a complex representation which is built out of primitive mental representations. Primitive mental representations are usually thought to be concepts which correspond roughly to lexical items in language such as "cat," "tree," "rock," "chair," "water," etc. Fodor thinks that the semantic properties of complex mental representations are derived from the semantic properties of primitive mental representations compositionally. So, the task of the semantic naturalist is primarily to naturalize the semantic properties of these primitive mental representations.

The use of the term "representation" in this context might be confusing. This term is also widely used in theories of cognitive science. It should be noted that what Fodor is primarily interested in is not the concept of representation as it is used in the scientific theories of cognitive science, but the folk concept of representation as it occurs in folk psychological explanations. Some concepts are used both scientifically and by the folk. For example, take the concept of disease. It is clear that the scientific concept of disease and the folk concept of disease are different things. One might be interested in the concept of disease as it is used scientifically or one might be interested in the concept of disease as it is used by the folk. Of course, the term "representation" does not directly occur in folk psychology but it is the representational theory of mind which links the two. So, what I will be mainly interested in will be the folk concept of representation. But this does not mean that the concept of representation as it is used in cognitive science is totally irrelevant to my concerns (and also to Fodor's). Since I think

that the folk and the scientific concept of representation overlap in many interesting ways, the scientific concept of representation will be helpful to develop my proposal as will become clear in the chapters to follow.

Another point that I want to mention is the relation between the semantics of mental representations and the semantics of natural language expressions. The former is usually a subject for philosophers of mind while the latter is usually a subject for philosophers of language. Nevertheless the two issues are closely related. There are two central questions regarding the semantics of natural language expressions: (1) What is the reference of a term? (2) In virtue of what is the connection between a term and its reference established? Analogous questions can be asked about mental representations. What is the reference of a mental representation? (2) In virtue of what is the connection between a mental representation and its reference established? The natural language expressions considered for these questions are generally centered on two types: proper names (such as "Aristotle," "Ankara," "Pegasus," etc.) and natural kind terms (such as "water," "cat," "gold," etc.). In the discussions on the semantics of mental representations, on the other hand, the major focus is on natural kind terms. It is assumed that there are mental representations which represent natural kinds. So, in this respect the answers to the two questions for both natural language expressions and mental representations are quite related. The major difference is that philosophers of language may freely invoke intentional states in their explanations. For example, they may account for how the connection between a term and its reference is established by invoking the beliefs and desires of a speaker. But of course this is not possible for the theorist whose aim is to explain the nature of mental representations which are supposed to constitute intentional states.

My contribution in this thesis consists of two parts. My first contribution is to develop an understanding of naturalization. According to me, naturalization consists of two steps. In the first step the analysis of the notion under study (such as "representation") is given by conceptual means. And in the second step, a naturalistic system is proposed which purports to satisfy the conditions set out in

the first step. My second contribution is to propose a naturalistic account of representation. To do this, in accordance with my understanding of naturalization, I first analyze the notion of "representation" as a set of conditions and, secondly, I describe a naturalistic system which is claimed to satisfy these conditions.

The structure of this thesis is as follows: In chapter II, I discuss the nature of folk psychology and intentional states conceived of as theoretical entities. I also discuss the possible philosophical positions concerning intentional states, such as, eliminativism and intentional realism. In chapter III, I try to develop an understanding about naturalization. In particular, I deal with what naturalization amounts to, the criteria of a successful naturalization, and naturalization as a reply to eliminativism. In chapter IV, I introduce Fodor's representational theory of mind. I also discuss two naturalization attempts based on the notions of causation and similarity. In discussing these theories I will give some of the important conditions which a successful naturalization attempt has to satisfy. In chapter V, I develop my own naturalistic proposal for an account of representation and evaluate its success with respect to the conditions that are given in the previous chapter. Finally, in chapter VI, I discuss some further issues which are related to the naturalization project and which can be raised as objections to my account.

# CHAPTER II

# FOLK PSYCHOLOGY

## 2.1 What is Folk Psychology?

Folk psychology (FP hereafter) can be defined as the set of statements ordinary people use to describe, explain, and predict human behavior by positing mental states such as believing, desiring, having pain, etc. The statements used in FP can be grouped in three types: the first type links external stimulus with mental states, the second type links mental states with other mental states, and the third type links mental states with behavior. For example, the statement "She felt great pain when she put her hand on the stove" links an external stimulus (putting her hand on the stove) with a mental state (feeling pain), the statement "Since he thinks that he will not be successful in the exam tomorrow, he hopes that the exam will be postponed" links a mental state (believing that he will not to be successful in the exam tomorrow) with another mental state (hoping that the exam will be postponed), and finally the statement "Since he believes that it will rain, he took his umbrella with him" links a mental state (believing that it will rain) with behavior (taking the umbrella with him).

To develop an understanding of the nature of FP seemed important for many philosophers; at least, for understanding the nature of mind. One influential construal of FP, initially proposed by Sellars (1956) and popularized by Lewis (1972), views FP as a folk theory. Sellars imagines a myth which explains how FP developed. In the first stage of the myth, to explain others' behavior, people

were using terms only for public and spatiotemporal properties. That is, no terms were used designating mental states. But in the second stage, they developed their theory so that it allowed using terms which refer to mental states or inner episodes. They introduced mental state terms probably because they could then make better explanations and predictions. And later they learnt to apply the theory to themselves. Sellars' aim was primarily to attack the view that the statements we use about our own mental states were due to a private and introspective ability of our minds. But his attack resulted in an original construal which basically claims that FP is just like any other empirical theory that humans developed to explain and predict natural phenomena.

Simply stated, an empirical theory is a set of empirical laws for explaining and predicting certain phenomena. Just as classical mechanics is a set of laws for explaining and predicting the motions of bodies, FP is a set of laws for explaining and predicting the behavior of other people. I have given examples of these laws (or statements) above. FP is a folk theory because, different from classical mechanics, it was not developed explicitly as a scientific theory. But this characteristic of FP does not preclude considering it as a genuine empirical theory, for, like scientific theories, it involves a set of laws which is used to predict and explain certain phenomena, namely, the behavior of people.

### 2.2 Theoretical Terms

The nature of theoretical terms is a widely discussed topic especially in philosophy of science. One influential elaboration of theoretical terms in the context of philosophy of mind is given in Lewis (1970) and Lewis (1972). Lewis' analysis primarily aims to understand the nature of mental states construed as theoretical terms embedded in a theory. To discuss the nature of theoretical terms Lewis (1972) gives the following example (a *theory* proposed by a detective to be the best explanation of a murder case):

> X, Y and Z conspired to murder Mr. Body. Seventeen years ago, in the gold fields of Uganda, X was Body's partner … Last week, Y and Z

> conferred in a bar in Reading … Tuesday night at 11:17, Y went to the attic and set a time bomb … Seventeen minutes later, X met Z in the billiard room and gave him the lead pipe … Just when the bomb went of in the attic, X fired three shots into the study through the French windows … (pp.249-50):

Lewis dubs the terms X, Y, and Z as theoretical terms (T-terms) and the rest of the terms in the story as old or original terms (O-terms). Lewis warns us that O-terms have nothing to do with observational terms; similarly T-terms should not be thought of as unobservable terms. He states that he is not making an observable/unobservable distinction which is notoriously problematic. The T-terms should be thought of as *new* terms introduced by a theory.

Lewis argues that the meanings of the T-terms are definable functionally or by their causal roles. That is, the meaning of T-terms in the story can be defined as the occupants of the causal roles as specified by the story. For example, the meaning of X can be defined as the entity which conspired to murder Mr. Body and who fired three shots into the study through the French windows and so on. The position of Lewis is very close to what is known as functionalism in philosophy of mind. According to functionalism the essence of mental states is functional: what makes something a particular mental state is its functional (causal) relations to external stimulus, other mental states, and behavior.

Let me give two analogies useful in grasping the notion of a theoretical term.

Consider Mendelian genetics. Mendel, in order to explain inheritance, posited genes in his theory of genetics. Before the identification of genes with segments of DNA molecules, if someone asked the question "What are genes?" the Lewisian answer would be like this: "Genes are the occupants of the causal roles specified by Mendelian genetics." Later when DNAs were discovered, genes were identified as the segments of the DNA molecule. What makes this identification possible is the fact that DNA segments occupy (roughly) the same causal roles as genes, which are defined implicitly in Mendelian genetics.

I think that this notion of theoretical term is not limited to empirical theories. Lewis' understanding of T-terms and O-terms is quite similar to that of

undefined terms of axiomatic systems. The undefined terms used in the axioms of an axiomatic system can be grouped in two categories (Wilder, 1965): undefined technical terms and undefined universal terms. For example, in an axiomatic system of Euclidian geometry, *point* and *line* are the undefined technical terms and terms such as *collection*, *there exist*, *every*, *not*, *at least*, etc., are the undefined universal terms. The difference between these two categories is that the undefined universal terms are assumed to have a universally known fixed meaning. Whereas the meanings of undefined technical terms are not like this; we are free to give any meaning to (or interpret) them, given that the interpretations are compatible with the meanings of these terms defined implicitly by the statements made in the axioms. That is, the axiomatic statements implicitly define a meaning for the undefined technical terms. I think that we can make this useful analogy here: the semantics of undefined technical terms is like that of the notion of T-terms and the semantics of undefined universal terms is like that of the notion of O-terms as introduced by Lewis.

### 2.3 Eliminativist Argument Type I

To understand FP as a theory is a useful suggestion. But if FP is a theory then it is possible that FP is a false theory. Eliminativism with regard to the mind claims that mental states[1] do not exist. In other words mental state terms intend to refer but what they intend to refer to does not exist. In this section we will look at the eliminativist thesis in some detail.

I will discuss the thesis of eliminativism as developed by Churchland (1981) who is one of its principal defenders. The argument in that paper can be stated as follows:

---

[1] Mental states are usually divided into two groups: intentional states and phenomenal states. Churchland is explicit that he is questioning the existence of intentional states. It might be possibe to extend the same argument to phemomenal states also. However, most philosophers think that phenomenal states have an essential subjective component so that their semantics cannot be captured as theoretical terms.

Premise 1: FP is an empirical theory.

Premise 2: The semantics of the terms of FP should be understood along the same lines as the semantics of theoretical terms.

Premise 3: FP is a false theory.

Conclusion: Items in the ontology of FP do not exist.

In the preceding sections we have seen that the first two premises are also shared by other philosophers. In fact these two premises form a common ground for philosophers who have opposite views (eliminativists and realists) to discuss the fate of FP. Similar to Lewis, Churchland claims that the construal of FP as an empirical theory "entails that the semantics of the terms in our familiar mentalistic vocabulary is to be understood in the same manner as the semantics of theoretical terms generally: the meaning of any theoretical term is fixed or constituted by the network of laws in which it figures" (p.69).

So Churchland thinks that if he can show that Premise 3 is true then together with premise 1 and premise 2, that will imply eliminativism, that is, the ontology of FP does not exist.

Churchland puts forward three claims to show that premise 3 is true. The first one considers the lack of explanatory success of FP. As Churchland states, "when one centers one's attention not on what FP can explain, but on what it cannot explain or fails even to address, one discovers that there is a very great deal." (p.73). According to him, FP cannot explain the mental phenomena such as the faculty of imagination, nature of sleep, perceptual illusions, and dynamics of mental illness. His second claim considers FP in comparison to respectable sciences. As Churchland states, "we must evaluate FP with regard to its coherence and continuity with fertile and well-established theories in adjacent and overlapping domains—with evolutionary theory, biology, and neuroscience, for example—because active coherence with the rest of what we presume to know is perhaps the final measure of any hypothesis." (p. 73). In this respect, according to him, although many areas of the sciences, such as biology,

neuroscience, organic chemistry, particle physics, etc., show a great deal of coherence and integration, FP is not part of this picture. Its categories stand alone and have very little chance, if any, for reduction or integration with the rest of the sciences. Thirdly, considered from a historical perspective, Churchland claims that FP shows a declining character. Ancient people were explaining many phenomena, such as the motion of the stars, natural disasters, etc., by appealing to the anger, love, desires, of the stars, gods, etc. But all of these explanations were eliminated throughout the history as the sciences developed. Now, intentional explanations in terms of anger, love, desire, etc., are only used to explain the behavior of humans. That is, FP's application area has been narrowed down. And furthermore, this narrowed FP has remained almost the same for more than two millennia. Borrowing a term from Imre Lakatos, Churchland calls this situation of FP a "stagnant" or "degenerating research program." According to Churhland FP will share the fate of eliminated theories of the past such as the phlogiston theory of combustion, Aristotelian cosmology, or the vitalist conception of life.

Churchland thinks that if FP had been a true theory it would have explained many other important mental phenomena, would have shown a degree of coherence and integration with the adjacent sciences, and would have developed to a greater extent by now. Since none of these has happened, FP is likely to be a false theory: that is, its T-terms are not likely to be realized.

### 2.4 Eliminativist Argument Type II

It is important to see that Churchland's argument given in section 2.2 did not claim to show that intentional mental states cannot exist in any physically possible world. The argument tries to show that intentional states most likely do not exist in *this* world. Another way to argue that intentional states do not exist might be like this. One can claim that FP as a theory cannot be realized in any physically possible world, i.e., there can be no set of physical entities, in this world or in any physically possible world, which satisfy the roles of intentional states set by FP.

Fodor (1987) thinks that this is, in fact, the main motivation for being an eliminativist regarding the intentional states:

> … the deepest motivation for intentional irrealism derives not from such relatively technical worries about individualism and holism as we've been considering, but rather from a certain ontological intuition: that there is no place for intentional categories in a physicalistic view of the world; that the intentional can't be *naturalized.* (p.97, emphasis his)

I think Fodor is right in his suspicion. Many philosophers have thought that intentional states resist integration in the natural order. The reason many philosophers have expressed such a worry is that intentional states possess properties which seem not to be possessed by any other physical thing. One such property is called intentionality. This property refers to the fact that intentional states are *about* other things beyond themselves. For example, when we ascribe to a person the belief that it will be sunny tomorrow, we ascribe to him an intentional state (believing in this case) which is about the proposition "Tomorrow it will be sunny."

Note that this second form of argument is stronger than the first one. For if something does not exist in any physically possible world then it cannot exist in this world, but the reverse is not true. Let me give some examples to make this point clear. We are all capable of entertaining quite many different types of concepts, such as, TREE, CHAIR, BEAUTY, NUMBER, GOOD, DEMOCRACY, etc.[2] It is evident that our just having a concept does not imply that there exists something which falls under that concept. For example, we can have the concepts of UNICORN, PEGASUS, or GOLDEN MOUNTAIN. But, to the best of our knowledge, we know that there are no such things, that is, they do not exist in this world. However, in some physically possible world a unicorn or a golden mountain might exist. On the other hand, some of our concepts correspond to entities which (presumably) cannot exist in any physically possible

[2] Following the convention of Fodor, I will hereafter denote the names of concepts in capital letters.

world at all. For example, GHOST, MIRACLE, EVIL, WITCHCRAFT, SANTA CLAUS, etc., are examples of concepts in this category. The reason why there cannot be anything which satisfies the conditions of a concept of this kind is that it is highly improbable, given our best sciences, that any physical entity fulfills the conditions for such concepts. For example, according to our common sense conception of ghosts, ghosts can pass through walls, but given our current knowledge of physics, this is something impossible.

According to this second type of eliminativism, intentional states such as beliefs, desires, etc., have similar status as the concepts such as GHOST, MIRACLE, EVIL, etc. So, in this respect naturalization can be understood as a project which tries to show that this second type of eliminativist claim is false. In order to do this, naturalists try to specify naturalistic systems (that is, systems whose descriptions involve naturalistically respectable terms) which realize the theoretical terms of folk psychology. If this project succeeds then naturalists would have shown that intentional states are compatible with the natural order or, in other words, in a physically possible world intentional states can exist.

Note that even if the naturalization project succeeds, that will not show that intentional states exist in this world. It might turn out that the status of intentional states is similar to the status of the concepts such as UNICORN, PEGASUS or GOLDEN MOUNTAIN. That is, the first type of eliminativist argument is not affected by a successful naturalization whose aim is only to show that the existence of intentional states is just a possibility. In order to show that intentional states exist in this world one needs to show that the naturalistic system, which is proposed to fulfill the conditions for being an intentional state, is realized in the brain. And I think this is where the philosophical and scientific research on the human mind coincides. For, in order to show that the proposed naturalistic system is realized in the brain the naturalist needs to appeal to the scientific theories of the brain.

## 2.5 Possible Positions

As I stated before, intentionality is the property of mental states to be about something beyond themselves. Modern discussions of intentionality usually cite Brentano as the philosopher who attracted attention to this characteristic of intentional states and who, famously, claimed that intentionality is the mark of the mental and no physical phenomena can exhibit such a characteristic. The following paragraphs quoted from Brentano (1874) nicely summarize his understanding of intentionality and his dualist attitude, respectively.

> Every mental phenomenon is characterized by what the Scholastics of the Middle Ages called the intentional (or mental) inexistence of an object, and what we might call, though not wholly unambiguously, reference to a content, direction toward an object (which is not to be understood here as meaning a thing), or immanent objectivity. Every mental phenomenon includes something as object within itself, although they do not do so in the same way. In presentation, something is presented, in judgment something is affirmed or denied, in love loved, in hate hated, in desire desired and so on.
>
> This intentional inexistence is characteristic exclusively of mental phenomena. No physical phenomenon exhibits anything like it. We can, therefore, define mental phenomena by saying that they are those phenomena which contain an object intentionally within themselves (pp. 88-89).

In the first paragraph Brentano defines the notion of intentionality. Although this paragraph is not wholly clear, according to the notion of intentionality accepted by most of the contemporary philosophers, intentionality can be explicated by an analogy to natural language expressions. It is thought that the class of things which exhibit aboutness includes words, sentences, signs, or symbols in general. For example the word "Aristotle" is simply a string having certain syntactic properties (such as consisting of nine letters). But the same string has also a semantic property, namely, it is about one of the great philosophers of ancient times. But if natural language expressions also exhibit intentionality then one can object to Brentano by noting that intentional states are not the only things that exhibit intentionality, and that intentionality is also

possessed by natural language expressions. But this move generally is thought to be hopeless, for natural language expressions have only derived intentionality (Dennett and Haugeland, 1987). In other words, natural language expressions do not have original (or intrinsic) intentionality. The word "Aristotle" is just a string which has a certain shape. Only thanks to our minds that string possesses the power of aboutness. If there were no minds then "Aristotle" could not be about Aristotle. But it is generally accepted that our minds have original intentionality.

Now given these considerations of intentionality the possible positions are as follows. One might be a dualist (e.g., Brentano and Descartes) and hold that the mental and the physical are different realms and they cannot be unified. But today most philosophers assume physicalism (except may be on the question of phenomenal states) and reject dualism. If one assumes physicalism then one has three main alternative positions on the nature of intentional states.

One might be an eliminativist (e.g., Churhland, 1981, Stich, 1983) and hold that intentional states do not exist. Note that eliminativism is a very radical position. It is intuitively very hard to accept that we have no beliefs, desires, etc. Moreover although it is true that the physical sciences (such as physics, chemistry, biology, etc.) do not invoke intentional notions in their explanations, most of the higher-level sciences (such as economics, politics, sociology, history, and of course, psychology) invoke intentional notions widely in their explanations. Eliminativism regarding intentional states also jeopardizes the explanations these sciences make.

The second possibility is called intentional realism (Fodor, 1987, 1990a). The philosophers in this camp hold that the human mind is part of the nature. The human mind is just a complex physical system (just as a tree or a rock) and while it is true that intentionality is only exhibited in humans (or may be in some other animals) this does not show that it cannot be exhibited by any complex physical system. Fluidness is exhibited only by liquids but this does not show that liquids are something superphysical. The difficult task facing these philosophers is to show how something physical can be intentional.

Eliminativism and realism do not exhaust all the possibilities. There are different versions of eliminativism and realism. Some philosophers (e.g., Dennett, 1971) hold a third position called instrumentalism. According to these philosophers intentional states do not exist but this does not imply that the explanations invoking intentional notions are useless. These philosophers try to make sense of the heavy use of intentional notions in our everyday life and in the social sciences without committing themselves to their reality. But as Fodor emphasizes, the difficulty for these philosophers (just as for the philosophers who are instrumentalists regarding scientific theories in general) is to explain how intentional explanations seem to work if intentional states are not real.

The possible positions on intentionality are not limited to the positions that I have mentioned. Fodor (1990b) gives a useful introduction to other various possible philosophical positions.

### 2.6 More on Elimination, Vindication, and Reduction

Some clarifications on the notions of elimination, vindication, and reduction are appropriate at this point. One important distinction to note is that elimination is not reduction. When some upper level theory (such as the kinetic theory of gases) is reduced to a lower level theory (such as statistical mechanics), the terms and laws used in the upper level theory are not thereby eliminated but are vindicated. That is, the ontology of the upper level theory is as real as the ontology of the lower level theory. For example, according to the identity theory of mind, mental states are brain states. If the identity theory is true then it will be possible to reduce mental states to brain states. But, again, that will not show that mental states do not exist; instead that will vindicate that mental states as known by commonsense are real states of the brain. To give another example, with the advance of molecular biology when genes were identified with (or reduced to) segments of DNA, that had shown that genes, as mentioned in Mendelian genetics, refer to real things. Eliminativists are not reductionists; according to eliminativists the reduction of intentionality is out of the question, since you cannot reduce something which does not exist.

Another important point about elimination and vindication is that deciding whether some theory is vindicated or eliminated is not a yes or no matter. Like many other concepts the concepts of vindication and elimination admit of degrees and vagueness. Lewis (1972, p.251) states this complication in connection with a detective's theory as follows: "What if the theorizing detective has made one little mistake? … We can say that the story as told is *nearly realized*." As Lewis admits the notion of near-realization is hard to analyze. But in reality this is almost always the case. For example, it is generally accepted that Mendelian genetics is a typical example of a theory which is vindicated and the phlogiston theory is a typical example of a theory which is eliminated. However, this does not mean that all parts of the Mendelian genetics are vindicated. What is important is whether the *essential* parts of a theory are vindicated or not. But again, this does not help to remove the vagueness of the concept of vindication, since being essential is itself a vague notion which depends on context and is interest-relative.

However, the element of vagueness does not preclude the possibility of discussing whether some theory is vindicated or eliminated. We just need to keep in mind that some vagueness is inevitable when discussing whether a theory is vindicated or eliminated.

# CHAPTER III

# NATURALIZATION

## 3.1 What is Naturalization?

In this section I will discuss in some detail what the naturalization project amounts to. Especially, I want to consider Fodor's (one of the leading naturalists) understanding of it. Below is a passage from Fodor in which he defines the project of naturalization:

> The worry about representation is above all that the semantic (and/or the intentional) will prove permanently recalcitrant to integration in the natural order…. What is required to relieve the worry is therefore, at a minimum, the framing of *naturalistic* conditions for representation. That is, what we want at a minimum is something of the form *'R represents S' is true iff C* where the vocabulary in which condition C is couched contains neither intentional nor semantic expressions. (Fodor, 1990c, p.32, emphasis his).

Although Fodor does not state it explicitly, it seems that in the passage above Fodor wants a conceptual analysis.[3] As traditionally conceived, philosophers engaged in conceptual analysis try to define the meaning of a concept under study with other terms (or concepts). This is not an easy task.

[3] In fact, for a number of reasons (such as Quine's criticisms of the analytic/synthetic distinction and Fodor's conceptual atomism) Fodor is opposed to conceptual analysis. But I think that Fodor can't escape from conceptual analysis even though he does not explicitly state it.

Even if we use a concept in our daily lives quite easily, when it comes to define it, it is not easy to give a satisfactory definition. This is especially true for abstract concepts such as KNOWLEDGE, FREE WILL, CAUSATION, BEAUTY, MORAL EVIL, etc. Throughout the history of philosophy many philosophers tried to give an analysis of such concepts. The method employed in conceptual analysis is usually this: First a definition of the concept in question is offered and then both factual and counterfactual examples are used to test whether the concept and its definition are in conformity. If one comes up with an example in which the concept but not the definition applies or vice versa then the definition is rejected. The aim is to formulate a final definition which satisfies the intuitions and which is resistant against the counter-examples.

Now let us look at a well-known example of a conceptual analysis: the analysis of the concept KNOWLEDGE. Traditionally, knowledge (propositional knowledge) is analyzed with three conditions as follows:

A subject S knows a proposition that P if and only if

- P is true.
- S believes that P.
- S is justified in believing that P.

Now let us see how these three conditions work. Consider the sentence "Mary *knows* that John killed Tom." If P is not true (if, for example, Tom was killed by somebody else or Tom is still alive) then the intuition suggests that although Mary believes that John killed Tom, it is not appropriate to say that Mary knows that John killed Tom. Hence the first condition for genuine knowledge: if a subject S knows that P then P must be *true*. But being true is not enough for being known. There are infinitely many true propositions but it is odd to say that we know all of them. So we need a second condition. In order for Mary to know that P, in addition to the truth of P, Mary must *believe* that P. Mary cannot be said to know that P if she does not believe that P. But it turns out that these two conditions are still not enough for knowledge. Suppose that Mary

came to believe that John killed Tom in an irrational way. She thinks that she has superphysical powers. Most philosophers agree that even if it is the case that John killed Tom and that Mary believes that John killed Tom, that will not imply that Mary *knows* that John killed Tom. What is needed is some kind of *justification* of the part of Mary and hence the third condition. But what kind of justification is needed is not easy to spell out. For example it might be the case that Mary in fact saw the crime or a friend of her whom she trusts told her so. Things seem to get complicated as we scrutinize the analysis. Moreover, Gettier (1963) proposed counterexamples to the analysis of knowledge as justified true belief. He showed that there are cases which satisfy the justified true belief criterion but which, intuitively, cannot be counted as knowledge. Today nobody thinks that a successful account of the concept KNOWLEDGE has yet been given, and given the failures of conceptual analyses in other domains, many philosophers think that there is something inherently wrong with the idea of conceptual analysis. But of course many (such as Jackson, 1998) still think that despite its problems conceptual analysis is one of the central aims of philosophy.

Philosophers who pursue the project of intentionality generally do not like to talk about their methodology. This must be either because they think that the issues about methodology are trivial or because they fear entering into such a topic from which it might be hard to get out. Or perhaps they think that to do philosophy you need to start somewhere; so they start with an attempt at naturalization without enough discussion of what naturalization amounts to. Whatever the reason for this situation may be, I think that it is necessary to have some idea about the methodology and the nature of naturalization in order to make sense of the naturalization proposals on the table.

To repeat, Fodor's way of approaching the problem (that is, find a condition C such that "R represents S" is true iff C') seems to be that he is looking for a conceptual analysis of the notion of representation. But when we look at the outcomes of naturalization produced by philosophers working on the subject the situation seems somewhat more complex. These outcomes include analyses based on notions such as causation, similarity (or isomorphism), functional role,

and natural selection. The thing to notice is that there is no conceptual link between representation and these latter notions. In other words, the notions such as causation, natural selection, or isomorphism are not *conceptually* necessary for something to represent.

Although naturalization involves conceptual analysis, it seems more than that. I think that naturalization can be seen as a two-step process. The first step is conceptual analysis, that is, the concept of INTENTIONALITY (or REPRESENTATION) is given necessary and sufficient conditions for its application by checking our intuitions on factual and counterfactual examples. And in the second step a naturalistic system is described which is claimed to satisfy these conditions. It is possible that there might be more than one naturalistic system which satisfy the conditions for intentionality. The discussions among philosophers on naturalization of intentionality have been generally focused on the second step. The first step, that is, the conceptual analysis, is largely ignored. But I think that although conceptual analysis does not appear explicitly, it is implicit in many of the discussions. The reason for this might be due to the unsuccessful history of conceptual analysis in philosophy. But I think that conceptual analysis is an important part of naturalization and should be explicitly discussed.

The output of a naturalistic analysis of the concept of intentionality can contain only naturalistic terms in the analysis. In other words, the analysans should not contain semantic or intentional terms. But this involves an implicit assumption that we have already criteria for distinguishing between expressions which are intentional (or semantic) and which are not. But what are these criteria? Without such criteria it is not possible to give a naturalistic analysis of intentionality. In the footnote of the above quote Fodor tries to answer this kind of question: “Since we haven’t any general and satisfactory way of saying which expressions are semantic (or intentional), it’s left to intuition to determine when a formulation of C meets this condition” (p. 48). It is generally accepted that expressions such as “believes,” “desires,” “thinks,” “is about,” “refers,” and “is

true" are semantic (or intentional) expressions and these expressions should not occur in the analysis of intentionality.

Let me illustrate this two-step process with an example. Take the folk concept of GHOST. Suppose that our aim is to show that ghosts are physically possible creatures. To do this we need first to define what it is to be a ghost, i.e., we should give a conceptual analysis of the concept GHOST. Suppose that after some conceptual analysis we decided that one condition for being a ghost is to be able to pass through walls. In the second part we need to describe something physical which has this property, that is, some physical entity which can pass through walls. If we can do this and can also describe something physical which satisfies all the other conditions for being a ghost, then we can say that we have successfully naturalized the concept of GHOST. As I said before, a successful naturalization does not show that ghosts really exist; all it shows is that ghosts are physically possible creatures.

### 3.2 Eliminativism, Physicalism and Reductionism

As I stated before, the motivation of the naturalistic project arises largely because of the eliminativist threat. A successful naturalization is supposed to count as a reply to eliminativism. In this section I will look at this idea a bit closer. Let us look at a quote from Fodor on this point:

> 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 the 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 the intentional are real properties of things, it must be in virtue of their identity with (or maybe of their supervenience on?) properties that are themselves *neither* intentional *nor* semantic. If aboutness is real, it must be really something else. (Fodor, 1987, p.97, emphasis his)

Here Fodor endorses a special version of a general metaphysical claim called physicalism. Simply put, according to physicalism everything that exists is physical. Physicalism also helps the eliminativist claim in the following way: If some entity can't be shown to be a physical entity then it is legitimate to doubt its reality. The same claim also applies to intentionality. If a physicalist claims that intentionality is something real then she should show how it is to be given an account in physical terms. If intentionality cannot be accounted for in physical terms and if physicalism is true, then the conclusion we have to draw is that intentionality does not exist.

Although many contemporary philosophers assume that some version of physicalism is true, it proved difficult to spell out what exactly physicalism is without committing to circularity or triviality. Generally, philosophers endorsing physicalism assume that physical entities are those posited in the science of physics (and may be chemistry and biology). So basically what you do is to divide the set of entities in the world into two large groups. Usually the entities posited in lower-level sciences (such as physics, chemistry, biology, etc.) are included in one group and the entities posited in upper-level sciences (such as, psychology, sociology, economics, etc.) are included in the other group. The physicalist claim then says that the entities in the latter group can be accounted for by (or reduced to) the entities in the former group. This construal of physicalism is nicely summarized in the following (somewhat long) quote by Petitt (1993) which gives a definition of physicalism as consisting of four claims:

> *Claim 1. There are microphysical entities.*
>
> A. There is an empirical world of the sort that physics posits.
> B. Different kinds of thing in the empirical world share (subatomic) levels of composition of the kind that physics – specifically, microphysics – posits: there is a realm of smaller and simpler, microphysical entities.
>
> *Claim 2. Microphysical entities constitute everything.*
>
> A. Everything in the empirical world is composed in some way – composed without remainder – out of (subatomic) entities of the kind that microphysics posits, or it is itself uncomposed and microphysical.

B. The composition involved is conservative or non-creative in this sense: absent the introduction of a new source of higher-level laws or forces, two microphysically composed entities cannot differ intrinsically without some difference of a microphysical kind – without some difference in the character or configuration of their microphysical components.

*Claim 3. There are microphysical regularities.*

A. Microphysical entities are subject to certain law-like regularities, in virtue of their microphysical properties and relations.
B. The laws at work in the microphysical realm do not obtain because they are required to obtain by the obtaining of certain laws at a macro level: perhaps the same laws (e.g., the same conservation laws), perhaps laws of a different character; the microphysical laws, as you may say, are primitive.

*Claim 4. Microphysical regularities govern everything.*

If there are macro-level laws, as there surely are, then:

A. They do not complement micro-level laws, taking up some degree of slack left by those laws; and
B. They are not independent of micro-level laws: they do not have the potential to conflict with them and they do not serve to reinforce them, representing an extra booster for sequences of events that are established in accordance with those laws.

Let me give an example which, I believe, helps to grasp the basic idea of physicalism. The example I have in mind is the Chinese Tangram puzzle. The aim of this puzzle is to construct a given big shape out of smaller pieces by obeying the rules of the game. Traditionally there are seven pieces: two large triangles, an intermediate triangle, two small triangles, one square, and one parallelogram. The rules of the game are simple: one must use all seven pieces, pieces must lay flat, and they must be adjacent without overlapping. Using the seven pieces and obeying the rules it is possible to construct a large number of big shapes. However there exist some big shapes which cannot be so constructed. For example a circle cannot be constructed out of the seven pieces and by obeying the rules.

I think that the Tangram game is a useful example to illustrate the idea of physicalism. The seven pieces correspond to the basic particles posited by micro

physics and the rules of the game correspond to the basic regularities (laws) involving the basic particles. We can formulate the thesis of physicalism with respect to the Tangram game as follows: Everything in the Tangram world is composed of, without remainder, the basic seven pieces. And the rules that govern the seven pieces govern everything. The eliminativist thesis, then, can be formulated as follows: If one cannot show how a big shape can be constructed out of the smaller pieces then the big shape does not exist in the Tangram world.

Physicalism as formulated above is not immune to objections. For example Crane and Mellor (1990) remark that physicalism as formulated by many of its defenders is either false or trivially true. The problem is how to define what constitutes the physical (or microphysical) entities. If physical entities are defined as the entities posited in *current* physics, then physicalism is false since current physics is surely incomplete to account for all aspects of the natural world. On the other hand if physical entities are defined as the entities in the *complete future* physics then physicalism is trivially true. Because if we assume a *complete* physics then it is an analytic truth (that is, true by definition) that it will account for any aspect of the natural world. Secondly, what kind of entities the complete future physics will posit is not known. It might turn out that in the future mental entities will be counted as physical entities, which implies that physicalism will be trivially true for the mental.

The problem of giving a definition of physicalism which is not false and not trivially true is a real problem. But I don't think that it has no solution. For example Jackson (1998) proposes the following. He thinks that we can define physical entities and physical laws as the ones which are needed to give a complete account of the ordinary objects such as rocks, trees, water, etc., around us. Then we can define physicalism as follows: the physical entities (and physical laws) which are sufficient to give a complete account of ordinary objects are sufficient to give a complete account of everything. One problem with this definition of physicalism, as noted by Jackson, is panpsychism. That is, if rocks, trees, and the like turn out to have mental states then according to this

definition, physicalism becomes trivially true. Nevertheless Jackson thinks that we can safely assume that panpsychism is false.

One might think that Jackson's proposal still has problems. My primary aim here is not to give a definition of physicalism as a substantial claim (that is, a claim which is not false and trivially true). All I want to show is that both intentional eliminativists and realists assume some version of physicalism, and the naturalist project, as a reply to eliminativism, is understood better with physicalism in the background.

### 3.3 Naturalism and Conceptual Analysis

In this section I will argue against Stich and Laurence's (1992) attack on the project of naturalization. Stich and Laurence think that the project of naturalization is misguided and that philosophers who are engaged in naturalization do not have a coherent idea of what naturalization amounts to. Since according to Stich and Laurence there is no clear statement of what naturalization is, they propose three alternative accounts of naturalization and try to show that none of them is satisfactory. It is important to note that Stich and Laurence do not claim that there cannot be a satisfactory account of intentionality. All they claim is that if naturalization is unpacked along the three alternatives that they give, then naturalization is not a coherent project. They do not deny the possibility that there might be a fourth alternative which is satisfactory. According to one of the alternative accounts, which Stich and Laurence consider, naturalizing demands a conceptual analysis of the notion of intentionality.[4] And they go on to argue that this view is problematic. According to my construal, although naturalization cannot be equated with conceptual analysis, conceptual analysis is an integral part of it. So, Stich and Laurence's

[4] The second alternative account construes naturalization as an a posteriori identification and the third construes it as a supervenience claim. Since my account of naturalization explicitly involves conceptual analysis I will only consider Stich and Laurence's argument for the case of conceptual analysis.

argument against naturalization construed as conceptual analysis poses a threat also to my account of naturalization. Below I will first give Stich and Laurence's argument and then argue against their view.

According to Stich and Laurence (1992) any account of naturalization has to satisfy a pair of constraints:

> First, it will have to sustain an argument from the premise that intentional notions can't be naturalized to the conclusion that intentional irrealism or some other deeply troubling doctrine is true. Second, there must be some reason to think that, when 'naturalizing' is unpacked along the lines proposed, it is in fact the case that the intentional can't be naturalized. (p. 160)

Stich and Laurence admit that when naturalization is unpacked as conceptual analysis it is indeed the case that intentionality can't be naturalized. That is, no satisfactory conceptual analysis of intentionality can be given. Stich and Laurence give two well-known reasons for this result. The first is the unsuccessful history of conceptual analysis. Although numerous attempts have been made no satisfactory analysis of concepts such as JUSTICE, TRUTH, CAUSATION, FREEDOM, KNOWLEDGE, etc. have been given. Secondly, conceptual analysis is thought to presume the view called the classical theory of concepts according to which concepts are mentally represented as a set of individually necessary and jointly sufficient conditions. However, studies on categorization have revealed that human categorization involve typicality effects and fuzziness. These results are usually considered to show that apart from philosophically interesting concepts, concepts such as CAT, TREE, BIRD, etc., cannot be analyzed in terms of a set of individually necessary and jointly sufficient conditions. Stich and Laurence rightly argue that if naturalization is unpacked as classical conceptual analysis their first constraint cannot be satisfied. That is, unnaturalizability does not imply irrealism. As Stich and Laurence remark it is simply absurd to conclude that cats don't exist from the premise that the concept CAT cannot be given a classical conceptual analysis. So

the conclusion they draw is that if naturalization is unpacked as classical conceptual analysis then naturalization is deeply misguided.

However, I think that the conclusion Stich and Laurence draw is too hasty. I do not deny that classical conceptual analysis has problems; however, nothing in philosophy is immune from problems. Maybe what is needed is to modify the classical form of conceptual analysis. This is what I will argue below.

Consider the concepts of CAT and GHOST. Intuitively, there is a difference between them: while there are many objects which fall under the concept CAT, presumably there are no objects (in this world or in any physically possible world[5]) which fall under the concept GHOST. How can one argue for this intuitive difference? One way to argue for this might be as follows. Ghosts do not exist because the concept of GHOST has such properties which cannot be instantiated by a physical thing. Such a property, as I mentioned before, might be the property of being able to pass through the walls. But how do we ascribe such a property to ghosts? I can see no other way other than conceptual analysis. The concept of GHOST is not an exception to prototype theory. There might be typical or atypical ghosts. Or there might not be a set of necessary and sufficient conditions for being a ghost. But what is obvious is that being able to pass through the walls is associated with the concept GHOST. So, I think that in order to save the intuitive difference between the concepts of CAT and GHOST one needs to appeal to conceptual analysis. On the other hand, Stich and Laurence must either deny this intuitive difference (which is absurd) or show another way to save it.

Stich and Laurence might reply that what they argue against is only the classical form of conceptual analysis. They may not insist that all forms of conceptual analysis are hopeless.[6] To repeat, according to classical conceptual analysis a concept is analyzed as a set of individually necessary and jointly

---

[5] If you think that ghosts are not physically impossible creatures then you can just take another one from the set {EVIL, MIRACLE, SANTA CLAUS, etc.}.

[6] Whether Stich and Laurence think that all forms of conceptual analysis are hopeless (that is, whether there is something wrong with conceptual analysis in general) is not clearly stated in their paper.

sufficient conditions. Prototype theory (Rosch and Mervis, 1975), on the other hand, analyzes a concept as a weighted list of typical features. Not all features in the list need to be possessed by an object in order for it to fall under a concept. Whether the classical or the prototype (or some other) view is correct is not resolved yet. But my point is that even if the prototype view turns out to be the ultimate truth this does not lessen the importance of conceptual analysis. For the essence of conceptual analysis is to consult our intuitions under factual or counterfactual cases in order to reveal the properties that we associate with a concept. So I agree with Stich and Laurence in that when naturalizing is unpacked as classical conceptual analysis then it is hard to satisfy their first constraint. But other forms of conceptual analysis such as prototype view can satisfy their first constraint. But this time Stich and Laurence might reply that if naturalization is unpacked along the lines of the prototype view then their second constraint can't be satisfied. The second constraint says that it is the case that intentionality can't be analyzed in accordance with the prototype view. If intentionality can be given such an analysis then there is nothing to worry about. However, naturalizing, at least my construal of it, is not just to give an analysis of the notion of intentionality. What, in addition to analysis, is needed is to show that the analysis can be satisfied by a physical system. I think this is the real challenge of naturalization. And there is something to worry about if this cannot be achieved since then intentionality would be placed among ghosts, miracles, evils, etc.

Once the analysis of concepts admits of degrees and fuzziness the results of naturalization will also reflect this fact. For example, rather than to say that ghosts exist or ghosts do not exist, we would have to say that there are typical ghosts or there are no typical ghosts. But this is not a problem. Prototype effects can be detected for almost all kinds of concepts (even for mathematical concepts such as EVEN NUMBER) and I don't see why the concept of NATURALIZATION needs to be immune from it. Once again, the analogy of Mendelian genes might be helpful here. As I remarked earlier, the definition of genes implicit in Mendelian theory may not be totally satisfied by DNA

segments but this does not prevent us to conclude that Mendelian genes exist. I do not think that it is possible to escape from fuzziness altogether. But fuzziness should not lead us to give up what we are doing.

# CHAPTER IV

# REPRESENTATIONAL THEORY OF MIND

## 4.1 Three Main Conditions on Intentional States

As I have explained in the previous chapter I see naturalization as a two-step process. The first step is to specify the conditions for the application of the concept of REPRESENTATION (or INTENTIONALITY). However, there seems to be a problem. Naturalization primarily is an attempt to vindicate folk psychology. And to do that we need to naturalize the concepts of folk psychology. However, neither the term "representation" nor the term "intentionality" figures in the statements of folk psychology. It seems to be that putting the naturalization project in this way is a bit misleading. What Fodor actually wants (and does) is the naturalization of intentional states, such as beliefs, desires, hopes, fears, etc. The concept of REPRESENTATION appears only in the analysis of intentional states. But since representation is not thought to be a naturalistic notion, the project of naturalization then goes on to naturalize the concept of REPRESENTATION. This might be a bit confusing but I hope it will become clear as we proceed.

In this section I will give three main conditions on intentional states and in the next section I will give Fodor's theory which is designed to satisfy these conditions. The conditions I will mention appear in Fodor (1987, 1990a).

*Condition 1: Intentional states are semantically evaluable or they have aboutness.*

This is one of the most important properties of intentional states. Intentional states are about certain other things (objects, properties, states of affairs, etc.). For example, in the sentence "Tom believes that Mary is at home," Tom is claimed to be in a certain intentional state, namely, "believing that Mary is at home." The semantic value of this intentional state is the state of affairs that Mary is at home. If it is the case that Mary is at home then we say that Tom's belief is true, if it is not the case that Mary is at home then we say that Tom's belief is false. The proposition "Mary is at home" is also called the *content* of the intentional state. Like the intentional state believes that P, other intentional states (such as desires that P, fears that P, thinks that P, etc.) are semantically evaluable too. This property of referring to some other thing is a very special property that no other thing seems to have. Trees, rocks, galaxies, etc., all have many properties but they don't have the property of being about some other thing.

*Condition 2: The interactions among intentional states, usually, respect the logical relations of their content.*

If a person believes that P $\rightarrow$ Q and believes that P then that person, usually, believes that Q. For example, if a person believes that too many people will die if the war begins and she believes that the war began then we expect that that person believes that too many people will die. In other words, thinking is not just random interactions among intentional states; interactions among intentional states are semantically coherent (i.e., are truth preserving).

*Condition 3: Productivity and systematicity of thought.*

Thought is productive, that is, there is no limit to the number of thoughts we can entertain. For example, you probably did not think before that "No grass

grows on kangaroos" (Fodor's example). But it is not difficult to entertain it. Similarly we can think about (or believe) an infinite number of thoughts. (Not at the same time of course.) Thought is also systematic, that is, if a person can think that P then she can also think that Q, where P and Q are logically related. For example, if one can think that John loves Mary then she can also think that Mary loves John.

### 4.2 Representational Theory of Mind (RTM)

RTM is proposed by Fodor as a naturalistic theory designed to satisfy the three conditions that I have given in the previous section. I have said that one of Fodor's main motivations for proposing a naturalistic theory is to show that intentional states can exist in a naturalistically possible world. But this is not the only reason behind proposing RTM. Fodor also thinks that RTM is true in the actual world. According to Fodor RTM is the best theory (in fact, according to Fodor, the only theory) which satisfies these conditions. Fodor thinks that, the absence of rival theories is a strong reason to think that RTM is true in the actual world.

Although Fodor revised his formulation of RTM in his later writings (Fodor, 1998), the original formulation of the theory consists of two claims (Fodor, 1987, p.17):

Claim 1 (The nature of propositional attitudes)

- For any organism O, and any attitude A toward the proposition P, there is a ('computational'/'functional') relation R and a mental representation MR such that
  - MR means that P, and
  - O has A iff O bears R to MR.

Claim 2 (the nature of mental processes)

- Mental processes are causal sequences of tokenings of mental representations.

In the quote above, the expression "propositional attitudes" is just another name for intentional states. Fodor's analysis of propositional attitudes construes

them as consisting of two aspects: *state* and *content*. States are characterized by the words belief, desire, etc., and contents are what the complement clause in propositional attitudes refers to. For example, in the sentence "She believes that it will rain," the state of the propositional attitude is "believing" and the content is "It will rain." Claim 1 asserts that for each propositional attitude there is a corresponding mental representation which has the same content as the propositional attitude. And the relation R distinguishes what the state of the propositional attitude is. So what distinguishes a belief from a desire when both of them have the same content is the relation of the mental representation to the organism. The functional role played by the same mental representation is different in believing versus desiring.

Now let us see how the three conditions I have given above are satisfied by RTM. I will start with condition 2, which says that the interactions among intentional states, usually, respect the logical relations of their content. How can something physical manage to do this? It is important to note that according to claim 1, having a belief that P is to have a mental representation (or mental symbol) in the belief box[7] which means that P. To grasp the depth of the question we should distinguish two aspects of a mental symbol: its syntax and semantics. The syntactical properties of a mental representation are physically intrinsic properties like its shape whereas the semantic properties of a mental representation are given by its truth conditions. Intuitively, the interactions betweens mental representations are sensitive only to their syntactic properties. This point can be illustrated nicely with an example (Dretske, 1989). Suppose that a soprano breaks a glass by singing. Intuitively the meaning of the song is causally irrelevant for the effect. Accordingly, the question unpacks as follows: how can interactions between mental representations which are sensitive only to their syntax manage to be truth-preserving if the meanings of mental

[7] The "belief box" metaphor is just a convenient way to designate the functional role of belief states. To say that one has the proposition P in one's belief box is to say that the mental representation which means that P bears the functional role of a belief in one's mind.

representations have no effect on their interactions? Note that this is not the same problem as the problem of mental causation. The main problem in mental causation is to explain how semantic properties can be causally relevant for an effect. But in our case the problem is to explain how syntactic interactions between mental representations can respect their semantic properties. According to Fodor two great achievements in the twentieth century make this possible: formal systems theory and Turing machines. Let us look at them in turn.

A formal system consists of a set of symbols and a set of rules for manipulating and transforming the symbols. The name "formal" designates the fact that the rules operating on the symbols are sensitive only to the symbols' formal properties (the term "formal" is used with the same meaning as that of the term "syntactic"). That is, the meaning of the symbols (if there is any) has no effect on the operation of the rules. Formal systems are mainly developed to make mathematical theorizing as precise as possible, free of hidden meaning assignments. The difficult part of formalizing a mathematical domain is to specify formal rules which are truth-preserving. But once you specify the truth-preserving formal rules and the initial symbols (axioms) to start with, the rest is just a mechanical process: you simply run the rules on the initial symbols and it is guaranteed that the symbols which are produced by the system (theorems) will also be true.

On the other hand the works of Turing showed that it is possible to implement a formal system on a Turing machine. Formal systems theory together with Turing's work show us that we can build general mechanical systems which can have truth-preserving syntactic operations (of course within some limitations such as shown by Gödel's incompleteness theorems).

The problem raised by the semantic coherence of thought for naturalization is also related with another problem which is known as Descartes' challenge: How can something material be rational? For Descartes no other thing in nature other than the human mind can be rational which leads him to his well known dualist position. I have said that formal system theory and Turing's work showed how to explain semantic coherence naturalistically. But although this is necessary to

give a naturalistic account of rationality it is not sufficient. That is, with formal systems and Turing machines you can build mechanical theorem provers or chess playing machines but this is not enough for rationality. For rationality we need, at least, to show how to implement mechanically valid reasoning. To do this we need to formalize logic and this takes us to another great achievement. The works on symbolic logic, especially the works of Frege, showed us how to formalize a substantial part of logic (the predicate calculus) and since this formalized logic can be implemented on a Turing machine, it is possible to build physical systems which can logically reason (logic theorem provers are such systems).

Returning to condition 2, given that beliefs are nothing but mental symbols which bear the appropriate relation to the organism (this relation is generally thought not to pose any threat to naturalization), it is possible to build physical systems whose belief states interact with respect to their content as well as which can make logically valid reasoning.

Now let us proceed with condition 3. Recall that condition 3 was productivity and systematicity of thought. RTM postulates a system of mental representations (called language of thought or Mentalese) which has a compositional syntax and semantics. Mentalese works like a language. And the problem of productivity of thought is solved much like language: a finite set of mental representations together with a compositional syntax and semantics is sufficient to form an infinite number of complex mental representations and hence infinite number of thoughts or beliefs. The problem of systematicity is solved similarly. If mental representations have a compositional syntax and semantics then it easy is to explain systematicity.

To illustrate the idea of how a compositional system can solve the problem of productivity consider an example automaton. Let us assume that the automaton takes as input a string (of any length) which consists of 0s and 1s and outputs 1 if the number of 0s is equal to the number of 1s and outputs 0 otherwise. The behavior of the automaton can be said to be productive since there is no upper limit on the length of strings it can process. One possible explanation of the

behavior of the automaton is to assume that it contains two very large lists of strings one of which holds the strings with equal number of 0s and 1s and the other list holds the strings with unequal number of 0s and 1s. When the automaton takes a string as input it searches both lists and according to the result of this search it outputs either 1 or 0. Though this might be a possible explanation it seems implausible. For it takes too much space to store all these strings and too much time to compare the input string with the members of the list. Moreover, if the automaton can correctly classify an infinite number of strings then this explanation is hopeless. One other solution is to postulate a recursive grammar. A grammar which can process the strings that we are considering might be as simple as this: $S \rightarrow 0S1S$, $S \rightarrow 1S0S$, and $S \rightarrow \varepsilon$. So, if this recursive (hence compositional) grammar is implemented in the automaton it can accept an infinite number of strings with equal number of 0s and 1s.

We have seen how the properties of FP that are specified by condition 2 and condition 3 can be derived from the properties of the naturalistic framework posited by RTM. However, the property specified in condition 1 is just assumed. That is, the semantic properties of intentional states are derived from the semantic properties of mental representations. But RTM does not provide an account of how mental representations acquire their semantic properties, i.e., RTM does not answer the question "In virtue of what mental representations represent?" And, as Fodor thinks, it is unacceptable to have unanalyzed mental representations at the foundations of philosophy of mind. So, now the problem is to naturalize the notion of representation.

### 4.3 Naturalizing Representation

The naturalization of representation has turned out to be a recalcitrant problem. For although quite a number of proposals have been made, none of them has enjoyed wide acceptance among philosophers.

How can one evaluate whether a given naturalization proposal for representation is successful? In other words, what are the desiderata of a successful naturalization? As I have tried to explain in the previous sections, the

first thing is to analyze the concept of REPRESENTATION and reveal the conditions for its application. I will not list all of the conditions mentioned in the discussions but consider some of them which are, I think, shared widely and which have played an important role in the literature. Also I will not immediately give a list of conditions for the application of the concept REPRESENTATION. These conditions gradually appear as we go through some of the naturalization proposals. I will not consider all of the naturalization proposals made to this date.[8] My focus will be on two fundamental notions, namely, causation and similarity. Both notions are widely proposed by different philosophers with different add-ons as a naturalistic base for representation.

A note will be appropriate here. When trying to set out these conditions, philosophers do not content themselves only with the concept REPRESENTATION as it appears in folk psychology. They also demand that naturalization proposals should also be compatible with (or explain) the nature of concepts and meaning in general. The reasons are as follows: According to some philosophers concepts are nothing but structured (or atomic) representations. That is, the capacity to represent is the essential aspect of being a concept. Accordingly, a naturalized account of representation naturally also becomes a naturalized account of being a concept. Hence, the issues that are needed to be explained by a theory of concepts should be explainable (to a certain extent) with a naturalized account of representation. This is why the naturalistic accounts of representation are sometimes challenged to explain such phenomena as concept learning, categorization, innate concepts, etc. Some other desiderata come from meaning. Symbols, such as signs, words have a meaning, i.e., they refer to (or represent) other things (objects, events, or states of affairs, etc.) different from themselves. It is generally thought that the capacity of symbols to refer is not intrinsic but derived. Their capacity to refer is inherited from mental representations. Accordingly, semantic properties of natural language terms are

[8] For example I will not consider teleological (Godfrey-Smith, 1998) and conceptual role (Block, 1998) theories.

supposed to be explainable by a naturalistic account of representation. But one should be careful not to demand too much from a naturalistic account of representation. A naturalistic account of representation will help to explain many phenomena related with concepts and meaning in general. But a naturalistic account of representation cannot explain everything about concepts or meaning. What it is supposed to explain and what it does not is not very clear and depends on many other assumptions one holds. One should be careful not to be unfair when evaluating a naturalistic account of representation.

### 4.4 Naturalistic Accounts of Representation

In this section I will consider two of the influential naturalistic accounts of representation. Since the literature on this area is huge, I will confine myself to introduce the basic ideas put forward in these accounts and point out their major difficulties. As we go through these accounts, I will also identify the conditions which a naturalistic account is supposed to satisfy.

Fodor's theory of asymmetric dependence grounds the notion of representation in causation. The basic idea (which is dubbed by Fodor as "the crude causal theory") is this:

A symbol "S" expresses (or represents) the property P if there is a law that instantiations of P cause tokenings of "S."

The first and the foremost condition which any successful naturalistic account has to explain is the fact that representation is an asymmetric relation. In order for a true representation to obtain we need two things related in a certain manner. Moreover representation is an asymmetric relation since the fact that X represents Y does not imply that Y also represents X. (The term "asymmetric" as it is used in this context should not be confused with its occurrence in the name of Fodor's theory which I will explain next.) Causation has seemed to be a promising candidate to fulfill this condition. For causation is a natural relation between two things (objects, events, facts, etc.). Moreover causation is an

asymmetric relation just as representation is. Note that the fact that representation is an asymmetric relation creates a problem for those who support an account of representation based on similarity. For similarity, unlike representation, is a symmetric relation. Now we can state our fourth condition:

*Condition 4: Representation is an asymmetric relation.*

The metaphysics of causation is itself a deep and complex issue. An account which relies on causation has to face the difficult problems that the literature on causation revealed. One such problem is discussed by Putnam (1992, pp. 47-55). There is usually more than just one single cause responsible for the occurrence of a given effect. Suppose that rubbing a match causes fire. It seems to be that the causal relation is between the event rubbing a match and the event fire. But rubbing a match is not the only cause of the fire; there are plenty of other causes, or causal contributories, such as the presence of oxygen and absence of wetness of the match, which were all responsible for that fire. We normally don't cite the presence of oxygen, for example, as the cause of a fire. But imagine a planet where there is ordinarily no oxygen in the atmosphere but lots of match rubbings. In that planet when some oxygen leaks into the atmosphere in the vicinity of a match rubbing, fire is caused. In such circumstances we may regard the presence of oxygen as the cause of the fire. Selecting one of the causal actors as "*the* cause" and regarding the others as "background conditions" or "standing conditions" is interest-driven and depends on the context.

This point can be made more precise by using Mackie's (1974) analysis of causation in which he introduced the notion of an "inus" condition. An inus condition for some effect is an insufficient but non-redundant part of an unnecessary but sufficient condition which can be formulated as below:

$$(C_1^1 C_1^2 \Lambda\ C_1^p) \vee (C_2^1 C_2^2 \Lambda\ C_2^r) \vee \Lambda\ \vee (C_n^1 C_n^2 \Lambda\ C_n^s) \rightarrow E$$

According to Mackie, causes ($C_j^i$) are inus conditions. That is, they can only cause an effect in the presence of other conditions. Moreover, there can be more than one sufficient condition for a given effect. Accordingly, rubbing a match is an inus condition for fire.

This creates a problem for causal accounts of representation in the following way. The intuitive appeal of causation is the fact that we answer the question "In virtue of what does X represent Y?" by saying that "Because Y causes X." However if Y is an inus condition for X then it seems unprincipled to select X as the cause of Y since there needs to be other conditions, other than X, for Y to occur.

The nature of causation is a difficult metaphysical topic. One might think that it is unfair to expect from philosophers of mind to solve the deep problems of metaphysics. Fodor seems to be thinking in a similar way, for he feels free to use causation as the basis of representation with little discussion about the metaphysics of causation[9] and about the "the cause" problem I have been considering. Nevertheless, I think that this is an important and unresolved problem for causal accounts of representation.

*Condition 5: The possibility of misrepresentation.*

Another important and widely discussed problem of the crude causal theory is that it leaves no room for misrepresentation. Misrepresentation is a problem especially for theories based on the notion of causation (but also for similarity-based accounts). Let us see how the problem of misrepresentation arises in causal theories of intentionality. Recall that according to the crude causal theory a symbol "S" expresses (or represents) the property P if there is a law that instantiations of P cause tokenings of "S." So, for example, "cow" expresses the property *cow* because there is a law that cows cause "cow" tokens. But this

[9] Fodor assumes a counterfactual account of causation. But counterfactual accounts do not solve the "the cause" problem.

assertion immediately leads to the problem of misrepresentation. We do not always represent things in our environment in the right way; sometimes we misrepresent them. For example, sometimes (e.g. under poor visual conditions), we might misrepresent a horse as a cow (i.e., horses sometimes cause "cow" tokens). And, given certain background conditions, this can even happen regularly. The capacity of misrepresentation is usually regarded as one of the important features of our representational system. So it is important for a theory of content to explain our capacity of misrepresentation, i.e., to explain how we can have false beliefs as well as true ones. But the crude causal theory leaves no room for misrepresentation. Consider again the horse-caused "cow" tokens. According to our intuitions, horse-caused "cow" tokens are misrepresentations and horses should not be included in the meaning of "cow." But the crude causal theory cannot achieve this, since it puts, by definition, anything that causes "cow" into the meaning of "cow." So if horses are also capable of causing "cow" tokens, according to the crude causal theory, "cow" will also mean *horse*. Hence, horse-caused "cow" tokens will not be a case of misrepresentation. This is a result that we don't want.

*Condition 6: Explanatory value*

Cummins (1996) puts yet another condition, which he calls *explanatory constraint*, on a successful theory of representation: "The theory should underwrite the explanatory appeals that cognitive theory makes to mental representation" (p. 2). Note that contrary to some philosophers such as Fodor, Cummins does not think that a successful theory of representation should explain the folk notion of representation. His aim is mainly to explain the explanatory role of mental representations as used in cognitive science. So, he may not commit himself to some of the conditions that I am placing on a successful theory of representation. My view, however, is that the philosophical theory of representation should both seek to vindicate folk psychology and explain the explanatory role of mental representations.

This condition is also problematic for causal/informational theories. According to these theories to represent is simply to detect. Or to put in other words, to represent is to carry information. Animals and especially humans are good detectors. We can detect many features of the world related with colors, shapes, events, etc. Detection, no doubt, is an important feature of our cognitive system. But the point here is that representation is something more than detection.

To see this consider Fodor's theory of content which postulates a compositional set of primitive symbols. The critical thing about this theory is that most lexical concepts are nothing but atomic symbols which constitute the primitive symbols of this compositional system. For example, the concept COW is just an atomic symbol which is caused by cows. It has no internal structure relevant to cognition. But atomic symbols are arbitrary symbols, that is, the content of a symbol is independent of its intrinsic properties. For example, assume that you want to build a cow detector. All you want is that the detector tokens a symbol when confronted with cows. You can build the machine such that any symbol might be tokened when the detector is confronted with a cow. For example a red light may turn up or a switch may change its position. So, when the detector is confronted with a cow what you get is only an arbitrary symbol tokening.

To see how this is problematic for cognitive explanation consider the example given by Cummins (1996, p. 70). Cummins considers a case in which you are asked to go milk the cow. Let us assume that you have managed to go to the barn. When you look around and see a cow you will token COW. But COW is an atomic symbol, it carries no information about the properties of cows. How are you going to locate the udder? Cummins thinks that a cow image might be useful here but if you assume a causal theory all you have is just an arbitrary symbol. The inevitable strategy for a Fodorian will be to cite stored knowledge. That is, the COW symbol activates the stored knowledge about cows which contains information about the udder (its location, etc.). But the complaint here is that the concept COW does not have any explanatory role except activating some

stored knowledge. The entire explanatory burden is on this stored knowledge. Also, why we should identify the concept COW with an atomic symbol but not also with the stored information it activates is not clear. In short, causation alone seems not to be enough to account for the explanatory appeals of cognition.

*Condition 7: Reference is determinate.*

Another condition which a successful account of representation should satisfy is the fact that reference is a determinate (not ambiguous) relation. Ambiguity is especially a problem for similarity (or isomorphism) based accounts of representation.

To understand this condition better let us look at how it appears in the context of an analysis of representation based on similarity. The appeal of an account based on similarity can be grasped easily with an example. Suppose that a new car is produced with a navigation system. The navigation system is composed of two parts: an electronic map of the streets of Ankara and a computer-controlled mechanical system which navigates the car by inspecting the electronic map and turning the wheel accordingly. Intuitively the electronic map represents the streets of Ankara. But why? Because, as one might reply, while describing the car it is already specified that the electronic map is a map of the streets of Ankara. So, one might think that the map represents what it represents because of this specification made by a human being. Of course, this is not the answer we want, for that will be a circular answer which explains the representational capacity of the map in virtue of the intentions of the designers. What we want is an account of representation which specifies the content of a representation independent of the intentions of a human being. One plausible reason why the map represents the streets of Ankara might be that there is a similarity between the map and the streets of Ankara. This is one of the oldest replies that had been given to the question "Why does a symbol represent what it is supposed to represent?" And naturally it has been subjected to many criticisms (see, for example, Cummins, 1989, pp. 27-34, Goodman, 1972). One such

criticism which motivates an account of representation based on isomorphism rather than similarity is this. Similarity is usually understood as sharing of properties. That is, an object is judged to be similar to another object if they share certain properties. However, it is absurd to expect that the electronic map shares properties (first-order properties) with streets of Ankara, such as having a gray color, made of concrete, etc. What is a more plausible suggestion is that the electronic map shares second-order properties with the streets of Ankara. In other words there is an isomorphism between the electronic map and the streets of Ankara.

Isomorphism is an important notion utilized in cognitive science (Gallistel, 2001) to explain how humans can think about events, situations, etc., that exist outside the mind in their absence. The explanation roughly goes like this: since humans possess mental representations which are isomorphic to external situations, the mental representations act as proxies of the situations in the external world whose informational content is utilized by our cognitive system. The explanation of representation based on isomorphism has many supporters among cognitive scientists. However, from a philosophical standpoint, isomorphism cannot be sufficient to give a naturalistic account representation. Because it is subject to an important criticism: ambiguity. To state this problem clearly let me first define the mathematical concept isomorphism more formally.

A structure $S = (O, R)$ consists of a set, O, of objects and a set, R, of relations defined on the members of O. Given two structures $S_1 = (O_1, R_1)$ and $S_2 = (O_2, R_2)$, $S_1$ is isomorphic to $S_2$ iff

- There is a mapping, $f_1$, from $O_1$ to $O_2$ and
- There is a mapping, $f_2$, from $R_1$ to $R_2$ and
- For all $r \in R_1$, if r holds on objects $o_1, o_2, o_3, \ldots, o_n \in O_1$ then $f_2(r)$ holds on objects $f_1(o_1), f_1(o_2), f_1(o_3), \ldots, f_1(o_n) \in O_2$.

Figure 1, illustrates the notion of isomorphism with an example.

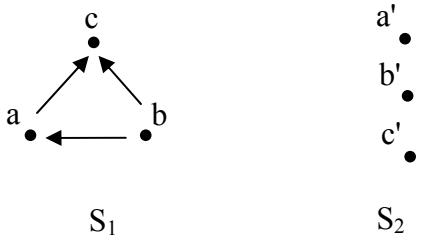


Figure 1: An illustration of isomorphism.

The two structures $S_1=(O_1, R_1)$ and $S_2=(O_2, R_2)$ are isomorphic because the following three conditions hold: (1) there is a mapping from the members of $O_1$ to $O_2$ ($a \rightarrow a'$, $b \rightarrow b'$, $c \rightarrow c'$), (2) there is a mapping from the members of $R_1$ to $R_2$ (directed-connection $\rightarrow$ below), and (3) for the relations that hold on the members of $O_1$ ({c, a}, {c, b}, {b,a}) there are corresponding relations that hold on the members of $O_2$ ({c', a'}, {c', b'}, {b',a'}).

The biggest problem of the notion of isomorphism is the fact that it is ambiguous just like the notion of similarity. Just as any thing can be similar to any other thing, any physical thing can be isomorphic to almost any other thing. For example, it is possible to specify an isomorphism between the structure $S_1$ and three cars lined up in a parking area or three books on a shelf or a host of other things.

Note that there are two kinds of ambiguity in isomorphism which I will call as *structure ambiguity* and *mapping ambiguity.* Let me first explain structure ambiguity. Recall that a structure consists of a set of objects and a set of relations defined on the objects. Consider a map. The intended structure usually consists of a set of points on the map and the spatial relationships between the points. However, one can also determine the structure as follows: a set of points on the map and the color relationships between the points. If not infinite, one can determine quite many structures on a given physical thing. On the other hand, even if a particular structure S is determined on a physical object there is still an

ambiguity problem. S can be isomorphic to more then one structure. Consider a map of Ankara. An isomorphism can be defined from the map to Ankara. However, it is also possible to define another isomorphism from the map to another map which is a duplicate of it. So, if the only thing we have to ground representation is isomorphism then we have to conclude that the map of Ankara represents both a city and another map (and many other things). A successful account of representation should handle these two kinds of ambiguities.

# CHAPTER V

# A NEW ACCOUNT OF REPRESENTATION

I will now begin to develop my naturalistic account of representation. Any theory of representation can be seen as a set of integrated parts which are designed to account for different problems. My account is no exception. I should say that not all parts of my account are new and some parts are similar (although involve different aspects) to other previous proposals. For example, my account assumes that RTM's explanation of the semantic coherence and the productivity/systematicity of thought is correct. But I can say that some new proposals that I am going to offer and the way the parts are put together constitute a new account.

## 5.1 A Critique of Fodor's Causal Account

As I said before Fodor's theory of asymmetric dependence has been widely criticized. I will not repeat those criticisms here. Instead I will criticize one aspect which is at the foundations of Fodor's theory and which, I think, has not been paid enough attention by his critics. This criticism will also start off my development of my account of representation. Here is Fodor's theory of asymmetric dependence:

A symbol "S" represents the property P if,

(i) instantiations of P lawfully cause "S" tokenings.

(ii) sometimes tokens of "S" are lawfully caused by non-Ps.

(iii) non-P caused "S" tokenings are asymmetrically dependent on P-caused "S" tokenings.

As we have seen before RTM leaves an important gap: how mental representations get their semantic values is unexplained. Fodor's asymmetric dependence theory (ADT) comes into play to fill this gap, that is, to provide semantics for the atomic symbols of Mentalese. The essential characteristic of ADT is to derive the semantic properties of atomic symbols of Mentalese from the causal connections that the symbols have with the external world. This is done basically as follows: the causal connections of a symbol determine a set of candidate properties that the symbol might refer to and the asymmetric dependence condition selects one of them as the reference of the symbol. The part of ADT that has received much attention is the notion of asymmetric dependence that occurs in (iii). Almost all of the discussion on ADT has focused on this notion. Quite many papers are devoted to discuss questions such as how we should understand the notion of asymmetric dependence or whether asymmetric dependence solves the disjunction problem (see for example papers in Loewer and Rey, 1991).

However, philosophers have paid little attention to condition (i). One reason might be the belief that condition (i) is basically a metaphysical assumption and discussing it will take one away from doing philosophy of mind into doing metaphysics. In order to do philosophy of mind one needs to start at some point. And that there are causal laws between properties is a good starting point which does not go against, in any extreme way, the considerations in metaphysics of causation. I also share this belief; however, I think that there is a serious problem with condition (i) which has gone unnoticed, and that this problem does not stem from some metaphysical considerations on causation.

In his discussion of the disjunction problem, Fodor frequently uses the example of a horse causing the symbol "cow." Here '"cow"' is the Mentalese term which is stipulated to refer to the property *cow*. He points out that both cows and horses (under poor visual conditions, for example) are capable of

causing "cow" tokens. But then, according to clause (i) taken in isolation (which is called by Fodor "the crude causal theory") only, "cow" refers to the disjunctive property *cow or horse*, which is a result we don't want. We need to find some way to prevent the property *horse* (or in general semantically irrelevant causes) to enter into the meaning of "cow." This is where the asymmetric dependency comes in. According to Fodor, since the *horse*-to-"cow" law depends asymmetrically on the *cow*-to-"cow" law, "cow" means only *cow*. Here my aim is not to discuss the notion of asymmetric dependence or to discuss whether it succeeds in solving the disjunction problem or not. What I want to do instead is to discuss another assumption of Fodor's.

In the above example, it is clear that Fodor assumes that there is a causal law between the property *cow* and the symbol "cow" (and between property *horse* and the symbol "cow"). But what is the justification for this assumption? As I said above, it is plausible to assume that there are causal laws between some properties, but to assume that there are causal laws between some properties is one thing and to assume that there is a causal law between any two properties is another thing. So, Fodor owes us an explanation of his assumption that there is a causal law between *cow* and "cow." To put it in a more general way, Fodor assumes that for each lexical concept there is a corresponding property with which the concept is causally connected. But why there can always be found such a causal connection needs to be explained.

I do not think that our brains are causally connected to *horse*. Just think about the possible situations one can token the symbol "horse." If we consider the visual cases only, it is possible to token "horse" by just seeing a horse from a certain angle, i.e., we do not need to see a horse from all possible viewpoints to token "horse." But in none of these cases the property that is causing "horse" is the property *horse*; rather it is *parts* of the property *horse* which are responsible for causing "horse" tokens.

Fodor says that there is a law connecting *horse* and "horse" because when *horse* is instantiated, given the background conditions, "horse" gets tokened. This is true but it does not show that there is a law between *horse* and "horse."

Consider a conjunctive property $P = A\&B\&C$. Assume that the following three laws are in force: $A \rightarrow$ "S," $B \rightarrow$ "S," and $C \rightarrow$ "S." Now, it is true that whenever the property $P$ is instantiated "S" gets tokened but it does not show that there is a law between $P$ and "S."

To give a concrete example, consider a photocell-receptor-controlled door. The door opens when a human comes near it. So, there seems to be a lawful relation between a human and the door. However, what exactly is the property which causes the door to open? Is it the property *being a human*? If it is the case that the door opens when a (non-human) large enough object is placed near it then we cannot conclude that *being a human* (qua *being a human*) is the property which causes the door to open. Probably some other property such as *having a certain size* is causally responsible for the opening event. And humans cause the door to open because they instantiate this property.

Similarly for the horse case. It seems that there is no law between *horse* and "horse"; all we have got are laws between parts of the property *horse* and "horse" (such as the *side-of-a-horse* $\rightarrow$ "horse" or the *front-of-a-horse* $\rightarrow$ "horse"). What exactly these properties are which are responsible for causing "horse" tokens can only be specified by empirical research. But for our purposes it is enough to assume that there are laws between "horse" tokens and some properties instantiated by horses.

### 5.2 What might concepts be?

In the previous section I tried to show that there are causal laws between the concept HORSE and some properties instantiated by horses. I share Fodor's view that the causal relations of a concept play an important role in determining its content. However, contrary to Fodor, I think that it is problematic to claim that there is straightforwardly a law between *horse* and HORSE. So what is needed is to find a way to construct the content of HORSE out of the properties that cause HORSE.

My inspiration comes from theories of object recognition (Edelman, 1997, Palmeri and Gauthier, 2004). Although there are differences between these

theories, almost all of them assume a two-level processing. It is assumed that in the first level some of the properties of an object are mentally represented and in the second level these representations are compared against a set of pre-stored object-category representations. Finally the best matching stored representation is tokened, i.e., the object is recognized as the best matching stored object-category representation. Figure 2 below illustrates this process.

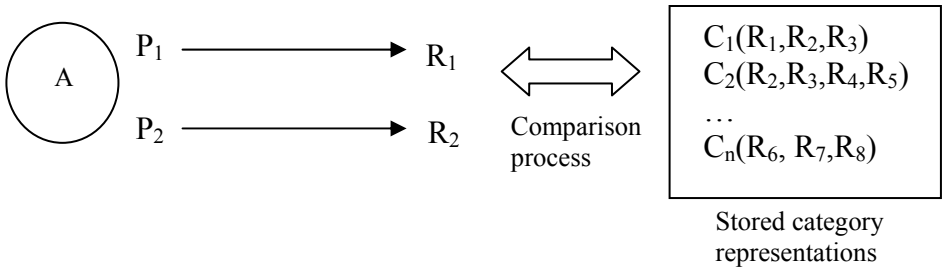


Figure 2: Two-level processing in object recognition.

In Figure 2, $P_1$ and $P_2$ are properties that are instantiated by the object A. $R_1$ and $R_2$ are representations that are caused by $P_1$ and $P_2$. I will call $R_1$ and $R_2$ "primitive representations." $C_1$, $C_2$, …, $C_n$ are representations of categories. Each category representation is composed out of primitive representations; for example $C_1$ is composed of $R_1$, $R_2$, $R_3$. When the object A is presented to a subject, the properties $P_1$ and $P_2$ cause tokens of $R_1$ and $R_2$, respectively. Then these primitive representations are compared against the previously stored category representations and the best matching category, $C_1$ in this case, is tokened.

This is a highly abstract model of the processing involved during categorization of an object which lacks many details. But this abstract model is enough for my purposes. Nevertheless, to make it more concrete let me briefly review one popular theory of object recognition. The theory I have in mind is Biederman's (1987, 1995) recognition-by-components model.

Object recognition is defined by Biederman (1995) as "the activation in memory of a representation of a stimulus class – a chair, a giraffe, or a mushroom – from an image projected by an object to the retina" (p.121). The problem of object recognition is that every time we view an object, a different image is projected on the retina. We can view the object in different orientations, from different distances, or under poor visual conditions. Our visual system is normally very successful in this respect; we can easily classify objects seen from different viewpoints. All theories of object recognition essentially try to account for how our visual system successfully manages this variability in the retinal image.

For example in template-matching models, the retinal image is supposed to be compared against a template which is a representation of a specific view of an object. To manage variability template matching models have two options (Hummel and Biederman, 1992): the model either stores a large number of templates (2-D images) for different views of an object or it should store a small number of 3-D images for an object which are then compared against retinal images by means of transformations.

Template-matching models have a number of difficulties. For example it is time consuming to produce transformations of 3-D images but response times of human subjects to classify an object under different views shows no time variance. But as Hummel and Biederman (1992) note, the main difficulty of template matching models is that they do not explicitly store the information about object parts and their relations, which is critical to the representation of objects by humans. For example in Figure 3 three objects are shown.

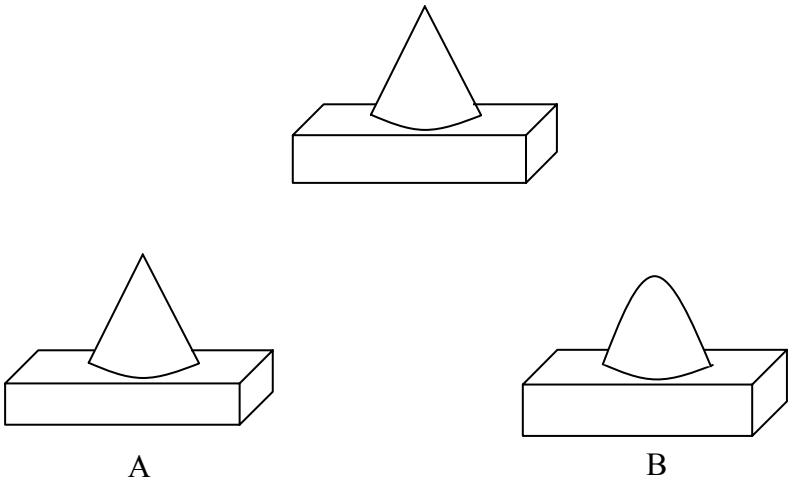


Figure 3: The importance of structure in object recognition.

When a subject is asked which of the two objects at the bottom of Figure 3 are more similar to the one at the top, most subjects will judge that object A is more similar because the tip of the cone in object B is rounded rather than pointed. However a template matching model would select object B as more similar because the rectangle at the bottom of object A is slightly thinner than the rectangle at the bottom of the object at the top. This is because a template matching model just compares the number of mismatching pixels between the objects. And the number of mismatching pixels between object A and the object at the top is more than the number of mismatching pixels between the object B and the object at the top.

There are other problems with template matching models and there are other theories which overcome some of these problems (Edelman, 1997, Palmeri and Gauthier, 2004). My aim is not to cover all of the theories of object recognition and their features. What I want to do instead is to emphasize the point that in object recognition explicitly representing the *structure* of an object is important. This is the idea that object categories are represented as structural descriptions, that is, object properties and the relations between these properties are both explicitly represented.

Such a model is proposed by Biederman (1987). According to Biederman's recognition by components (RBC) model, object categories are represented as configurations of simple, primitive volumes which he calls *geons*. These geons are recognized from viewpoint-invariant properties of retinal images (such as whether a contour is straight or curved or whether a pair of contours is parallel or not). Figure 4 shows a sample of geons used in Biederman's model.

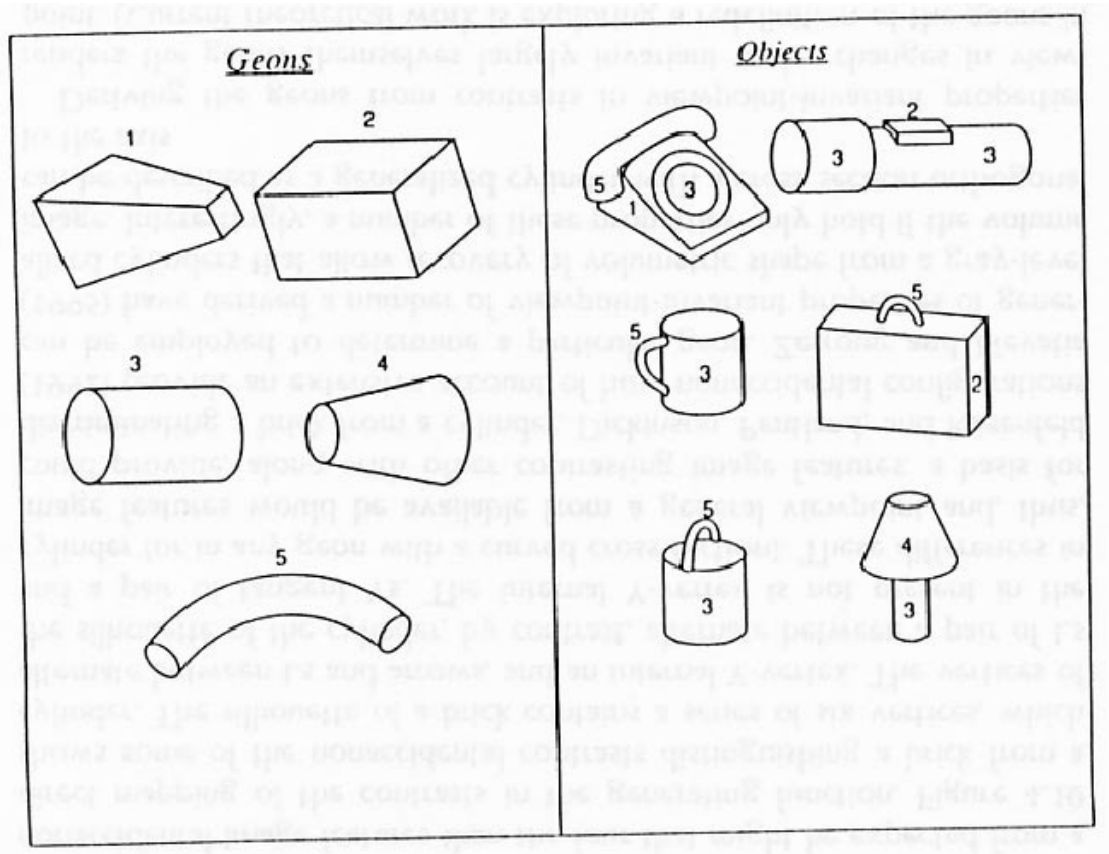


Figure 4: Geons and objects constructed out of them.

At the left of Figure 4 there are five geons and at the right there are objects which are constructed out of them. RBC theory also hypothesizes a set of relations which bind the geons together. Examples of these relations include vertical position (above, below, beside), join type (end-to-end, end-to-middle

centered, end-to-middle off-centered), relative size (larger, smaller, equal to), and relative orientation (parallel, orthogonal, oblique). With a small number of geons and relations it is possible to produce a huge number of objects. Biederman states that "with twenty-four possible geons, eighty-one combinations of relations, and fifteen attributes, the variations in relations and aspect ratio can produce 10,497,600 possible two-geon objects (242 x 152 x 81). A third geon, with possible attributes and its relations to one other geon, yields over 306 billion possible three-geon objects" (Biederman, 1995, p.143).

The model I have in mind also shares the basic ideas of Dretske's (1981) information theoretic account of representation. This account (as described in Aydede and Güzeldere, 2005) postulates an architectural distinction between sensory systems and central cognitive system. The primary job of sensory systems is to provide information to the organism about the external world. The outputs of sensory systems serve as inputs to the central cognitive system which are then used for different cognitive tasks. Primitive representations in my model roughly correspond to sensory representations. Their job is to detect external properties. That is, primitive representations are causally connected to external properties in the world. On the other hand, concepts constructed out of these primitive representations roughly correspond to the elements of the central cognitive system.

To a first approximation, then, I claim that a concept is a structured mental representation whose constituents are primitive representations. And the reference of a concept is the set of things which instantiate the properties represented by (or cause) the primitive representations which constitute that concept. For example, the concept $C_1(R_1,R_2,R_3)$ in Figure 2 represents those objects which instantiate the properties $P_1$, $P_2$, and $P_3$.

Below I will explain how my proposal satisfies the conditions that I have given in chapter IV. This discussion will help further explain and elaborate my proposal. Also in section 5.4 I will show that my proposal is a biologically plausible model by giving a (possible) connectionist implementation.

### 5.3 Conditions Satisfied

In this section I will return to the conditions I set out in chapter IV and discuss how my proposal satisfies these conditions.

#### 5.3.1 Conditions 1, 2, 3, and 4

For Fodor most lexical concepts acquire their content by standing in certain causal relations with the external world. COW refers to *cow* because there is a lawful correlation between *cow* and COW. The symbols denoting lexical concepts are the primitive symbols in a language of thought (or Mentalese). In addition to lexical concepts we can entertain an infinite variety of thoughts. For example we can think that COWS GIVE MILK or HORSES ARE AMONG THE ANIMALS WHICH RUN FAST. How are the content of these thoughts determined? Fodor raises the property of compositionality at this point. Thoughts get their content from the primitive symbols they are composed of together with the way these primitive symbols are combined. Compositionality is also supposed to explain the productivity of thought. We are finite beings but we can entertain an infinite number of thoughts. This is possible because the finite number of primitive symbols with some rules for combination (recursive rules for example) can yield an infinite number of complex thoughts.

Also the semantic coherence of thought is explained by analogy to formal systems and computation as we have seen before. Thoughts are composed out of structured symbols. These structured symbols have both syntax and semantics. Trains of thought can be semantically coherent just because it is possible to mimic the semantic relations between thoughts with the syntactic relations between the symbols.

I agree with these explanations of productivity and semantic coherence of thoughts. My proposal differs from Fodor's at the level of primitive symbols. For Fodor the primitive symbols roughly correspond to lexical concepts. For me symbols corresponding to lexical concepts are also structured. As I said before, I do not think that there are lawful connections between symbols and properties

which correspond to lexical concepts. That is, there is no lawful connection between *cow* and COW. I proposed that there are nomic connections between finer properties than *cow* and primitive mental representations. And the symbols corresponding to lexical concepts are composed out of these primitive mental representations.

Apart from this difference I agree with Fodor's explanations. Our thoughts are productive because they are composed out of primitive symbols with certain rules of combination. And our thoughts are semantically coherent because the semantic relations between thoughts can be mimicked by the syntactic relations between primitive symbols. My proposal, then, can be seen as a modification of Fodor's theory of language of thought in which the primitive symbols of Mentalese are replaced with structured symbols. My proposal does not bring new constraints which blocks the explanation of productivity and semantic coherence of thought. However, according to Fodor structured concepts such as prototypes are not compatible with the compositionality principle. I will discuss this issue in section 6.2.

Recall that condition 1 expresses the fact that representations have the property of aboutness. This condition is further unpacked as conditions 4 to 7. Since to explain how condition 4 is satisfied is relatively easier, let me start with it. Condition 4 says that representation is an asymmetric relation. Recall that accounts based on similarity have problems with this condition. My account on the other hand is partly based on causation. That is, the reference of primitive representations is grounded in causation. So this secures the asymmetric character of representation. In other words my account inherits its asymmetric character from the causal relations between the primitive representations and the external properties.

### 5.3.2 Condition 5

Before explaining how my proposal satisfies the misrepresentation condition, I will criticize Fodor's conception of misrepresentation which will also help us to understand my proposed solution.

It is important to note that misrepresentation is treated by Fodor as a special case of the disjunction problem. That is, according to Fodor both horse-caused "cow" tokens and milk-caused "cow" tokens are cases of tokening of a symbol by a semantically irrelevant cause. Since the essential problem for Fodor is to prevent the semantically irrelevant causes of a symbol from entering into that symbol's meaning, he does not need a special solution for the misrepresentation cases. If he can formulate a theory which will handle the semantically irrelevant causes that will also take care of the causes of a symbol which give rise to misrepresentations.

The case in which horses cause "cow" tokens is classified by Fodor as a case in which tokening of a symbol is caused by a property which is not expressed by that symbol. But why does Fodor think in this way? Why not think that horses cause "cow" tokens in virtue of some of their properties (such as being large, being four-legged, etc.) that they share with cows? Note that if we think in this way misrepresentation does not turn out to be a special case of the disjunction problem. For the properties of horses (such as, *having a large body*, *having four legs*, etc.) which are responsible for "cow" tokenings are not outside the extension of "cow." In fact, this kind of situation is very common. For example, sometimes a mirage can cause a "water" token, a rat can cause a "mouse" token, or a piece of rope can cause a "snake" token. In all of these cases a common pattern can be discerned. Mirages sometimes cause "water" tokens because some of the properties of water, such as the property of reflecting light in a certain way, are shared by mirages. Rats sometimes cause "mouse" tokens because both rats and mice have a pointed nose, a hairy skin, a thin long tail, etc. And ropes sometimes cause "snake" tokens because both ropes and snakes have a long-cylindrical shape, etc.

Let us illustrate the general situation I have in mind with an example and a somewhat more formal notation.[10] Consider three objects, viz. milk, cow, and

---

[10] Figure 5 is inspired by Cram's (1992) interpretation of Fodor's notion of asymmetric dependence.

horse, and four properties, viz. $P_1$, $P_2$, $P_3$, and $P_4$, instantiated by those objects, as shown in Figure 5.

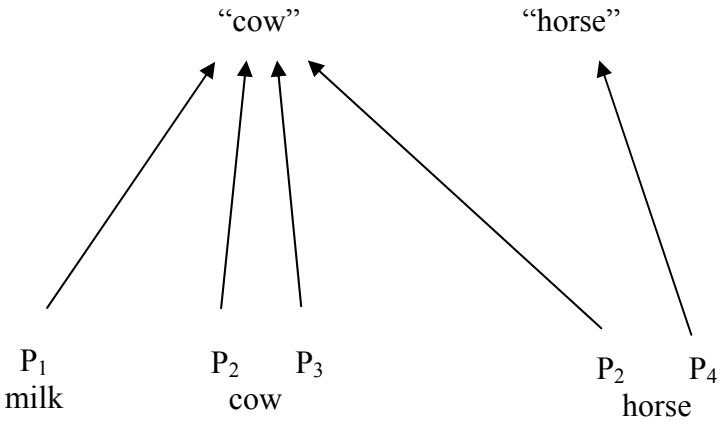


Figure 5: Misrepresentation and disjunction problem.

As depicted in Figure 5, milk instantiates the property $P_1$, cows instantiate the properties $P_2$ and $P_3$, and horses instantiate the properties $P_2$ and $P_4$. Each one of $P_1$, $P_2$, and $P_3$ is a property which, when instantiated in an object is capable (given certain background conditions) of causing a token of "cow." The arrows indicate the nomic connections between properties and mental symbol types. What the properties $P_1$, $P_2$, $P_3$ and $P_4$ correspond to in reality is an empirical question. All we need, for our purposes, is the existence of such properties.

Note that viewed in this way, horse caused "cow" tokens do not create a disjunction problem. This is because, *pace* Fodor, horses sometimes cause "cow" tokens in virtue of instantiating the property $P_2$. To put it in other words, properties which are expressed by a symbol can give rise to misrepresentations. But of course the disjunction problem is still with us. This is indicated, in Figure 5, by the property $P_1$. That is, sometimes some properties which are not expressed by a symbol S can cause a token of S.

Note that we now face two problems. First, we need to prevent the property $P_1$ from entering into the meaning of "cow." But this is not enough. We also need to explain why "cow" means $P_2\&P_3$ but not $P_2VP_3$. This is the problem which Dretske (1986) tried to overcome by appealing to the notion of associative

learning.[11] In that paper Dretske claims that a simple organism can have a primitive misrepresentation capacity if it has (at least) *two* ways of detecting the presence of some substance plus has a form of associative learning.

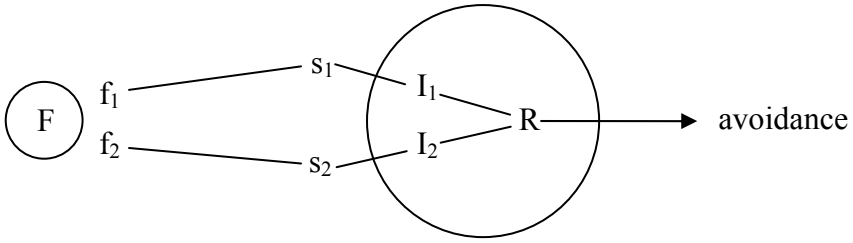


Figure 6: Misrepresentation. Redrawn from Dretske (1986).

Dretske explains what he has in mind as shown in Figure 6. The big circle at the right of Figure 6 represents a simple organism, $f_1$ and $f_2$ are properties of some substance F and $s_1$ and $s_2$ are proximal stimuli. $I_1$ and $I_2$ are internal states of the organism each of which is, given the background conditions, caused by a different chain of events (namely, $f_1 \rightarrow s_1 \rightarrow I_1$ and $f_2 \rightarrow s_2 \rightarrow I_2$). Also both $I_1$ and $I_2$ cause the occurrence of a third state, namely, R.

I will not discuss Dretske's proposal at full length here. I will discuss only his conception of misrepresentation which is important for our purposes. Recall that I have criticized Fodor for thinking of misrepresentation as a special case of the disjunction problem. In other words, he thinks that misrepresentation is a result of tokening of a symbol by a property which that symbol does not express. But as can be seen from Figure 6, Dretske's view of misrepresentation is similar to my view which I have explained above. According to Dretske, misrepresentation occurs when the organism in Figure 6 is presented with an

[11] It is not necessary for my purposes to explain how this problem arises in the context of Dretske's discussion of misrepresentation. But I want to note that it is exactly the same problem faced here. I also think that his solution, based on associative learning, is not successful.

ersatz F (that is, something which instantiates only some of the properties of the real F, say $f_1$). Since the property $f_1$ is sufficient, given the background conditions, to cause R, an ersatz F which instantiates $f_1$ only can cause R and give rise to a misrepresentation. But $f_1$ is not a property which is not expressed by R; it is part of the meaning of R. I think that this conception of misrepresentation is the right one.

Although this is a promising suggestion, the problem of misrepresentation is not solved yet. To illustrate the problem consider a hypothetical 1-dolar detector. Suppose the device accepts both paper money and coins (for simplicity, assume that the device accepts either a single paper banknote or a single coin). When a 1-dolar is inputted, the red light on the device turns to green. If the inputted money is not a 1-dolar then the red light does not change. The question is what does the green light represent? I think the answer should be that the green light represents either a 1-dolar coin *or* a 1-dolar paper banknote. But now consider Figure 6 again. Why shouldn't we say that R represents $f_1$ or $f_2$? But if R represents $f_1$ or $f_2$ then the misrepresentation problem is still with us. I think that my proposal which construes concepts as structured representations can handle this problem.

Let us assume that $R_1$, $R_2$, $R_3$, and $R_4$ are primitive mental representations caused by the properties $P_1$, $P_2$, $P_3$, and $P_4$ respectively. Mental symbols are stored as structured complexes built out of these primitive mental representations. The "cow" symbol, for example, is constituted by two mental representations, namely, $R_2$ and $R_3$. Note that this is a simplified version; in reality the number of constituents of the symbol "cow" is probably much larger. A symbol is tokened by a process similar to the ones postulated by the theories of object recognition. For example, when a person is confronted with a cow, first the visual system computes the mental representations corresponding to some of the properties of the cow (such as $R_2$ and $R_3$ as shown in Figure 2). Then these mental representations are compared with the stored representations of mental symbols and the best matching symbol ("cow" in this case) is activated.

Our two problems, remember, were to explain how misrepresentation is possible and to solve the disjunction problem. Let us start with misrepresentation. As I stated previously the main reason behind misrepresentation is that objects in our environment share some of their properties. So a natural suggestion might be that misrepresentation occurs when an object O, in virtue of having a property P, causes a symbol S, and if O does not possess all the properties which are encoded by S. For example horses sometimes cause "cow" symbols because sometimes (say, under poor visual conditions) our visual system can only detect some properties of horses which are shared by cows. This symbol in turn causes "cow" to be tokened as a result of the matching process. The importance of the matching process for the explanation of misrepresentation is its ability to token a symbol from partial information. In other words the matching process behaves like an inductive inference mechanism.[12] But, as is well known, induction is prone to error. And this seems to be the source of misrepresentation in humans.

I have said that misrepresentation occurs when an object O causes a symbol S and if O does not possess *all* the properties encoded by S. But this is too restrictive. A person's concept of cow might encode the property *has a tail*. When this person confronts a cow which has no tail (say, as a result of an accident) and tokens the "cow" symbol, we do not want to say that a misrepresentation occurs. In fact, like many other concepts, misrepresentation is not a yes or no matter but a matter of degree. There are clear cases of misrepresentation as well as cases which we hesitate to classify it as a case of misrepresentation. But my suggestion can be naturally modified to fit this fact. The modified formulation can be like this: misrepresentation occurs when an object O causes a symbol S and if O does not possess *most* of the properties

[12] By an inductive inference mechanism I mean inferring new (unperceived) properties of an object based on category membership. For example, when we see a red and round object on a tree, we first categorize it as an apple and infer, for example, that it has seeds.

encoded by S. This is a vague definition but so is the notion of misrepresentation.

Even if the account of misrepresentation that I have given above is correct, there still remains the disjunction problem. The disjunction problem is exemplified in Figure 5 with the property $P_1$. To solve the disjunction problem, one has to tell why $P_1$ is not expressed by "cow" even if $P_1$ causes "cow" tokens. It is tempting to say that $P_1$ is not expressed by "cow" because the primitive mental representations which constitute "cow" do not express $P_1$. The situation is depicted in Figure 7.

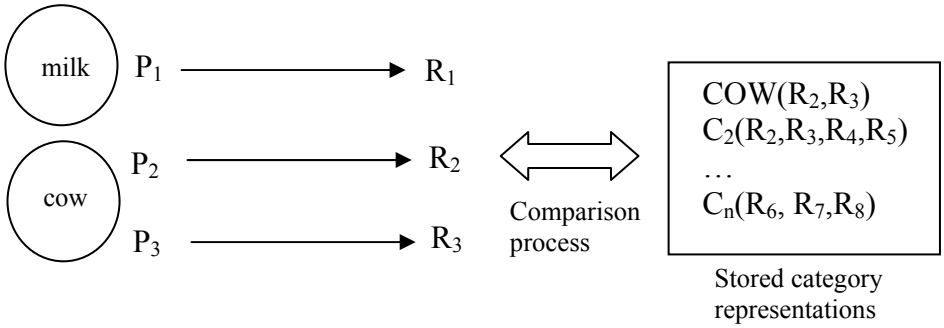


Figure 7: The disjunction problem.

As shown in Figure 7 the primitive mental representation which is caused by $P_1$, $R_1$ in this case, is not part of the concept COW. So even if in some circumstances $P_1$ is capable of causing COW, $P_1$ is not part of the meaning of it.

A careful examination of the above proposal will reveal that this solution to the disjunction problem, in fact, does not depend on the postulation of a complex conceptual structure. Even if the conceptual structure had been atomistic (i.e., even if concepts had no semantically interpretable parts), the same solution would still apply. This is because the heart of the solution depends on dividing mental representations into two levels. There are lower level mental representations (or primitive mental representations) and there are higher level mental representations (such as COW and HORSE). The mental representations at the higher level inherit their semantic properties from the mental

representations at the lower level. This also means a divergence from the causal theories of content. This is because the semantic properties of mental representations at the lower level are determined by their causal relations to the external world. But the semantic properties of mental representations at the higher level are determined only by the semantic properties of their constituents not by their causal relations with distal properties.

One other difference between these two levels of mental representation is the fact that there can be no misrepresentation at the lower level. Lower level mental representations express whatever property they are caused by. Misrepresentation, as a capacity, emerges as a result of some higher level complex processing such as depicted in Figure 2. Similarly the disjunction problem does not arise at the lower level mental representations. Like misrepresentation, the disjunction problem arises only at the higher (or conceptual) level.

### 5.3.3 Conditions 6 and 7

Mental representation as used in cognitive theories has explanatory value. Cognitive science generally explains the behavior of humans and animals by postulating mental representations and processes manipulating these representations. So a theory of representation needs to account for the explanatory role of mental representation.

If, as Fodor thinks, ordinary objects, such as cows, rocks, trees, etc. were represented by arbitrary symbols then almost nothing could be achieved with these representations. According to my proposal concepts corresponding to these ordinary objects are not atomic but structured mental representations. These structured representations carry rich information about the object represented which can be used in various cognitive processes. My proposal does not explain how cognitive processes utilize this information but I can say that the structured representations that I propose at least have the potential resources which can be utilized by cognitive processes.

The last condition in my list is the fact that representation is a determinate relation. When something represents it represents a definite object (or event or situation). I have mentioned two kinds of ambiguity before: structure and mapping ambiguity. Both of these kinds of ambiguity create trouble for isomorphism based accounts. Since my proposal does not rely on isomorphism these ambiguities do not apply to it. However, there is another kind of ambiguity which is relevant for causation based accounts. Below, I will consider this issue.

An object causes the corresponding mental representation through a chain of events. Let us denote the proximal stimulus a cow causes on a particular occasion as P. The causal chain between *cow* and COW can be represented as: *cow* $\rightarrow$ P $\rightarrow$ COW. That is, a cow first causes a proximal stimulus and then this proximal stimulus causes the concept COW. Now, the question is this. Why does COW represent cows but not P? To put the same point in general terms, is there a principled reason to select one of the mediating causes in a causal chain as the content of a mental representation? If we cannot find a principled reason then what a concept represents will be ambiguous. This problem is called the distality problem (Dretske 1981, pp. 156-68).

Remember that according to my proposal a concept represents the objects which instantiate the (distal) properties that cause its constituents. But if we think of the proximal properties that are caused by the distal properties then, again according to my proposal, a concept represents the objects which instantiate these proximal properties. Hence it seems that my proposal leads to ambiguity: it cannot distinguish between distal and proximal properties that cause the same concept. However, I think that the following might be a possible way out. If we consider all the proximal properties (including all the modalities such as vision, audition, etc.) which can cause a concept it is hard to find a single object which instantiates all of these proximal properties. If there are no such objects then the ambiguity created by proximal properties will be prevented. I should say that I am not totally satisfied with this solution, i.e., the distality problem is still an unsolved problem for me.

### 5.3.4 The Pan-semanticism Problem

In this section I will discuss one last condition which is not discussed in the literature as much as the previous ones but which, I think, is important to account for. Antony and Levine (1991) point out a tension between the naturalization project and intentional realism. Naturalization project aims to show that the mind is just a part of the nature; on the other hand, intentional realism is the thesis that there is something special about the mind which other parts of nature lack. So the task facing the naturalist who is also an intentional realist is to explain the properties of mind with naturalistic properties and at the same time to preserve the special status of the mind.

As we have seen before, according to the crude causal theory a symbol S represents the property P if instantiations of P cause S. But this simple formulation (apart from its problems we have seen before) leads to pan-semanticism, that is, the thesis that meaning is everywhere. The argument simply goes like this: causation is everywhere, if causation is enough to have representation then representation is everywhere. Of course this is an undesirable result. Intuitively we feel that representational capacity is peculiar to humans or maybe to higher animals. Crude causal theory, although a naturalistic formulation, cannot manage to preserve the intuition that there is something special about meaning.

Another way to see the problem is to look at simple organisms or devices. For example, the mercury level in a thermometer is caused by the ambient temperature. So, according to the crude causal theory it follows that the mercury level represents temperature. Or consider simple organisms. Dretske (1986) points out that some marine bacteria have internal magnets which align themselves parallel to the earth's magnetic field. These internal magnets are used by the bacteria to avoid oxygen-rich surface water and to move towards water where there is less oxygen. Whatever the function of these magnets, it is clear that these simple organisms have internal states which are nomically connected to some environmental properties. In both cases a simple theory of

representation, such as crude causal theory, will force us to say that very simple organisms or devices have representational capacities. Fodor (Loewer and Rey, 1991, p.256) says that a theory assigning representational capacities to such simple organisms or devices does not create a problem as long as it does not assign beliefs and desires to them. And he thinks that representational capacity alone is not enough to have beliefs and desires.

Intuitions might vary on this point but I think that it is better for a theory not to assign representational capacities to such simple organisms. Or to put it somewhat differently, representational capacity might come in degrees. That is, there might be simple or developed representational capacities. Simple organisms might have simple representational capacities but I do not think that they have representational capacities which are as developed as that of humans. There should be a difference between the representational capacities of humans and simple organisms.

The theory I proposed in this section fares well with this problem. Recall that I have distinguished between primitive and structured mental representations. The pan-semanticism problem applies only to the primitive level. But when the model I have proposed is considered as a whole, no such problem occurs. The semantics of the representations at the conceptual level (that is, the structured mental representations) are not simply given by nomic connections. As I have explained, there should be a two-leveled structure. And this two-leveled structure is something which is *not* abundant in nature.

Also as I have explained, the two-leveled structure works as an inductive inference mechanism. That is, we infer about the objects around us from incomplete information. Induction might sometimes lead to errors, as in the case of misrepresentation, but in general it is a powerful mechanism. It is something we do unconsciously almost all the time. I think that this might also solve the tension between naturalization and intentional realism pointed out by Antony and Levine. The two-leveled architecture that I have proposed does not use any semantic or intentional terms but at the same time it is not something which can be found everywhere.

## 5.4 A Connectionist Implementation

In this section I will give empirical and theoretical evidence in order to show how the model I have given in section 5.2 can be implemented in the brain. The empirical evidence will cover some of the findings in cognitive neuroscience which show how environmental features are encoded in neural patterns. By theoretical evidence I mean the theories of perception which try to construct a general framework based on such empirical findings. Also in the end of section 5.2 I have claimed that the set of primitive representations are rich enough to form the base of (or many of) our concepts. This section can also be seen as supporting this claim. In the context of the connectionist implementation, primitive representations and conceptual representations roughly correspond to neural encodings and combinations of neural encodings, respectively. Since neural encodings can represent a rich class of environmental features, combinations of neural encodings (hence conceptual representations) can represent a rich class of complex environmental features. Finally, in this section I will also illustrate the disjunction problem which will clarify both the disjunction problem itself and my solution to it.

### 5.4.1 Neural Encodings

It is a widespread assumption in perceptual processing that sensory information reaching the senses from environment undergoes several processing phases. The processing begins from the sense organs (such as the eye or the ear) and ends in the sensory-motor areas of the brain. At each stage of this processing the human nervous system constructs representations of the environment. These representations can be thought of as feature detectors. The feature detectors in early processing detect relatively basic features and, as the processing continues, more complex features of the environment are detected.

One of the earliest such discoveries is the work on the ganglion cells in the human retina (Anderson, 1995, p.40, Hubel, 1988, chapter 3). The human retina is composed of several layers. Ganglion cells form the last layer of the human

retina which is connected to the human visual cortex. These cells react in a certain manner to light received by the retina. When there is a steady and diffused light or when there is no light at all, ganglion cells fire at a spontaneous rate. There are two types of ganglion cells which respond differently to the incoming light: on and off type ganglion cells. When light falls on a small sensitive retinal region, on-type ganglion cells increase but off-type ganglion cells decrease their firing rate. On the other hand, when light falls on the surrounding area of the sensitive region, this time, off-type ganglion cells increase but on-type ganglion cells decrease their firing rate.

As the processing continues through the visual system, neurons react to more complex patterns. For example, in one of the earliest such discoveries Hubel and Wiesel (as cited in Anderson, 1995, p.40), in their study of the visual cortex of the cat, found that certain neurons in the cat visual cortex respond to edge shaped figures and certain neurons respond to bar shaped figures, which are known as edge and bar detectors.

More recent research revealed many such cortical neurons which selectively respond to a rich class of features in the environment. In what follows I will briefly summarize some of these findings that are given in Goldstein (1996).[13] The findings that I will review, I think, are enough to show that cortical neurons represent (by responding to) various features of the environment.

Some neurons in the visual cortex respond to specific orientations, to movement, and to the direction of movement (p. 97).[14] There are neurons which respond to specific colors. Area V4 is specialized for color. About 60% of the neurons in area V4 respond to color irrespective of the lighting conditions (p. 108). There are even neurons (in the inferotemporal area of the cortex) which respond to highly specialized features such as pictures of faces (p. 109).

---

[13] I have tried to give examples from all of the five human senses.

[14] This and the following page numbers refer to Goldstein (1996).

In the auditory cortex some neurons are found to respond best to lower frequencies and some others respond best to higher frequencies (p. 356). Irrespective of frequency some neurons respond only when a sound source is located in a particular area relative to the animal's head (p. 360). Also similar to the visual cortex, there are neurons in the auditory cortex that respond to specialized stimuli (p. 364). For example, some cells are found to respond neither to pure tones nor to complex tones but only respond to noises such as jingling and paper tearing. Also some cells are found to respond only when the tone is swept from low to high frequencies and others respond to a tone which is swept from high to low frequencies.

The somatosensory cortex contains neurons that respond to stimulations of the receptors in the skin. It is found that there are neurons in the somatosensory cortex of monkeys' that respond best to an edge oriented horizontally but respond poorly to other orientations (p.475). Also there are other neurons which respond to a movement across the fingertip from right to left but do not respond to a movement from left to right (p. 475).

The senses of smell and taste respond to chemical stimuli. Experiments have shown that different areas in the rats' olfactory bulb are active as a result of stimulation with camphor and with amyl acetate (p. 512). However, the neural code in the olfactory bulb is much more diffused and overlapped compared to other sensory systems. It is thought that maybe higher processing centers have more clear-cut response patterns to different odors. To this end, another experiment has shown that few of the cells in the olfactory bulb respond to only one odor, whereas half of the cells in the orbitofrontal cortex responded to only one odor (p. 513). This implies that neurons in the orbitofrontal cortex are tuned to respond to much more specific odorants than the ones in the lower levels.

Similar experiments are done for the taste system. One experiment showed that different molecules (ammonium chloride, potassium chloride, and sodium chloride) when given as a stimuli to a rat's tongue produced different patterns of responses in the fibers of the chorda tympani nerve (p. 523). Different response patterns of fibers therefore are thought to represent these three different

molecules. Also it was shown that salty, sweet, sour, and bitter substances cause a change in the firing rates of certain neurons in the nucleus of the solitary tract (p. 525).

Although the exact nature of neural encodings are yet to be discovered, it is secure, in the light of recent research, that cortical neurons respond to (or detect) a very rich class of environmental features (such as varieties of color, orientation, movement, luminance, frequency, temperature, pressure, and certain chemicals).

The assumption that human neurons in the human visual cortex respond to environmental properties is also assumed by some of the theories of object recognition. For example Biederman's (1987) recognition by components model, which I described in section 5.2, assumes that volumetric parts and their relations are detected by the visual system. Detection of some primitive shapes lies at the heart of Biederman's theory. Biederman claims that by combining these small set of volumetric parts and relations it is possible to construct billions of objects. And, according to Biederman, stored representations of objects are just combinations of these primitive parts.

Another important set of experimental and theoretical study that relies on neural encodings of environmental features is the feature integration theory of attention (Treisman & Gelade, 1980). According to this theory, humans initially register the features of a given visual stimuli in feature maps. There are different feature maps for color, orientation, luminance, shape, and for other preattentive features. Note that this initial preattentive stage only encodes features of the visual scene. No binding between features is made. In the second stage features are bound together to form coherent objects. How the features are bound together is a difficult problem. One promising idea is temporal binding. Neurons representing the features are bound together if they are correlated in time. But temporal binding does not explain how the neurons that are bound together are selected. According to Treisman & Gelade (1980) attention is the key factor here. The features that are bound together are selected by focused attention.

One important class of evidence supporting feature integration theory of attention comes from the so called illusory conjunctions. Illusory conjunctions

are cases where features of objects are bound wrongly. For example, suppose an image is composed of a red circle and a blue square. Illusory conjunction occurs when a person who sees this image perceives a red square or a blue circle. Treisman (1998) gives two types of examples to support illusory conjunctions. One is about a patient who has severe problems of binding features. The patient is shown very simple displays containing just two colored letters selected from T, X, and O in red, blue, or yellow, and asked to tell the first letter he saw. Even with exposures as long as 10 seconds the patient makes binding errors in more than 35% of the trials. He reports one letter in the color of the other. The second type of example comes from normal people. According to Treisman since attention is needed to bind the features together, she conjectured that if normal people lack enough time for attention they too will too will make binding errors. Treisman's experiments show that even normal subjects make binding errors by putting features (such as shape, color, size, etc.) together in wrong conjunctions when there is not enough time for attention.

### 5.4.2 Binding of Features

The bottom up approach to perception has one important problem that has to be solved. Various features of objects appear to be processed by different parts of the brain forming a distributed representation. Yet we do not perceive a set of features floating around but we perceive coherent objects whose properties are bound together.

Although the exact nature of neural encoding (the representation of environmental features by neural impulses) is not known yet, researchers distinguish between two general types of encoding. The first type of encoding occurs as a change in the firing rates of individual (or groups of) neurons in specific areas of the cortex. I have reviewed some findings in this regard in the previous section. However, this type of encoding is transitory. That is, the change in the firing rates of neurons in response to certain environmental features is not permanent. On the other hand, since it is undisputable that the human brain can store information, there should be another type of encoding

which is permanent. This second type of encoding of information is usually thought to be in the form of permanent changes in the synaptic connections between neurons. For example, according to Hebbian learning, if one neuron is stimulating another neuron repeatedly and persistently, then the strength of the synaptic connection between the two neurons will be increased (which is sometimes stated as "Neurons that fire together, wire together"). Changes in synaptic connections, in contrast to changes in firing rates, are permanent and are thought to be the neural mechanisms underlying long term memory.

In the rest of this section I will give a neurological model of the two-leveled architecture that I proposed in section 5.2. This model will be based on convergence zones framework (Damasio, 1989, Damasio and Damasio, 1994). Convergence zones are neural mechanisms of binding neural activities. Features that are simultaneously activated are bound together. Convergence zones, as proposed by Damasio, do not consist of a single layer but are composed of hierarchical layers. First level convergence zones bind sensory neural encodings caused by environmental features, second level convergence zones bind first level convergence zones, and so on.

How convergence zones (if they exist at all) are actually implemented in the neural architecture of the brain is not known. But it is not impossible that they can be implemented. I will describe one specific computational implementation (Moll and Miikkulainen, 1997) in order to be as precise as possible. Figure 8 depicts the architecture of convergence zones which is used in the simulations of Moll and Miikkulainen.

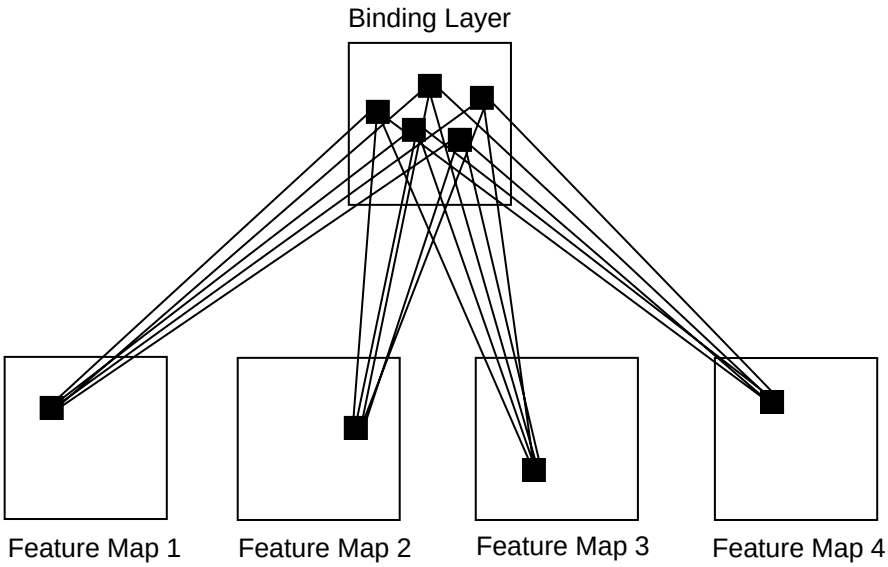


Figure 8: A connectionist implementation of convergence zones.

In Figure 8, there are two layers of neurons: binding layer and the feature map layer. Feature map layer consists of neurons whose activation signals the presence of certain external features. There are separate feature maps for each domain of features. For example, feature map 1 might correspond to color domain and the specific feature shown as a small square in it might correspond to the color *red*. Other feature maps might correspond to other feature domains such as shape or size. The binding layer consists of neurons which have both feed forward and feed back connections to the neurons in the feature map layer. The basic function of the neurons in the binding layer is to bind (in terms of their connection weights) together the simultaneously activated features. How this binding mechanism works is described next.

Activation values of all the neurons and the weight values of all connections take discrete values (0 or 1). Binding proceeds in three steps as follows: initially all connection weights are set to 0. First, some of the neurons in the feature map layer which detect external features are activated (set to 1). Second, a subset of

the neurons in the convergence zone layer is selected randomly[15] and activated (set to 1). Third all the connections between the active neurons in feature map layer and the binding layer are set to 1. The third step binds the neurons in the feature map layer which are active simultaneously. Note that binding takes a single step in this particular simulation.

Retrieval of a stored pattern proceeds as follows. First, all of the neurons in the binding layer are set to 0. A subset of neurons in the feature map layer which correspond to a subset of a stored pattern to be retrieved is activated. These activated units further activate neurons in the binding layer through the connections which are activated during the binding phase. The activated neurons in the binding layer, in turn, activate neurons in the feature map layer. Since a neuron in the feature map layer takes part in representing multiple patterns, at the end of the retrieval process the activated neurons in the feature layer correspond to more than one pattern. The final pattern is selected by retaining only the most activated neuron in each feature map.

What I am trying to show, by giving this connectionist model, is that the two-leveled model that I have given in section 5.2 can be implemented by a cognitively plausible neural architecture. Primitive representations in my model correspond to neurons in the feature map layer. They represent environmental features which cause them. Conceptual representations, on the other hand, correspond to collections of primitive representations which are caused as a result of a kind of similarity matching mechanism. Similarity matching mechanism in the special case of the neural model above is implemented in terms of the special configuration of neurons in the binding layer and their connections to the neurons in the feature map layer.

---

[15] Random selection of neurons might seem to be cognitively implausible, but Moll remarks (in e-mail correspondence) that he "believe[s] there is some evidence that the representations formed in hippocampal area CA3 are seemingly random (although that could also mean that we just don't understand what is going on)."

The basic problem confronting the causal theorist is to solve the disjunction problem. There is usually more than one cause of a mental representation. But, not all of these causes are semantically relevant. So, the causal theorist should find a principled way to distinguish semantically relevant causes from the semantically irrelevant ones. Now let us see how the disjunction problem arises in the context of the neural model that I have given above.

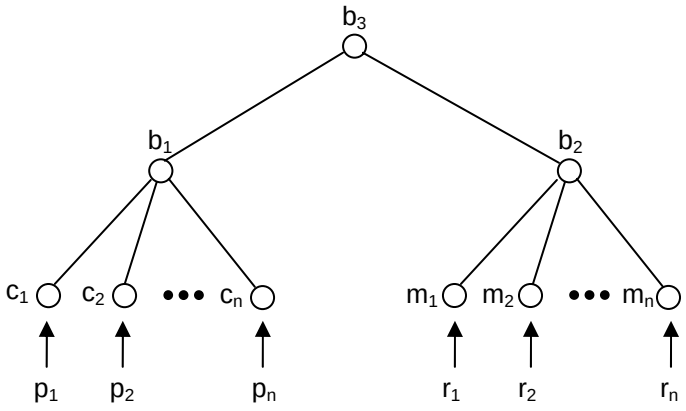


Figure 9: An illustration of the disjunction problem.

In Figure 9, $p_1$, $p_2$, …, $p_n$ and $r_1$, $r_2$, … $r_n$ are environmental features which cause firing of the corresponding neurons $c_1$, $c_2$, …, $c_n$ and $m_1$, $m_2$, …, $m_n$. Neurons $b_1$ and $b_2$ are the binding neurons which bind a set of neurons. Neuron $b_3$, on the other hand, is a higher level binding neuron which binds the binding neurons $b_1$ and $b_2$.[16]

[16] Although, in Figure 5, there is only one level of binding neurons, Damasio's convergence zone framework proposes a hierarchy of binding neurons which is compatible with the connectionist model in Figure 9.

The neural model works as described before.[17] When a subset of environmental features, i.e. a subset of $p_1$, $p_2$, …, $p_n$ (or $r_1$, $r_2$, … $r_n$), is present, they activate neurons in the feature map layer, i.e. a subset of $c_1$, $c_2$, …, $c_n$ (or $m_1$, $m_2$, …, $m_n$), and the activation propagates to the binding neuron $b_1$ (or $b_2$) and then propagates back to the feature map layer and activates all the neurons $c_1$, $c_2$, …, $c_n$ (or $m_1$, $m_2$, …, $m_n$).

Now let us first connect this discussion to concepts and illustrate how the disjunction problem arises. Let us assume that, the set of neurons $c_1$, $c_2$, …, $c_n$ and $m_1$, $m_2$, …, $m_n$ represent two different concepts (for example, COW and MILK respectively). That is, according to this construal, a concept is a set of neurons in the feature map layer. The concepts COW and MILK are further bound together because of being, occasionally, simultaneously activated in a way similar to the binding of neurons which constitute COW and MILK. Let me now illustrate the disjunction problem. Each neuron in the feature map layer represents the properties which cause them (for example $c_1$ represents $p_1$). And the set of neurons which constitute COW (that is, $c_1$, $c_2$, …, $c_n$) represents the properties $p_1$, $p_2$, …, $p_n$. However, sometimes *milk* causes the activation of COW. This is implemented in the neural model in Figure 9 as follows. First a subset of $r_1$, $r_2$, …, $r_n$ causes the activation of a subset of neurons $m_1$, $m_2$, …, $m_n$. Then the activation propagates through $b_2$ and $b_3$ and activates the MILK (that is, the set $m_1$, $m_2$, …, $m_n$) and the COW (that is, the set $c_1$, $c_2$, …, $c_n$) concept. Hence according to the crude causal theory COW not only represents the properties $p_1$, $p_2$, …, $p_n$ but also represents the properties $r_1$, $r_2$, … $r_n$. Intuitively, what we want is a criterion to block the properties which cause COW through the binding neuron $b_3$, from the set of properties which COW represents. By looking at Figure 9 this might seem to be a simple task. However, it is not so easy.

The most intuitive criterion might be something like this. The properties $p_1$, $p_2$, …, $p_n$ cause COW *directly* but the properties $r_1$, $r_2$, …, $r_n$ cause COW

[17] Except that Figure 9 is a simplified version (for only one binding neuron is used for binding each set of features) of Figure 5 in order to focus on the disjunction problem.

*indirectly*. However, the terms 'directly' and 'indirectly' need to be clarified. After all, the subset of properties $p_1$, $p_2$, …, $p_n$ also cause COW indirectly through the activation of $b_1$. One can say that the causal path through the properties $p_1$, $p_2$, …, $p_n$ to COW involves only one binding neuron but the causal path through the properties $r_1$, $r_2$, …, $r_n$ to COW involves more than one binding neurons. This might be a principled distinction but I think that it is too much implementation specific. For example, it might turn out that Damasio's convergence zone model is wrong. In that case the criterion does not work for there would be no binding neurons.

One might also come up with the following criterion. The binding connections between COW and MILK are formed later than the binding connections among their constituent neurons. However, again this criterion is also implementation specific. There might be an implementation in which both types of binding connections are formed simultaneously. Also, any solution which depends on the history of the organism confronts the Swampman problem. Suppose that you are walking in the woods and a lightning strikes a dead tree in a swamp and turns it into a molecule by molecule replica of you entirely by coincidence and out of different molecules. The intuition is that any explanation about mental representations (or concepts) should apply both to you and to your replica. However, if you give an explanation that depends on history (as above) then it cannot apply both to you and to your replica since they have totally different histories.

My proposal is this. The difference between the causation of COW by a subset $p_1$, $p_2$, …, $p_n$ and a subset of $r_1$, $r_2$, …, $r_n$ is that only the former properties cause COW through an inductive inference mechanism. More precisely, the inductive inference mechanism here is a similarity matching process. A subset of $p_1$, $p_2$, …, $p_n$ first activates some of the neurons in the feature map layer. These neurons in turn cause COW through a similarity matching process. In other words, COW (rather then some other concept) is caused by a set of activated neurons in the feature map because COW is the most similar concept to the activated neurons.

My proposal is not specific to the neural modal in Figure 9. Many theories of perception assume such inductive inferences in their models. For example, most of the object recognition theories as I have described in section 5.2, despite their differences, assume that object recognition begins with some cues and ends with the activation of the object category. Perception infers the most likely object category from partial information. Also the connectionist model of Moll and Miikkulainen (1997) assumes that partial information coming through our senses leads to the activation of a more complex stored pattern by undergoing an inductive mechanism.

The information presented in this section also shows that the two-leveled model I have proposed is capable of representing a rich class of concepts. Feature detecting neurons can detect a rich set of primitive features of the environment such as types of color, shape, size, illumination, frequency, odor, pressure, etc. And many types of entities in the world can be represented by a collection of these primitive features. Simmons and Barsalou's (2003, p.454) description of the visual representation of CAT is similar to what I have in mind: "During visual processing of a cat, for example, some neurons respond to line orientations, vertices, and planar surfaces. Others fire for colour, orientation, and direction of movement. The overall pattern of activation across this hierarchically organized distributed system represents the entity in visual perception".

In section 6.2 I will discuss conceptual atomism. I will try to overcome Fodor's arguments against the view that construes concepts as structured entities. Apart from the discussion in that section, I think that the empirical and theoretical evidence in this section is strongly against conceptual atomism and favors concepts as structured entities. This is because the evidence in this section shows that our brains can represent primitive features in the environment and bind these primitive representations to form more complex representations. If we want to find a place for concepts in this system with what are we going to identify the concept CAT, for example, other than a complex/structured representation? There seems to be no place for unstructured/atomic

representations except the primitive ones. But it is simply implausible to identify concepts with such primitive representations.

# CHAPTER VI

# FURTHER ISSUES

## 6.1 Weak and Strong RTM

Naturalization, as it is discussed here, is the attempt to find a naturalistic place for intentional states. What has to be done for this is to describe a physical system which satisfies the conditions placed by folk psychology. Recall that naturalization only shows a possibility; even if a physical system satisfies all of the conditions of folk psychology this does not show that that physical system really is implemented in our brains.

One other philosopher who holds an eliminativist attitude toward intentional states is Stephen Stich. In Stich (1983) his primary aim is to show that folk psychological concepts, such as beliefs and desires, have no role to play in cognitive theories. Stich is not interested in naturalization. He thinks that folk psychological concepts involve properties which do not have any scientific role. The semantic properties of intentional states are such properties for Stich. He thinks that semantic properties play no role in cognitive theories and claims that syntactic properties of mental states are enough (and indeed provide a better paradigm) for building theories about cognition. So he developed an alternative theory which he calls "The Syntactic Theory of Mind" (STM for short):

> The basic idea of the STM is that the cognitive states whose interaction is (in part) responsible for behavior can be systematically mapped to abstract

> syntactic objects in such a way that causal interaction among cognitive states, as well as causal links with stimuli and behavioral events, can be described in terms of the syntactic properties and relations of the abstract objects to which the cognitive states are mapped. More briefly, the idea is that causal relations among cognitive states mirror formal relations among syntactic objects. If this is right, then it will be natural to view cognitive state tokens as tokens of abstract syntactic objects. (Stich, 1983, p. 169)

I find Stich's theory quite a challenge for representational theories of mind. Stich's point can be made with an abstract example I think. Suppose that we have a block box and we know its input output behavior. But we do not know anything about its internal structure. To explain the input-output behavior of the black box we can postulate internal structures and states. If we succeed in explaining the input-output behavior of the black box with some particular internal structure then it is legitimate to claim that we have solved how the black box operates. Of course, there might be more than one type of internal structure which explains the input output behavior of the black box. The only way to eliminate competing hypotheses is to design careful experiments to refute one of the hypotheses. Still, theoretically it is not possible to eliminate all competing hypotheses except one because of the underdetermination problem. Nevertheless, I think, this is the way many theories in cognitive psychology are developed. The important point for our discussion is the fact that although postulating internal states has explanatory roles, postulating internal states with semantic properties seems to play no explanatory role. In other words, to scientifically explain the behavior of the black box, postulating internal states which have only syntactic properties is enough.

But according to folk psychology semantic properties of intentional states are important. Stich (1983) divides representational theories of mind (RTM) into two types according to the role they assign to semantic properties of intentional states. According to strong RTM, mental states have causal roles in virtue of their semantic properties (or content). But how semantic properties contribute to the causal roles of the mental states is problematic. Remember that semantic properties are relational properties. And in causation it is generally accepted that

only the intrinsic properties of objects are relevant. For example the mass of a coin is a causally relevant property of the coin. It might be that that coin is in your pocket. But being in your pocket adds nothing to the causal role of the coin.

Stich calls the second version of RTM weak RTM. Weak RTM drops the strong RTM's claim that mental states have causal roles in virtue of their content and agrees with STM that only syntactic properties of mental states are relevant for their causal roles. The distinguishing characteristic of weak RTM from STM is its insistence that mental states have semantic properties. Weak RTM claims that although the semantic properties of mental states are not causally relevant, they are correlated with the syntactic properties of mental states. Suppose that Mary believes that tomorrow's exam will be difficult. That is, according to RTM, Mary tokens a mental representation in her belief box which means that tomorrow's exam will be difficult. Further suppose that Mary's belief causes her to study harder. According to STM the behavior of Mary (studying harder) is caused by a certain mental state in virtue of its syntactic properties. Weak RTM shares this explanation. But it adds that the mental state which causes Mary's behavior also has a semantic property, namely, it is about the proposition that tomorrow's exam will be difficult.

The question regarding naturalization is whether folk psychology presupposes strong or weak RTM. If it presupposes strong RTM then the problem of how relational properties can be causally relevant needs to be explained. But if it presupposes only weak RTM then it does not jeopardize the naturalization project. Condition 2 (the semantic coherence of intentional states) above corresponds to weak RTM. And I explained how the naturalization problem it creates is solved by the developments in formal systems theory and computation.

### 6.2 Structure vs. Atomism

One of the fundamental characters of my proposal is the fact that concepts are structured entities. In fact, treating concepts as structured entities is the central approach in cognitive science. For example, theories like the classical

theory, prototype theory (Rosch and Mervis, 1975) or the exemplar theory (Medin and Schaffer, 1978) despite their differences, assume that concepts are structured entities.

The reason for this tendency to assume concepts as structured entities is simple. With structured entities you have the resources to explain many phenomena related with concepts, such as, concept learning and categorization. For example concept learning can be explained as building complex concepts out of simpler ones. Or categorization can be explained as checking whether the properties specified in a concept's structure are instantiated by certain objects. For example if the concept BIRD encodes the properties such as flies, has wings, has feathers, etc., we can explain the categorization of an object as a bird in the following way. If the object instantiates all (or some) of these properties then it is categorized as a bird. In this way we can also explain some categorization effects, such as, response time, errors, etc.

On the other hand Fodor is famous for holding an opposite view. He holds that (lexical) concepts have no structure at all, they are atomic. Of course, Fodor is aware of the benefits of a structured view of concepts, and that he has to face the challenge to explain some odd implications of his theory. One such implication is the fact that if lexical concepts have no structure they have to be innate. Fodor calls the argument that leads to this conclusion the "The Standard Argument" (Fodor, 1998). The argument can be put in this way. New concepts are learned by combining previously learned concepts. Previously learned concepts are also learned in a similar way but this process cannot go forever, that is, some concepts must be unlearned. Everybody agrees that there should be some unlearned concepts. But since conceptual atomism assumes that lexical concepts are not structured they cannot be learned by combining other concepts, which implies that lexical concepts (such as ROCK, TREE, CAR, etc.) must be unlearned (or innate). This is of course an odd result which is hard to accept. There have been attempts to solve this problem. For example, Margolis (1998) tries to show how learning might be possible within an atomistic framework.

Despite the problems of conceptual atomism, Fodor continues to defend it since he thinks that there are serious problems with approaches that assume concepts as structured entities. In the rest of this section I will review Fodor's arguments against structured approaches and try to give answers to the problems Fodor raises.

Fodor (1998) considers two popular theories of concepts which construe concepts as structured entities: The classical theory and prototype theory. Let us first look at the classical theory. The classical theory of concepts is the oldest theory of concepts. According to it, concepts are structured mental entities which encode necessary and sufficient conditions for their application. For example the concept of a BACHELOR is a structured entity which is composed out of the concepts UNMARRIED and MAN.

According to Laurence and Margolis (1999), the structure of concepts can be construed with the *containment model*. According to this construal, the constituents of a concept are literally its proper parts. So it is natural to say that the concepts UNMARRIED and MAN are parts of the concept BACHELOR. A natural question is to ask whether the concepts UNMARRIED and MAN are also structured, that is, are composed out of still simpler concepts. Traditionally it is held that concepts are eventually composed out of primitive concepts which are sensory based and have no structure.

The distinguishing character of the classical theory is its treatment of categorization. According to the classical theory an object falls under a concept if it satisfies *all* of the conditions that are specified by its structure. That is category membership is a yes or no matter. If an object does not satisfy one of the conditions then it is not counted in the extension of the concept. For example, if the concept BIRD has the constituents FLIES, HAS-TWO-LEGS and SINGS then the extension of BIRD are the objects which fly, have two legs and sing. If there is an object which flies and have two legs but cannot sing then it does not fall under the concept BIRD.

One serious problem of the classical approach is that it entails the fact that concepts can be given definitions. For example, it would be possible to define

BIRD in terms of the concepts such as HAS-TWO-LEGS, FLIES, SINGS, etc. But it is well known that repeated attempts of philosophers to define concepts such as, KNOWLEDGE, CAUSATION, FREEDOM, etc., have all failed. No philosopher succeeded in providing a definition which is immune from counterexamples. It might be thought that these concepts are too abstract and to define such abstract concepts is difficult. But it has turned out that to define many other concepts such as BIRD, ROCK, CAR, etc., is also problematic. Another problem for the classical theory is that it implies analyticity. For example if the classical theory were true, the statement "birds fly" would be analytic. But Quine's influential arguments against analyticity convinced many philosophers that there are no analytic statements. These kinds of criticisms led the decline of the classical theory of concepts. But it is possible to solve the problems created by the classical theory without giving up the assumption that concepts are structured entities.

The digital character of category membership has been criticized also by cognitive scientists. Empirical research on categorization (such as mentioned in Rosch and Mervis, 1975) has revealed that human categorization is not digital but graded and fuzzy. That is, humans judge some objects as better (or more typical) members of a category than some others. For example, sparrows are generally judged to be "better" members of the category BIRD then ostriches.

The Prototype theory has been developed to account for such findings. It is important to understand the differences between the classical theory and the prototype theory. Although the prototype theory has been developed as an alternative to the classical theory, it is not a complete denial of the classical theory. There are important commonalities between the two. For example both theories assume that concepts are structured entities. Also both theories assume that the constituents of concepts, in the final analysis, are sensory based. The essential difference between the two emerges in categorization. Both theories assume that categorization is a comparison process but whereas in classical theory the comparison process has a digital nature, in prototype theory the comparison process has a probabilistic nature. In prototype theory a given

instance is assumed to be compared against a mental representation of a prototype (which is a weighted list of features) according to some similarity measure. The result of the comparison process returns a result which indicates the similarity (or typicality) of the instance to the prototype. Different similarity measures can be specified in order to maximize the explanatory power of the theory. So according to prototype theory, to be classified in a certain category an object need not have all the features encoded in the prototype. Also the comparison process produces graded results. That is, some objects match more numbers of (or more weighted) features of the prototype than others. And this fact explains the subject's typicality judgments.

Fodor's main argument against the prototype theory is based on the principle of compositionality. Fodor assumes that concepts, like language, are productive (there are infinitely many concepts that one can entertain) and systematic (if you can think of Mary is taller than Tom, you can also think of Tom is taller than Mary). And compositionality is the best explanation of productivity and systematicity. Compositionality of concepts refers to the fact the content of a concept is determined by the syntax and the content of its constituents. I agree that compositionality is an important property of concepts but I do not agree that the prototype theory cannot satisfy the compositionality requirement.

The first claim of Fodor states that "indefinitely many complex concepts have no prototypes; a fortiori they do not inherit their prototypes from their constituents" (Fodor, 1998, p.100). Fodor gives examples of boolean concepts such a NOT A CAT. He claims that NOT A CAT has no prototype, that is, there is no object which is a typical NOT A CAT. For example clouds are good examples of the concept NOT A CAT but so too are phones or computers. The problem is that the features of the concept NOT A CAT are too large and heterogeneous and so there cannot be a weighted list of these features. But, here, Fodor treats "NOT" as a concept. And this is what generates the problem. It is true that prototype theory has difficulties explaining logical connectives but it is also true that any theory (maybe with the exception of conceptual role theories)

has difficulties in explaining logical connectives. Fodor (1990b, p.110) admits that his theory has problems too. On the other hand an alternative is not to treat logical connectives as concepts. For example, Prinz (2002, p.288) proposes that instead of a concept, negation can be understood as an operator. It just reverses the similarity measure. Then something falls under NOT A CAT if it does not fall under CAT. That is, NOT A CAT is derived from CAT by reversing the similarity measure. In this way the prototype theory predicts that NOT A CAT has no prototype: since many objects will not instantiate any of the properties of a CAT, those objects will have the same rank in terms of being NOT A CAT. Also if some objects, such as dog, instantiate some of the properties of CAT (such as being four-legged) then those objects will be less typical NOT A CATs then, for example, clouds. And this is an intuitive result: a cloud is a typical NOT A CAT then a dog. Lastly, there is no problem of storing the too large and heterogeneous list of features of NOT A CAT since what is stored is only the CAT prototype.

The other problem is formulated by Fodor in this way: "there are indefinitely many complex concepts whose prototypes aren't related to the prototypes of their constituents in the ways that the compositional explanation of productivity and systematicity requires" (Fodor, 1998, p.100). Fodor also calls this problem "The Pet Fish Problem." Fodor thinks that prototypes do not compose respecting the compositionality principle. For example, goldfish is a prototypical example of the PET FISH concept, however, it is a poor example of both PET and FISH. In other words the prototype of PET FISH is not determined by the prototypes of its constituents. So, the content of PET FISH is not a function of its syntax and the content of its constituents. Compositionality principle is violated. Some may think that it is normal for the compositionality principle to be violated in some cases (idioms are typical examples). But Fodor thinks that PET FISH is not an exceptional case. If it is indeed the general case it is a serious problem for the prototype theorist since she will have to give up the compositionality principle which is an important resource for explaining the productivity and systematicity of thought. At this point I agree with Fodor that

compositionality is an important principle. But I don't think that the PET FISH example is a general case. May be the concept PET FISH is not a compound concept. That is, we probably do not learn it by combining PET and FISH. Imagine a person who does not have a PET FISH concept but who has the concepts PET and FISH. Now if you ask her what kind of things might fall under PET FISH, probably she will not give goldfishes as a typical example, but rather try to combine some typical pets and fishes. Moreover, there are many compound concepts whose contents are determined by the prototypes of its constituents. Here is a few of them: SLEEPING DOG, WALKING CAT, THE CAT ON THE MAT, RED CAR, etc. For example, try to image a typical RED CAR and compare your idea with a typical RED and a typical CAR. I guess that your typical imagination of a RED CAR will be composed of your typical RED and your typical CAR. That is, I don't agree with Fodor that prototypes do not combine compositionally.

Theories which postulate concepts as structured entities surely are not unproblematic. But I think that most of these problems can be solved in one way or another. In short, Fodor's worries do not convince me to discard structural theories altogether and to support an atomistic theory.

### 6.3 Description Theories of Reference

I think that my proposal has some similarity to the description theory of reference since both of them claim that the reference of a symbol is the set of objects which satisfy a set of properties. However, description theories are not favorable these days because of some influential criticisms. In this section I will briefly introduce the basic idea of description theories and try to reply to some of the criticisms.

The central question for a theory of reference is this: "In virtue of what do the terms in a natural language refer to what we intuitively think they refer to?" Not all terms in a natural language refer; for example, prepositions are usually are thought not to refer. However it is generally agreed that some terms, especially proper names, are referring expressions par excellence. For example

the proper names "İstanbul," "Aristotle," or "Mount Everest" refer to certain objects or individuals.

According to description theories of reference (developed by Frege and Russell) proper names refer via the *descriptive content* associated with that name. In other words a proper name refers to an object in virtue of the fact that that object satisfies the descriptive content associated with that name. The descriptive content associated with a name may be different in different speakers. Also the same name might be associated with a different descriptive content in different times by the same speaker. For example one person might associate the descriptive content "The last great ancient philosopher" and another person might associate the descriptive content "The pupil of Plato" with the term "Aristotle."

Description theories of reference are especially successful when they are used as a theory of meaning. According to the description theory of meaning, the meaning of an expression is its associated descriptive content. This theory successfully deals with problems posed by co-referring expressions and expressions which refer to non-existent things compared to an older theory of meaning (known as Millianism) according to which the meaning of an expression is just its reference.

The problem of co-referring expressions is also known as the Frege's Puzzle. Frege argued that if meaning were just reference then the two expressions "the morning star" and "the evening star" should have the same meaning (assuming that the expressions "the morning star" and "the evening star" have the same reference, namely the planet Venus). But it is clear that they have different meanings, for the sentences which contain the expression "the morning star" will change their meaning if we replace the expression "the morning star" with the expression "the evening star" in intentional contexts.

But a description theory can explain this fact by saying that while the two expressions have the same reference they have different descriptive content (or if we use the Frege's term, they have different *senses*).

Secondly, consider the sentence "Pegasus does not exist." Since the term "Pegasus" does not have a reference, the sentence will be meaningless if we assume that meaning is only reference. But intuitively the sentence is a meaningful one and the description theory can explain this fact. For if the meaning of the term "Pegasus" is its descriptive content, namely "a winged horse," then we can say that the initial sentence is meaningful since it says that "A winged horse does not exist."

Description theories of reference have important virtues but these days they are not fashionable because of the powerful objections raised against them especially by Kripke (1980) and Putnam (1975).

Suppose that a particular person associates the descriptive content "The last great ancient philosopher" with the expression "Aristotle." If descriptivism were correct then a sentence such as "Aristotle was not a philosopher" would be a contradiction for that person. For it says that "The last great ancient philosopher is not a philosopher." But this is not the case at least according to the intuitions of some philosophers.

Or consider the objections that are grouped under the name "ignorance and error." Suppose a person associates the descriptive content "philosopher" with the expression "Aristotle." Then according to descriptivism that person cannot refer to Aristotle since his descriptive content refers to all of the individuals who are philosophers. Similarly suppose that the same person additionally believes that Aristotle is the pupil of Socrates. Then according to descriptivism, when that person uses the expression "Aristotle" he refers not to Aristotle but to Plato. But in both cases "Aristotle" manages to refer to Aristotle despite the ignorance and error in the descriptive content.

Clearly these objections rest on intuitions, as it happens all the time in philosophy. One might resist these intuitions. For example one might say that if a person associates the descriptive content "The last great ancient philosopher" with the expression "Aristotle" then to say that "Aristotle was not a philosopher" would indeed be a contradiction for that person. Or one might say that if a person associates the descriptive content "philosopher and pupil of Socrates" to the

expression “Aristotle” then that person really refers to Plato when he utters the word “Aristotle.” But some philosophers (for example, Searle, 1983, Jackson, 1998) think that description theories can cope with such objections. As Jackson points out, the problem of the objections to description theory of reference grouped under the name “ignorance and error” is that they under-describe the descriptive content. For example in the case of a person who believes only that Aristotle is a philosopher, the description content of the term “Aristotle” is not only being a philosopher. For that person also knows that other people in his community also use the term “Aristotle” to refer to Aristotle. So that person’s descriptive content for “Aristotle” must at least contain both “being a philosopher” and “the person who others in my community refer by the term ‘Aristotle’.” Or consider the case of error. I might believe that Aristotle was the pupil of Socrates. Then is it the case that I do not refer to Aristotle when I use the term “Aristotle”? Intuitions might vary at this point. But I am inclined to think, with Searle, that the descriptive content of proper names is not a single description but a cluster of descriptions. And the reference of a term is the object which satisfies *most* of the descriptions in the cluster. That is, although I may believe some false propositions about Aristotle, nevertheless, I can still refer to Aristotle because of the rest of the content that I associate with the term “Aristotle.”

I think the more important objection raised against description theories of reference is the “passing the buck” objection (Devitt, 1996):

> Description theories are *essentially incomplete*. A description theory explains the reference of a word by appealing to the application of descriptions associated with the word. So the theory explains the reference of the word by appealing to the reference of other words. How then is the reference of those other words to be explained? Perhaps we can use description theories to explain their reference too. This process cannot, however, go on forever: There must be some words whose referential properties are not parasitic on those of others… . Description theories pass the buck. But the buck must stop somewhere. (p. 159)

I agree with Devitt that the buck must stop somewhere. But although description theories seem to be incomplete in this respect it does not follow that description theories does not say anything valuable. The situation is very much like the classical or prototype theory of concepts. They too claim that concepts are structured representations composed of simpler representations. So they too explain the reference of a concept by appealing to the reference of simpler representations. But generally they are silent on how those simpler representations acquire references. But again this does not show that the classical theory or the prototype theory say nothing valuable. Construing concepts as structured entities has important explanatory virtues. For example you can explain categorization or concept acquisition. Similarly description theories are able to explain some semantic phenomena such as Frege's puzzle as we have seen before.

If we return to my proposal it can be seen as a species of description theory. The descriptive content of an expression corresponds to the structure of a concept. The structure of a concept contains primitive representations representing properties. If I stopped at this point the passing the buck objection would be applicable to my proposal too. But I claim that the buck stops at the primitive representations. The reference of primitive representations is not determined by still more primitive representations; rather their reference is determined by their causal relations to the external world.

# CHAPTER VII

# CONCLUSIONS

In this thesis my contribution consists of two parts. First I have developed an account of the methodology of the naturalization project and second I have proposed a naturalistic account of representation.

I have claimed that naturalization of representation consists of two steps. In the first step the analysis of the notion of representation is given by conceptual means. And in the second step, a naturalistic system is proposed which purports to satisfy the conditions set out in conceptual analysis. A naturalization proposal is then accepted to be a successful one to the degree that it satisfies these conditions.

I have also proposed a naturalistic system which I claim can satisfy some of the important conditions for representation. Surely these are not the only conditions which should be satisfied by a successful naturalization proposal. Mental representation lies at the heart of human cognition such as language, learning, concepts, etc. As such, my account is expected to explain many phenomena related to these issues. I do not claim that my two-leveled model can explain everything but I claim that it explains some of the important conditions and I am optimistic that it can be further developed to account for other aspects of human cognition.

Also it might be thought that my proposal is a species of empiricist theories of meaning since it identifies the content of a mental symbol in terms of its causal relations to the world. But this would be a hasty conclusion to draw. I do

not claim that *all* mental symbols acquire content in accordance with my proposal. In particular, I do not want to claim that my account extends to mental symbols for logical connectives, mathematical objects, or other abstract entities. I should be pleased if it did extend to them, but showing that would take a lot more further work. If my proposal can explain how an important class of mental symbols, viz., those referring to empirical items, acquire content, or if it can reveal at least *part* of the process of content fixation, my task will be accomplished.

## REFERENCES


Anderson, J. R. (1995). *Cognitive psychology and its implications* (4th ed.). New York: Freeman.

Antony, L. & Levine, J. (1991). The nomic and the robust. In B. Loewer & G. Rey (Eds.), *Meaning in mind: Fodor and his critics* (pp.1-16). Oxford: Blackwell.

Aydede, M. & Güzeldere, G. (2005). Cognitive architecture, concepts, and introspection: an information-theoretic solution to the problem of phenomenal consciousness. *Noûs*, 39(2), 197-255.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review*, 94, 115-147.

Biederman, I. (1995). Visual object recognition. In S. F. Kosslyn & D. N. Osherson (Eds.). *An invitation to cognitive science, 2nd edition, Volume 2, Visual Cognition*. MIT Press.

Brentano, F. (1874). *Psychology from an empirical standpoint*, London: Routledge and Kegan Paul.

Churchland, P. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy*, 78(2), 67-90.

Cram, H. R. (1992). Fodor's causal theory of representation. *The Philosophical Quarterly*, 42, 56-70.

Crane, T. & Mellor, D. H. (1990). There is no question of physicalism. *Mind*, 99, 185-206.

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

Cummins, R. (1996). *Representations, Targets, and Attitudes*. Cambridge: MIT Press.

Damasio, A. R. (1989). The brain binds entities and events by multiregional activation from convergence zones. *Neural Computation*, 1, 123-132.

Damasio, A. R. & Damasio H. (1994). Cortical systems for retrieval of concrete knowledge: the convergence zone framework. In C. Koch & J. L. Davis (Eds.), *Large Scale Neuronal Theories of the Brain* (pp. 61-74), Cambridge, MA: MIT Press.

Devitt, M. (1996). *Coming to our senses*, Cambridge, Cambridge University Press.

Dennett, D. C. (1971). Intentional systems. *Journal of Philosophy*, 68, 87-106.

Dennett, D. C. & Haugeland, J. (1987). Intentionality. In R. L. Gregory, (Ed.), *The Oxford companion to the mind* (pp. 383-386): Oxford University Press.

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

Dretske, F. (1986), Misrepresentation. In R.J. Bogdan, (Ed.), *Belief: form, content and function* (pp.17-34). Oxford: Clarendon Press.

Dretske, F. (1989). Reasons and causes. *Philosophical Perspectives,* 3, 1-15.

Edelman, S. (1997). Computational theories of object recognition. *Trends in Cognitive Sciences*, 1(8), 296-304.

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

Fodor, J. (1990a). *A theory of content and other essays*. Cambridge, MA: MIT Press.

Fodor, J. (1990b). Fodor's guide to mental representation. In Fodor, 1990a.

Fodor, J. (1990c). Semantics, Wisconsin style. In Fodor, 1990a.

Fodor, J. (1990d). "A Theory of Content II." In Fodor, 1990a.

Fodor, J. (1998). *Concepts: where cognitive science went wrong*. Oxford: Oxford University Press.

Gallistel, C. R. (2001). Mental representations: psychology of. In *Encyclopedia of the Social and Behavioral Sciences*, New York: Elsevier.

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

Goldstein, E. B. (1996). *Sensation & perception* (4th ed). New York: Brooks/Cole Publishing Co.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman (Ed.), *Problems and projects* (pp. 437-447), New York: Bobbs Merrill.

Hubel, D. (1988). *Eye, brain, and Vision.* Retrieved September 14, 2006, from Harvard University, Department of Neurobiology web site: http://neuro.med.harvard.edu/site/dh/bcontex.htm

Hummel, J. E. & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, 99, 480-517.

Jackson, F. (1998). Reference and description revisited. *Philosophical Perspectives*, 32, 201-218.

Jackson, F. (1998). *From metaphysics to ethics: a defense of conceptual analysis*, Oxford: Clarendon press.

Kripke, S. (1980). *Naming and necessity*. Oxford: Blackwell.

Lewis, D. (1970). How to define theoretical terms. *Journal of Philosophy*, 67, 427-446.

Lewis, D. (1972). Psychophysical and theoretical identifications. *The Australasian Journal of Philosophy*, 50, 249-258.

Loewer, B. & Rey, G. (1991). *Meaning in mind: Fodor and his critics*. Oxford: Blackwell.

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

Mackie, J. (1974). *The cement of the universe*. Oxford: Clarendon Press.

Margolis, E. (1998). How to acquire a concept? *Mind & Language*, 13, 347-369.

Medin, D.L. & Schaffer, M.M. (1978). Context theory of classification learning, *Psychological Review*, 85, 207-238.

Moll, M. & Miikkulainen, R. (1997). Convergence-zone episodic memory: analysis and simulations. *Neural Networks*, 10(6):1017–1036.

Palmeri, T.J. & Gauthier, I. (2004). Visual object understanding. *Nature Reviews Neuroscience*, 5, 291-303.

Petitt, P. (1993). A definition of physicalism. *Analysis*, 53(4), 213-23.

Prinz, J. (2002). *Furnishing the mind.* Cambridge, MA: MIT Press.

Putnam, H. (1992). *Renewing Philosophy*. Cambridge, MA: Harvard University Press.

Putnam, H. (1975). The meaning of 'meaning'." In K. Gunderson (Ed.), *Language, mind and knowledge* (pp. 131-193). Minneapolis, University of Minnesota Press.

Rosch, E. & Mervis, C. B. (1975). Family resemblances: studies in the internal structure of categories. *Cognitive Psychology*, 7, 573-605.

Searle, J. (1983). *Intentionality*, Cambridge: Cambridge University Press.

Sellars, W. (1956). Empiricism and the philosophy of mind." In H. Feigl & M. Scriven (eds), *Minnesota Studies in the Philosophy of Science* (vol. 1): University of Minnesota Press.

Simmons, K., & Barsalou, L.W. (2003). The similarity-in-topography principle: reconciling theories of conceptual deficits. *Cognitive Neuropsychology, 20*, 451-486.

Stich, S. (1983). *From folk psychology to cognitive science*. Cambridge, MA: MIT Press.

Stich, S. P. & Laurence, S. (1994). Intentionality and naturalism. In P. A. French & T. E. Uehling (Eds.), *Midwest Studies in Philosophy*, v. 19, *Naturalism* (pp. 159-182): University of Notre Dame Press.

Treisman, A. & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, 12, 97-136.

Treisman, A. (1998). Feature binding, attention and object perception. *Philos. Trans. R. Soc. Lond,* B 353, pp. 1295–1306 .

Wilder, R. (1965). *Introduction to the Foundations of Mathematics*. New York: John Wiley & Sons.

## CURRICULUM VITAE

### PERSONAL INFORMATION

Surname, Name: Aytekin, Tevfik
Nationality: Turkish (TC)
Date and Place of Birth: 14 March 1972 , İzmir
email: tevfik_aytekin@yahoo.com

### EDUCATION

| Degree | Institution | Year of Graduation |
|---|---|---|
| MA | METU Philosophy | 2003 |
| MS | Hacettepe U. Computer Science | 2001 |
| BS | Bilkent U. Computer Science | 1994 |

### WORK EXPERIENCE

| Year | Place | Enrollment |
|---|---|---|
| 2004 - Present | Yeditepe U. Computer Science Dept. | Research Assistant |
| 1997 - 2004 | METU Informatics Institute | Project Manager |
| 1994 - 1997 | Hacettepe U. Computer Science Dept. | Research Assistant |

### PUBLICATIONS

Aytekin, T., Sayan, E. "How Not to Misrepresent Misrepresentation," European Society for Philosophy and Psychology Conference, Lund University, Sweden, 11- 14 August, 2005.

Aytekin, T., “Modeling Multiplication Fact Retrieval: The Effect of Noise.” In Proceedings of the European Cognitive Science Conference, 2003.

Aytekin, T., “A New Theory of Content.” MA Thesis (2003).

Aytekin, T., “Computational Modeling of Cognitive Arithmetic.” MS Thesis (2001).

Aytekin, T., Korkmaz, E.E., Guvenir, H.A. "An Application of Genetic Programming to the 4-Op Problem Using Map-Trees." In Proceedings of the workshop on Evalutionary Computation. In Association with 7th Australian Joint Conference on Artificial Intelligence. Xin Yao (Ed.), Armidale, Australia, 1994.