I. INTRODUCTION

A. Psychological Explanation and the Contents of Our Thoughts

When we attempt to explain the behavior of human beings, we often make reference to the mental states of the humans whose behavior we are trying to explain. And when making reference to these mental states, we often describe the specific character of the states in question. Such descriptions normally specify what type of mental state the subject is in (e.g., a belief, desire, or fear state). In addition, such descriptions often contain a characterization of the mental state which is intended to differentiate the state in question from other mental states of the same general type. Typically, we accomplish this by identifying (in a public language) what the state in question is about (i.e., we identify the object[i] of the state), and in most cases, by describing what the state says about its object (what specific fact the state asserts of its object or how the state represents its object). Explanations of human behavior which invoke the existence of the kinds of mental states just described are called 'intentional explanations', and the description of the particular way in which a mental state represents its object is commonly called a description of the state's 'content'. Putting these two ideas together, the content invoked by intentional explanations is sometimes called 'intentional content'.

Intentional explanations are given in a variety of contexts. The ascription of mental states with specific content is a part of everyday folk explanations of behavior, e.g., "John opened the refrigerator because he wanted to get his sandwich." Content ascriptions also appear in the generalizations of social psychology, e.g., "The widespread acceptance of the institutionalization of sexism is explained by the fact that the sexist imagery prevalent in our society affects the content of people's beliefs about social institutions in such a way that people are inclined not to see sexism where it is built into our social institutions."[ii] Such ascriptions appear in cognitive psychology, as well, e.g., "Changes over time in infants' reaching and grasping behavior are best explained by assuming that infants are beginning to understand that individual physical objects persist through time." Similarly content-laden explanations are given in most, if not all, areas of psychology. But what are we to make of such talk? What is content, where does it come from, and in what way do all of the examples cited

above appeal to content in explaining people's behavior?

To begin answering the question how content is determined, we might take the content of a mental state to be a function (at least partly) of the content of the constituent elements of a subject's mental state. And we can begin to understand what the content of a particular constituent of a mental state is by asking what that constituent element refers to, i.e., what actual object(s) the constituent is supposed to correspond to or be about.

Returning to the sample explanations given above, consider the ways in which the referent of a mental state constituent (i.e., the thing to which the constituent corresponds) could play an explanatory role in a psychological explanation. In the first example, when John believes his sandwich (one which he previously made) is in the refrigerator, something in his mind corresponds in some way to the sandwich. This correspondence is part of the explanation of why he opens the refrigerator (there being a host of other factors, such as John's desire to get the sandwich, which also play a role in the explanation of why John opens the refrigerator door). A similar situation exists in the latter two examples. In the second case, subjects' belief that social institutions are just is the result of their being in, or having been in, mental states (specifically, perceptual states) with particular contents. The explanation assumes that the subjects have mental state constituents which correspond to men and women (as individuals and as groups) as elements of the relevant perceptual states. The reason the relevant perceptual states can affect the gender-related attitudes of the subjects, and consequently the subjects' beliefs about social institutions, is because the subjects have constituents in their perceptual states which correspond to the men and the women depicted in the sexist images which gave rise to those perceptual states. In the third sample explanation, the various, newly emerging behaviors of the infant subjects (e.g., looking in hiding place B for an object which they watched being hidden at B, but which had in the past been hidden in hiding place A instead[iii]) result from their changing ideas about the items which correspond to their mental constituent which refers to objects. The infants now know something new about a certain set of items (i.e., objects, in general) which they did not know before. I am not claiming that any of the preceding psychological explanations are adequate for any particular, actual cases. I only give these examples to illustrate how the appeal to the reference of a constituent of a mental state might make up part of a psychological explanation.

It is a bit cumbersome to talk about the constituents of mental states. For convenience' sake, I assume from here on out that the constituents of mental states are roughly analogous to the terms of a public

language. Further, it is assumed that subjects string these terms together in thought to construct mental sentences, and that these sentences express the contents of subjects' thoughts. The idea that there is a language of thought analogous in many ways to a public language is most closely associated with the work of Jerry Fodor (cf. Fodor 1975). A more detailed discussion of the assumption that there is such a language of thought (LOT, hereafter) is included below. For now, however, being able to speak of terms in LOT will make it much easier to talk about the reference of mental state constituents.

The sample explanations given above are incomplete. There is, for example, much more to the explanation of why John opens the refrigerator door than simply the correspondence between John's LOT term 'my sandwich' and the actual sandwich in the refrigerator. Still, the force of each explanation rests partly on the fact that a particular term in the relevant subject's LOT corresponds to a certain object or class of objects. Taking a page from the philosophy of language, the object or class of objects to which an LOT term corresponds is called the term's 'extensional content' or its 'referential content' (or, more simply, the term's 'extension' or 'reference'. More on this terminology below.)

We should note the narrowness of a focus on extensional content, for the term 'content' has been used in a variety of different ways. Especially within philosophy, there is an acute awareness of the different kinds of content an intentional item (a word, sentence, thought, etc.) might have. With respect to linguistic content, Gottlob Frege distinguished between three different kinds of meaning, or content, a term might have: the particular ideas an individual experiences when using or hearing a term, the abstract sense of the term (its intension), and the referent of the term (its extension) (Frege, 1980, pp. 202-203). In the philosophy of mind and philosophy of psychology, the content of a thought has been identified with analogues of each of these three. It is well recognized that when one thinks about horses, one has certain subjective ideas or experiences (e.g., memories of past rides). In addition, it is often assumed that a thought about horses involves a relationship to an abstract idea of horses (often called a 'concept'), which exists independently of anyone's individual psychological experience. Lastly, when one thinks about horses, it seems that one thinks about horses, i.e., about the things themselves. Here the collection of actual horses is called the 'extension' of the LOT term 'horse'.

Extensional psychology claims that we should include talk about the extensions of LOT terms (or states) in psychological explanations of human behavior. An extensional approach can do so in at least two ways. Extensional theorists can talk about the extensions of LOT terms to explain behavior in specific examples, e.g., 'Donna, who is blind, learned to use the term "ball" in an appropriate way by haptically

exploring the ball her father gave her' (cf. Landau and Gleitman 1985). However, we should also include as extensional a theorist's taxonomization of LOT terms, and the mental states constructed out of those terms, according to the types of things to which those LOT terms refer (cf. Sterelny 1990, pp. 101-104, or Burge's description of the way Marr's computational theory of vision individuates states, Burge 1986, pp. 32-33).[iv] Explanations which appeal explicitly to the actual extensions of terms may be importantly different than those which appeal only to an extension-based taxonomy of states or terms. This noted, I use the terms 'extension', 'reference', 'extensional content', 'referential content', 'intentional content' and 'representational content' interchangeably where I'm confident that doing so will cause no confusion. (Apologies to the reader for the length of the list, but it is merely a reflection of the diversity of terminology used by philosophers when discussing these matters.)

Pursued in either of its forms, extensional psychology obviously stands in need of a theory which assigns extensions to LOT terms. In the present work, I attempt to provide some of the necessary theoretical underpinnings for the use of content as an explanatory tool in psychology by developing such a theory of extension. The theory developed here, the Best Test Theory (BTT, hereafter), is not, however, wedded to one specific view regarding which type of content is most important in psychology. In particular, the reader should not take BTT's focus on extensions to imply that psychology is entirely extensional in nature. The Best Test Theory's subject matter is further limited in that BTT only explicitly identifies the extension of a limited class of LOT terms, natural kind terms.

In the process of identifying the extension of a natural kind term of LOT, BTT identifies the general types of things to which a given term refers. Thus, by the way of assigning extensions to natural kind terms in a subject's LOT, BTT also provides a way of identifying what many might think of as the intensions of these same LOT terms. (Intensions, recall, are abstract meanings, or Fregean senses, associated with terms.) As a theory of extension, BTT states that the extension of a natural kind term in LOT consists of the members of the natural kind for which the concept associated with that LOT term provides the best test. If it were construed as a theory of intension for natural kind terms in LOT, BTT would assign as the intension of term t the natural kind, considered as an abstract type, to which all of the members of the extension of t belong and all of whose current members belong to the extension of t. Thus, BTT can be seen as friendly to (or, at least, not unfriendly to) intensional psychology; for intensional psychology, characterized generally, is the view that in giving psychological explanations, one should make reference to the abstract ideas or meanings of a

subject's thoughts.

I should not paint too conciliatory a picture. There are surely versions of intensional psychology in which BTT has little or no theoretical place. An example of such a version of intensional psychology might be an intensional psychology which identifies the content of a mental state with the functional role of that state characterized without reference to the extension of any of that state's components (cf. Block 1986, Devitt 1989). According to such a view, the functional role of a state determines the state's extensional content as well as the extensional content of any of the state's components. Obviously, BTT has no role to play in such theories. I do not here offer any critical evaluation of functional role theories of mental content or the accompanying intensional approach to psychology. I am at pains only to point out that BTT leaves room for some types of intensions should we have use for them in the empirical enterprise of developing a psychology of human beings.

The psychological explanations with which the chapter began come from folk psychology, social psychology, and cognitive psychology, respectively. I have claimed that each of the sample explanations makes essential reference to the extension of certain of the subjects' LOT terms. This may suggest to the reader that the development of BTT is motivated by the very strong assumption that the areas of psychology from which these sample explanations come are entirely extensional. I will not argue that any one of these three types of psychological theories taxonomizes psychological states solely according to extensional content. Instead, I assume only that some of the explanations and generalizations of each of these theories taxonomize states by extensional content.

The sample explanations given at the outset serve three purposes. They are intended to show that a variety of psychological explanations rest for their force on the existence of content. This motivates a general interest in content. Secondly, the sample explanations are supposed to provide examples of psychological explanations in which the extensions of subjects' LOT terms appear to play an important explanatory role. This motivates the development of a theory of extensional content. And thirdly, the sample explanations illustrate an important criterion of adequacy which BTT must meet. By showing us what kinds of psychological explanations, from which areas of psychology, BTT is intended to serve, these examples provide us with concrete tests of BTT's success. The ability of BTT to cohere with the extensional explanations of the theories from which our sample explanations are drawn (especially folk psychology and cognitive psychology) is one test of BTT's success as a theory of extensional content. This criterion of adequacy is of great importance, for the project undertaken here is a straightforwardly naturalistic project; the

Best Test Theory is based on the assumption that the natural sciences provide an effective way (perhaps the most effective way) to gain knowledge. This naturalistic outlook also underlies BTT's assumption of the (at least provisionally) materialist position taken by the natural sciences.

B. Caveats

1. First caveat

In what follows, I do not argue that there are such things as contentful states, and I do not address the various skeptical arguments which call into question the existence of contentful states. (For examples of such skeptical arguments, see Stich 1983 and Churchland 1981.) These arguments surely deserve to be addressed, but to address them thoroughly would require a book-length project different in nature from the present one. Instead, I assume that contentful states exist and then attempt to develop a theory of extensional content for the natural kind terms in LOT.

2. Second caveat

In discussing the theoretical role of content, I have talked both about the content of mental states and the content of terms in LOT. However, BTT applies only to terms, not sentences, in LOT. For BTT to be of much interest, it would seem that some of the mental states invoked in extensional psychological explanations would have to have their content determined at least partly by the extensions of the states' constituent terms. However, BTT makes no claims regarding how the content of constituent elements of mental states combine to yield complete thoughts with determinate contents. Furthermore, BTT makes no commitments to Russellian propositions, Lewisian possible worlds, or any other entities as the semantic values of complete thoughts. Ultimately, psychology would seem to require a theory of content which explains how the meanings of complete thoughts are determined. Psychology faces a somewhat different set of problems than philosophy of language, though, and it may be that a theory of content for complete thoughts will not be strictly analogous to any of the most prominent philosophical theories of meaning for complete sentences in a public language.

3. Third caveat

When characterizing BTT as a theory of intentional content[v], we must be careful not to cast

BTT's net too broadly. It is not intended that BTT provide necessary conditions for reference or aboutness, generally speaking. The Best Test Theory is only intended to contribute to the definition of a notion of content which can play a fruitful role in a science of human psychology. We should be pleased if BTT also proves useful as part of a theory of content which would help explain the behavior of some non-human species. However, we must recognize the possibility that there may not exist a single, scientifically useful, species-general concept of referential content.[vi]

4. Fourth caveat

There are other uses for which BTT is not explicitly intended, but for which it might prove useful. For example, a theory of mental content might provide the theoretical basis for a theory of linguistic content. Also, if a theory of intentional content is naturalistic, it may shed light on traditional puzzles regarding how a physical mind/brain could ever have thoughts **about** other things. The Best Test Theory may, in the end, be useful in these respects. But it is not necessary that there be a single concept of content which satisfies all of the purposes for which one might reasonably want a theory of content.[vii] Ultimately, BTT's success should be measured according to its potential for providing a description of mental content which will be useful in explaining human behavior.

C. The Best Test Theory of Extension

The Best Test Theory's basic principle says that a term in a person's language of thought refers to the members of the natural kind for which the concept associated with the term provides the best test.[viii] In order to give the reader some idea of what this principle says, I begin this section by discussing what it is for a concept to provide the best test for a kind. This is followed by discussions of concepts, the language of thought, and natural kinds, in that order.

The primary function of concepts is to effect the efficient categorization of the items around us. In order to react to the presence of new items in our environment effectively, and in order to reason effectively in planning for such encounters in the future, we must be able to classify the items we encounter under general headings, headings to which we may attach bundles of relevant knowledge. One way of conceiving of the process of categorizing new items is to think of it as the application of a series of tests. Each concept is a test which may yield a positive or a negative result on any number of different trials.

The Best Test Theory assumes that if we test a particular individual to see whether a particular concept applies to that individual, and the result is positive, this causes the tokening of the associated term in the subject's LOT. So for example, if a person applies her tiger concept to the animal in front of her, and the trial yields a 'yes' output, she will token her LOT term 'tiger'. Whether her LOT term which corresponds to what I've called her 'tiger concept' really refers to tigers depends on whether this concept is more reliable as a test for tigers than it is as a test for any other natural kind.[ix] In general, a given natural kind concept may yield positive results on some occasions for a number of different kinds. But for most natural kind terms, the associated concept will provide a test which is a far more reliable test for one natural kind than for any of the others.

How do we find out which kind is the kind for which a particular term provides the best (or most reliable) test? In order to make precise the idea of being the best test for a natural kind, BTT employs the idea of a success rate. At any given time, each natural kind has a success rate relative to each term in a subject's LOT. The success rate of natural kind K relative to subject S's LOT term t is expressed by a mathematical function, the success rate function. This function yields an output ranging from 0 to 1 for each ordered quadruple of arguments consisting of a natural kind, a human subject, a natural kind term in that subject's LOT, and a specific time. The success rate of K at time m, relative to term t in S's LOT, is given by dividing the number of times that members of K have caused the tokening of t in S's LOT by the number of time members of K have caused the tokening of any term at all in S's LOT.[x] It will be convenient to talk about the value of the success rate function for a specific member of its domain as a percentage of success. So if the success rate of tigers relative to S's term 'tiger' is 0.93, I will sometimes say that, relative to S's term 'tiger', the success rate of tigers is 93%. Lastly, in order to identify the extension of a particular natural kind term, you must find the natural kind which has the highest success rate relative to that term.

It is key that BTT assigns extensions based on the comparison of success rates. BTT does not tell us to choose a term in S's LOT and then figure out which type of things have most often caused S to token that term. Instead, we find out which natural kind has the highest success rate, for subject S, relative to the term in question. This can be found out only when we first calculate, for S, the success rate of each natural kind relative to the term in question. To illustrate the type of interpretive mistake I am warning against, consider the following case. Imagine that S has seen fourteen horses and has identified all fourteen of them as horses. Imagine also that S has seen 1000 cows and has mistaken forty of them for horses. We should not be tempted

to think that BTT assigns cows as the reference of 'horse' just because, in this example, there are more cow-caused tokenings of 'horse', forty, than there are horse-caused tokenings of 'horse', fourteen. The Best Test Theory does not tell us to simply compare the number of cow-caused tokens of 'horse', forty, to the number of horse-caused tokens of 'horse', fourteen. Instead we are supposed to calculate the success rate of both natural kinds relative to S's term 'horse' (as well as the success rates of any other natural kinds that have caused the tokening of 'horse') and compare those success rates. The success rate of horses relative to 'horse' = 14/14 = 1. In contrast, the success rate of cows relative to 'horse' = 40/1000 = 0.04. Thus, regardless of the fact that cows have caused substantially more of S's tokenings of 'horse' than horses have, horses still have a much higher success rate than cows relative to S's LOT term 'horse'.

Throughout the dissertation, I talk of concepts in ways which may seem strange to some readers. Traditionally, concepts have been viewed as abstract entities, existing independently of people, which embody definitional essences of the kinds of which they are concepts. According to this view, often called the 'classical view' (Smith and Medin 1981, chapter 3, Markman 1989, pp. 39-42), people mentally grasp concepts. By doing so, people understand the essence of the items to which the concept applies, and also understand the meaning of the public language term which expresses the concept.

Influenced by the classical approach, philosophical theories of concepts often go to some lengths to address concerns about the metaphysical nature of concepts (cf. Peacocke 1992, Rey 1983). While concepts may or may not exist as abstract entities fitting the classical description of them, I use 'concept' in a way which deviates from standard philosophical usage (without, however, intending to imply that there is anything deficient about the standard usage). I use 'concept' as a catch-all term for all of the tests, procedures, recognition mechanisms, etc., which causally mediate the tokening of a specific term of LOT in an individual. Such concepts are not to be taken as abstract entities which the subject grasps; BTT in no way rests on claims about people grasping concepts. The concepts referred to by BTT should be understood as collections of concrete entities that exist in the mind/brains of individual subjects. These collections of recognition mechanisms, etc., may correspond to abstract entities, i.e., sets of recognition mechanisms, etc., but BTT does not assume either that people consciously grasp these concepts or that these concepts provide definitions of the kinds to which the terms associated with the concepts refer.[xi]

I emphasize that the notion of concepts employed here involves a certain amount of stipulation. This

is because I am not concerned with explication or conceptual analysis. I am not trying to elaborate on or analyze the pretheoretical notion of a concept. Instead, I am trying to define a working notion of a concept which can serve as a tool in the exposition of BTT. As usual, the proof is in the pudding, and the reader will have to judge the value of this notion of a concept in accordance with the relative successes of BTT. Readers who object to my use of 'concept' can imagine a '*' following each token of 'concept' printed herein ('*' being a common way to mark a non-standard usage of a familiar term.)

D. The Best Test Theory, the Language of Thought, and Natural Kinds

In its details, BTT should be seen as the development of a specific research program. The Best Test Theory assumes that there is an LOT and that there are such things as natural kinds, and then offers an explanation of how the members of a given natural kind can serve as the extension of a particular natural kind term in LOT. The assumption that there exist terms in a language of thought is substantive. The existence of a language of thought, consisting of syntactic categories, formation rules, expressions with semantic interpretations, etc., has been the focus of quite a bit of controversy (cf. Dennett 1978, Blackburn 1984, pp. 51-57, Sterelny 1990, 2.2, 2.3, and chpt. 8, Davies 1991, and Rey 1995). However, by adding the appropriate qualifications, I attempt to sidestep much of this controversy.

The Best Test Theory is consistent with, and would naturally serve, the LOT hypothesis in its strongest form (Fodor and Pylyshyn 1988, pp. 12-14). However, BTT depends directly on the truth of only a very small part of the full-blown LOT hypothesis. The Best Test Theory only assumes that there is a discrete class of terms in LOT which have natural kinds as their referents. It is convenient to talk about these terms as analogues to the common nouns or one-place predicates of a public language. However, it is not necessary that we define the types of terms to which BTT applies in the same way that we would define the syntactic categories Common Noun or One-Place Predicate.

In chapter IV, I argue that we can identify the natural kind terms of LOT without adverting to their membership in a syntactic category of LOT terms analogous to a syntactic category of public language terms. Assuming this argument is sound, BTT can remain agnostic regarding the claim that the cognitive processes in which the terms of LOT are said to play a role are syntactic in nature.[xii] The Best Test Theory may then be seen as consistent with connectionist claims that mental states do not have the constituent structure attributed to them by computational theories of cognition. (On the lack of constituent structure, see

Van Gelder 1990; on computational theories of cognition, see Haugeland 1981 and Pylyshyn 1984.) The Best Test Theory is also consistent with claims that reasoning does not proceed formally (as described, for example, by the inference rules of the predicate calculus), but instead by the use of specific examples, cognitive models, or cognitive reference points (cf. Lakoff 1987, Johnson-Laird 1983).

Given that BTT is offered as a naturalistic theory of content, it would seem uncontroversial that BTT assumes the existence of natural kinds. That natural kinds exist is a fundamental assumption of the natural sciences.[xiii] However, there is a surprising consequence which results from the assumption that natural kinds exist. The assumption that natural kinds exist allows BTT to solve (almost by stipulation, it would seem) the central problem faced by causal theories of reference for LOT terms.

E. Misrepresentation and the Disjunction Problem

Much of the literature on causal, informational, and covariational theories of content is concerned with solving the problem of misrepresentation (cf. Dretske 1981, 1986, Cummins 1989a, Fodor 1987, Fodor 1990b). Consider a crude version of a causal theory of reference for natural kind terms in LOT (Fodor 1987, p. 99), a theory based solely on the intuition that a natural kind term t of LOT refers to the causes of tokenings of t. A crude causal theory might claim that the extension of t in subject S's LOT consists of the members of any natural kind at least one whose members has caused (or is currently causing) the tokening of t in S. It seems that there can be no misrepresentation according to such a theory. Take S's LOT term 'horse'. Assume that most of the horses S encounters cause S to token 'horse'. According to our crude causal theory, horses are thereby in the extension of S's LOT term 'horse'. However, in addition to horses, our crude causal theory implies that the members of any other natural kind whose members have caused S to token 'horse' are in the extension of 'horse', even if such items are cows, fruit bats, or iguanas. Assume we want to exclude cows from the extension of S's LOT term 'horse', the crude causal theory will not allow us to do so, so long as at least one cow has caused S to token 'horse'. We cannot say that when S tokens 'horse' in response to a cow that S misrepresents the cow as a horse. S's tokening of 'horse' in response to a cow automatically places cows in the extension of S's LOT term 'horse', according to the crude causal theory. The crude theory thus assigns to our natural kind terms of LOT extensions with disjunctive structure in the sense that in order to be in the extension of t, an item need only satisfy one disjunct on the list of all of the natural kinds whose members have ever caused the tokening of t. Jerry Fodor calls this problem the 'disjunction

problem' (Fodor 1987, p. 102, the exposition here closely follows Fodor 1987), and rightly assumes that any causal, informational, or correlation-based theory of content will have to solve this problem.

The Best Test Theory solves the disjunction problem (DP, hereafter) by claiming, more or less, that natural kinds are the only appropriate reference classes for natural kind terms in LOT.[xiv] Imagine that a subject S sometimes tokens 'horse' in response to cows (say, as in Fodor's example, the cows appear on dark nights [Fodor 1990b, p. 121]). If you want to know what the extension of 'horse' is, BTT directs you to sift through the set of natural kinds and find the one with the highest success rate relative to 'horse'. The key assumption here is that success rates are only calculated for natural kinds, not disjunctive classes like horses or cows on dark nights (i.e., cows observed under the specific circumstances which cause S to mistake them for horses). When we compare the success rates of the various natural kinds, horse has a higher success rate than any other natural kind (including cow) relative to S's LOT term 'horse'. The Best Test Theory thus solves the DP.

It is one of the BTT's virtues that it solves the DP so easily. However, the ease with which BTT solves the problem may appear to result from ad hoc stipulation. Such concern is ill-founded. I argue in chapters V that we are justified in assuming that for the natural kind terms in LOT, success rates should be calculated for natural kinds only.[xv] The Best Test theory's specific role, then, is to tell us which, among all of the natural kinds whose members cause the tokening of a given natural kind term in LOT, is the natural kind whose members constitute the extension of that term.

F. Summary of Dissertation by Chapter

Chapter II begins with the proposal of criteria which any naturalistic theory of extension should satisfy, and progresses as a critical review of the two leading naturalistic theories of extension, those proposed by Fred Dretske and Jerry Fodor. Both theories are similar to BTT in their reliance on causal connections between natural kinds and terms in LOT to define the extensional content of LOT terms. However, both Dretske's and Fodor's theories have serious drawbacks. I reject Dretske's theory for a variety of reasons, one being that it does not define a notion of content according to which statements about content can properly enter into counterfactual-supporting generalizations of psychology. And while it is very powerful in some respects, Fodor's theory, as a theory of content for humans' LOT terms, suffers from the reliance on murky intuitions about what would happen if the nomic structure of the universe were altered.

Chapter III presents BTT as an alternative to the theories reviewed in chapter II. The Best Test Theory consists of a set of content-determining principles, each of which applies to different types of cases. The first principle, BT1, expresses BTT's basic idea, outlined above. Much of chapter III is spent explaining the details of BT1's application. The remainder of the chapter articulates the modified principles BT2 and BT3 which explain how people can use natural kind terms of LOT the extensions of which have been previously fixed to determine the extension of new LOT terms (or to modify the content of terms already in use).

Chapter IV consists of a review of empirical evidence in support of the claim that the representation of natural kinds is psychologically real in human cognition from a very early age. The empirical evidence seems to show that there is a certain class of terms of LOT (intuitively thought of as natural kind terms) whose members are systematically treated in a special way. On the grounds of this evidence, I conclude that the natural kind terms of LOT form a distinct, well-defined class of LOT terms, and thus that we are justified in treating them as a distinct class of LOT terms in developing a theory of extension for LOT terms in general.

The developmental data reviewed in the earlier sections of chapter IV raise questions about the role of implicit intentions in fixing the extension of LOT terms. The second half of chapter IV is an exploration of the role implicit intentions play in the child's identification of LOT terms as natural kind terms, in hopes of also understanding the role such intentions play in fixing the extension of natural kind terms in the child's LOT.

Having assumed from the outset that natural kinds exist, it remains to be proven that they are the right sort of things to which the relevant terms of LOT refer. In particular, if BTT is to provide an adequate solution to the problem of misrepresentation, we must show that for the natural kind terms in LOT, psychology's explanatory and predictive goals are better served by assuming that natural kinds are the appropriate reference classes rather than the disjunctive kinds defined by the kind of crude causal theory of extension considered above. Chapter V presents an argument to this end. The argument is based on the claim that if it is important to psychology to solve the DP at all, then BTT offers the right kind of solution. More explicitly, if psychology gains any advantage by having a theory of extension according to which extensions are not disjunctive in the way that gives rise to the DP, then such advantage is argument enough for dismissing disjunctive extensions (as BTT does) from the list of candidate extensions for the relevant classes of terms. To complete the modus ponens, I argue that psychological theories which assume that the

relevant terms in LOT refer to natural kinds are empirically superior to theories which accept disjunctive kinds of the sort which cause the DP as candidate reference classes for LOT natural kind terms.

Chapter VI consists of a review of some of the most potent criticisms that might be made of BTT, and gives responses to these criticisms. Among the criticisms are the following: (1) that BTT leads to phenomenalism, because, for example, the natural kinds for which many concepts provide the best test are natural kinds of proximal, not distal, stimuli, (2) that the kinds in nature are not discrete and therefore the relationships between kinds and terms which are required by BTT cannot exist, and (3) that BTT implies unacceptable, arbitrary changes in extension. It is hoped that the responses included in chapter VI will satisfy readers and further clarify the scope and the promise of the best test approach.

There are two versions of BTT, a weaker and a stronger. In its weaker form, the one which is developed in the present work, BTT consists of a set of principles which explain how the extensions of natural kind terms in LOT are fixed. In its stronger form, BTT assumes that the reference of all LOT terms rests fundamentally on reference to natural kinds (with the likely exception of terms from LOT's logical vocabulary, e.g., 'and' or '+' [Fodor 1990b, pp. 110-111]). As a comprehensive theory, BTT consists of an open-ended set of principles, those which explain how reference to natural kinds is achieved together with any principles necessary to explain the how successful reference to items or groups other than the collective memberships of natural kinds depends ultimately on successful reference to natural kinds. Throughout the dissertation, I am primarily concerned with presenting and defending the constituent principles of BTT as they apply to natural kind terms of LOT. Little time is spent discussing or arguing for the broader, more comprehensive version of BTT. However, at certain points, I briefly explore the ways in which BTT might be extended, and the reader is encouraged to consider how the core ideas of BTT might be applied to other types of LOT terms.

Notes to Chapter I

---

[i] The term 'object' is used here in a way similar to the way the term is used in grammar when we talk about the object of a transitive verb. Thus, the object of a mental state can be an individual, concrete item, an abstract entity, a set, a collection of individual objects, or anything else that a thought is about. It may seem that cases where thoughts whose objects are non-existent are an exception. I will assume that such thoughts have an object of some sort, the nature of the object in such cases being open to debate. For relevant discussions of the nature of the objects of such thoughts, see Geach 1967, Dennett 1982, and Edelberg 1986.

[ii] This sample explanation is not, strictly speaking, an explanation of why people behave in the way they do. Instead it is an explanation of why many people have the beliefs that they have about social institutions, in particular, beliefs that various social institutions are just. However, as is often the case in psychology, in order to explain the relevant behavior, we must first explain why the subjects have certain mental states which we take to be the cause of the behavior. The existence of certain mental states then becomes the short-term explanandum, with the long range goal being the explanation of the relevant behavior. The explanation of how the relevant mental states get their content is thus part the explanation of human behavior. In the case at hand, the fact that a subject believes that traditional marriage is a just institution can be explained by invoking the existence of other mental states with content (e.g., perceptual states which result from the observation of specific sexist imagery). Once we have explained why the subject believes that the traditional institution of marriage is just (and done so by reference to other contentful states), we can then invoke the content of the subject's belief that marriage is just to explain the subject's actual behavior. An example of such behavior would be the subject's criticizing friends for their non-traditional living arrangements even when the subject claims to oppose injustice.

[iii] Jean Piaget is well-known for having run experiments of this type. For recent critical discussions of such experiments, see Bower 1989, chpt. 5, and Diamond 1991.

[iv] The former type of explanation seems more basic than the latter. Before we can individuate an LOT term according to the type of thing to which the term refers, that term's extension must first be fixed. On the other hand, some theories of content which, generally speaking, taxonomize by extension do so even for terms which have no extension, e.g., 'unicorn' (cf. Fodor 1990b, pp. 100-101), which may make one doubt that a theory of extension truly is more basic than a theory which offers a generally extensional taxonomy of LOT terms and states. It seems that a theory of extension really is the more basic theory in that in order to assign the extensional content unicorn to 'unicorn', we must first have a theory that tells us that 'unicorn' would refer to unicorns were there any around. (Fodor remarks that his theory of intentional content tells us that "people would apply 'unicorn' to unicorns if there were any." [Fodor 1990b, p. 116].)

There are (at least) two ways for a theory of extension to accomplish this. One way is to construct a theory of extension so that it assigns extensions to terms across possible worlds (so that we know what 'unicorn"s extension is in worlds where it has an extension). (This first approach amounts to giving a theory of intensions, as intensions are often thought of as functions from possible worlds to extensions in those worlds.) The other way is to claim that all terms like 'unicorn' are complex terms whose parts have actual extensions. We would then taxonomize terms like 'unicorn' (and the states constructed from these terms) by referring to the extensions of their simple parts (which parts have actual extensions.) I prefer the latter of the two strategies, though this approach has implications which some find unacceptable (Baker 1991, p. 21, and

Boghossian 1991, p. 77). More on this in chapter VI.

[v] I have, at this point, returned to a general discussion of intentional (with a 't') psychology as opposed to intensional (with an 's') psychology, also discussed above. The distinction between intentional and non-intentional psychologies is, roughly, the distinction between psychologies which assume that mental states take objects and those which claim that mental states do not have objects. Extensional psychologies and explanations are normally intentional, for the mental states of an extensional psychology take the extensions of certain state constituents as objects. Intensional psychological theories are also sometimes characterized as intentional in that such theories sometimes assume that the mental states of which they speak are directed at, or are about, objects (even if these objects are themselves only abstract intensions). It is possible that an intensional psychology would not be intentional if the intensions of the relevant states and terms did not represent in any way. This may be the most accurate way to describe Stich's syntactic psychology (Stich 1983).

[vi] This is not to say that we should develop human psychology entirely independently of animal psychology. To the contrary, the various fields of study which fall under these headings have much to offer one another in the way of provocative ideas, explanatory strategies, etc. The point is simply that the development of human psychology, and BTT, should not be too heavily constrained by (as opposed to being informed by or inspired by) data collected by researchers in ethology and related fields. We should leave open the possibility that there are different intentional psychologies for different species. By talking about different psychologies, I mean psychologies which consist of different theoretical laws and which employ concepts which are either not present across theories or which are defined differently relative to different theories (the result being, for example, that we get two different theories of extension which apply in two different theoretical domains).

[vii] Robert Cummins makes a similar point in claiming that a theory of intentional content should be theory-relative, i.e., that we should only construct and judge a theory of content as it serves some particular psychological theory or theoretical approach (Cummins 1989b, pp. 12-14).

[viii] According to BTT, the (non-empty) extension of a natural kind term of LOT is the set of members of a natural kind. However, sometimes it is convenient to be able to speak simply of terms referring to natural kinds. Where context should resolve any ambiguity, I omit 'the set of members of'. It should also be noted that when I talk of the set of members of a natural kind, 'set' is used generically to mean 'collection'. This sense of 'set' is in contrast to its formal, mathematical sense in which talk about sets is talk about abstract objects.

[ix] On pains of circularity, we can not individuate LOT terms according to their extensions before those extensions have been fixed. Thus, a more neutral description of 'tiger' would be something bland like 'T'. By calling the LOT term of interest 'tiger', I only mean to let the reader know that the term of interest here is the LOT term which we would pretheoretically identify as referring to tigers.

[x] When no member of K has ever caused S to token any term at all, the rule as stated says that the success rate of K relative to any term in S's LOT = 0/0. However, because division by zero is not defined in arithmetic, we must treat these cases differently than we treat all other cases. In cases where a success rate = 0/0, BTT stipulates that the success rate is 0.

[xi] See Mind and Language, Vol. 4, for a useful forum discussion of concepts, in particular, of the

distinction between concepts as metaphysical entities and concepts as epistemological entities. Chapter III contains a more detailed examination of the nature of BTT's concepts together with a more detailed comparison of BTT's concepts to those discussed in the psychological literature.

[xii] For those to whom the argument of chapter IV does not seem satisfying, it might seem reasonable to embrace the LOT hypothesis entirely and interpret BTT accordingly. The present caveat is intended largely to protect BTT from the charge that it is inconsistent with connectionist results. A different way to avoid conflict between BTT and connectionist results may be to adopt a common interpretation of (at least some) connectionist research as an illumination of how a symbolic language is used by human brains (cf. Sterelny 1990, pp. 175-177, Clark 1989, chapter 7). When connectionist results are understood in this way (as part of cognitive theory at the implementational level), there is no conflict between connectionism and the full-blown LOT hypothesis.

[xiii] There is, however, controversy over the exact nature of natural kinds (cf. Lakoff 1987, Dupre 1981, Maudlin 1986). In so far as the controversy over natural kinds forms the basis of an objection to BTT, it will be necessary to further explore the controversy. Consequently, I discuss natural kinds in more detail in chapter VI when considering objections to BTT.

[xiv] This oversimplifies matters. There are cases of genuine disjunctive reference, where the extension of a term is made up of the members of more than one natural kind. Such cases are discussed in detail in chapter VI.

[xv] Fodor makes a similar assumption in claiming that only nomic connections are to be considered for the purposes of applying his theory of content (cf. Fodor 1990b, pp. 100-103, 121, Loewer and Rey, p. 257). However, Fodor allows a wide range of connections to count as nomic. For example, according to Fodor, the relationship between being a cow on a dark night and being a cause of 'horse' "is nomic on the operative assumption that cows on dark nights qua cows on dark nights are sometimes mistaken for horses." (Fodor 1990b, p. 121) Perhaps what Fodor has in mind here is that cows, in virtue of being cows, possess some properties (e.g., being big, being four-legged) which are such that they sometimes (on the occasional dark night) cause the tokening of 'horse' in humans. Calling this connection nomic (i.e., lawlike) seems to be stretching things a bit. If this is a lawlike relation, why do cows on dark nights only sometimes (in fact, very rarely) cause the tokening of 'horse'?

II. CAUSAL THEORIES OF CONTENT

A. Evaluative Criteria

This chapter consists of a critical review of Fred Dretske's and Jerry Fodor's theories of intentional content. Although Dretske and Fodor are not the only philosophers who have put forth naturalistic theories of intentional content for LOT terms, their theories provide the leading, most commonly cited examples of causal/covariational theories. Before turning to the details of Dretske's and Fodor's theories, it will be useful to have some idea of the standards by which theories of intentional content for LOT terms are to be judged. The following are three criteria which any naturalistic theory of intentional content for LOT terms should satisfy:

Criterion #1- The theory must explain how it is possible for people to misrepresent their environment.

Criterion #2- The theory must be consistent with, and better yet, provide some coherent explanation of, the behavior of humans as detailed by the relevant research sciences, and

Criterion #3- The theory must, roughly speaking, measure up to the standards for being a scientific theory.

Criterion #1 claims that any naturalistic theory of content must solve the problem of misrepresentation. In fact, the problem of misrepresentation is a problem that any theory of content, naturalistic or otherwise, must solve. However, the apparent fact that people sometimes misrepresent their environment poses a particular kind of problem for causal or covariation-based theories of content. The idea behind a causal theory of intentional content is that the extension of an LOT term is whatever causes the subject to token that term. But if the content of a term is defined as that which causes the tokening of the term, then it's difficult to see how one could ever have

mistaken thoughts (i.e., false beliefs). Imagine that one sees a cow and thinks (what we would normally characterize as) the mistaken thought, 'There's a horse'. The problem for a simple causal theory of content is that, since it was a cow which caused you to think 'horse', then 'horse' refers to cows as well as, perhaps, to horses. Consequently, what we would normally take to be a false belief is classified as true. A causal theory

of content must somehow explain how a type of mental structure can acquire a fixed meaning and thus be misapplied or 'mistokened' on a specific occasion. In the terminology of chapter I, Criterion #1 says that an acceptable theory must solve the DP.

If a theory of content for LOT terms is to be truly naturalistic, the theory must comport with the relevant research data from the social and brain sciences. Criterion #2 codifies this requirement. Where it does not conflict with the scientific data, some common sense or intuitive data should be taken into account. After all, common sense psychological theorizing (often called 'folk psychology') provides the basic theoretical framework for much of cognitive and social psychology. It is likely that some intuitions regarding what a given LOT term should refer to will not be shared by all. Accordingly, I do not weight such judgements too heavily. However, in cases where an intuitive judgement would seem to be widely shared and strongly felt, these judgements are accorded appreciable importance.

Of the three criteria, the third is the most difficult to make precise. The standards for good scientific theory construction are not widely agreed upon, and many skeptics object to the whole idea of a clearly demarcated realm of scientific knowledge. However, there are certain practices that are widespread in the scientific community, and it is expected that such practices will be adhered to as much as they can be in a philosophical development of a naturalistic theory of extensional content for LOT terms. This demand is justified by the fact that I am attempting to develop a theory of content which will be of theoretical use to cognitive psychology and, it is hoped, to a scientific psychology more broadly construed.

B. Dretske

1. Indication and assigned functions

According to Dretske, the content of a representation is fundamentally based on the existence of the indication relation. Dretske explains the indication relation in the following way:

> The power of signs to mean or indicate something derives from the way they are related to what they indicate or mean. The red spots all over Tommy's face mean that he has the measles, not simply because he has the measles, but because people without the measles don't have spots of that kind. In most cases the underlying relations are causal or lawful in character. There is, then, a lawful dependency between the indicator and the indicated.... Sometimes, however, the dependency between a natural sign and its meaning derives, at least in part, from other sources....[but] In order for one thing to indicate something about another, the dependencies must be genuine. There must actually be some condition, lawful or otherwise, that explains the persistence of the correlation. (Dretske 1988, pp. 56-57)[i]

Stated generally, X indicates Y because, were it not for Y, X would not be the case. The indication relation does not allow for misindication, or mistakes, as these are normally understood (Dretske 1988, pp. 55-56, 66).

Although indication does not, by itself, constitute representation, indication is a basic building block of representation. In order for a structure to have intentional, or representational content, the structure must have a function within a system. "Once C is recruited as a cause of M--and recruited as a cause of M because of what it indicates about F--C acquires, thereby, the function of indicating F." (Dretske 1988, p. 84) The visual appearance of an eagle indicates the presence of an eagle. But without having a particular function in the perceiving animal's cognitive system, the eagle appearance indicates other things besides just that an eagle is present, e.g., that a bird is present or that the universe is non-empty. The internal perceptual structure (or some LOT term associated with the perceptual structure) caused by the eagle can only come to mean eagle, in a full-blooded, representational sense, when it acquires the function of indicating the presence of an eagle (so that, for example, the animal who sees the eagle can take evasive action, thus avoiding death). "Only by using an indicator in the production of movements whose successful outcome depends on what is being indicated can this functional indeterminacy be overcome." (Dretske 1988, p. 70) For Dretske, indicator functions are indeterminate with respect to their content because a given type of indicating mental structure can indicate, or have its tokenings caused by, numerous different types of things. For a mental structure to have determinate representational content, it must have the function, within the cognitive system, of controlling behavior(s); and the mental structure must have this function because of what it indicates. Dretske summarizes his account of how content emerges in the following way:

> [D]uring the normal development of an organism, certain internal structures acquire control over peripheral movements of the systems of which they are a part. Furthermore, the explanation, or part of the explanation, for this assumption of control duties in not (as in the case of artifacts) what anyone thinks these structures mean or indicate, but what, in fact, they do mean or indicate about the external circumstances in which these movements occur and on which their success depends. In the process of acquiring control over peripheral movements (in virtue of what they indicate), such structures acquire an indicator function and, hence, the capacity for misrepresenting how things stand. (Dretske 1988, p. 88)

2. Objections to Dretske's view

a. The anthropology test and behaviorism revisited

On Dretske's view if a mental representation has never actively controlled behavior, the representation has no determinate content (and in fact, is not truly a representation at all). In light of common introspectively gathered data, Dretske's view seems implausible in the same way that some of the behaviorist claims of the 1940's and 1950's did. It is even more implausible when one considers that the establishment of content is dependent on successful behavior. However, putting questions of success aside for the moment, consider the following example.

You are an anthropology student who's been told to memorize all of the definitions in chapter 4 of your textbook for an upcoming exam. While reading chapter 4, you may acquire a number of highly detailed concepts (e.g., family relation concepts of the Southern African Kunda) just by reading definitions of the concepts in your textbook. When exam day comes, your instructor tests you on some of the definitions from chapter 4, but not on others. Much of what you learned when you were studying is likely to fade from your memory before long, whether you were tested on it or not. On the other hand, some of the definitions may stick in your mind, even some of those on which you were not tested. Of those definitions on which you were not tested, but which you remember, some may never control your behavior in any way. They may just sit there in your brain. In this way, it seems quite possible for a person to acquire highly detailed concepts (with determinate content) without those concepts ever controlling the person's behavior in any way. Dretske's theory denies this possibility, and thus seems deficient. (Criterion #2 is not satisfied.)

Dretske might respond to the preceding example by pointing out the complex nature of the concepts that were acquired. Dretske might claim that the only reason we feel confident that the concepts in question have determinate content is because their conceptual constituents have determinate content; those constituents, Dretske might claim, are in control of actual behavior, and thus, on his theory, have determinate content.

In response to Dretske, we might imagine an example analogous to the one above, using primitive, learned concepts instead of complex ones. Why couldn't we learn new simple concepts but never employ them in the control of behavior? There is a genuine risk, however, that a debate over the determinacy of content of simple, learned concepts would be fruitless match of intuitions. I leave matters here, then, noting the strength of the intuition that sometimes we learn things, and keep what we learn to ourselves.

In the passages quoted above, Dretske clearly rests content acquisition on the connection between an indicator mechanism and its control of actual (even successful) behavior. On the other hand, Dretske is

sometimes (slightly) less committal. "C [a token LOT term] gets the meaning or content F because past Cs (past tokens of the same type), by carrying the information that F, gave C its functional role in the regulation and the direction of output." (McLaughlin 1991, p. 215) Granted this passage ties content directly to behavior (or output), but it does so in a way that leaves room for a counterfactual-based interpretation of Dretske's theory. Dretske cites the change in the functional role of an indicating structure as giving rise to content, and functional roles are often defined according to their counterfactual powers not their actual causal histories. Perhaps when Dretske says that an indicator has been given a job to do in the explanation of behavior, he means to include explanations of what the subject would do, in addition to explanations of the subject's actual behavior. For example, when a rodent sees its first eagle, perhaps the eagle indicator can gain content before it ever controls actual evasive behavior, in virtue of the change in the indicator's functional role which is not yet manifest. The relationship between the rodent's eagle indicator and eagles may still be explanatory in the sense that it explains why the rodent would run away were it to see another eagle.[ii]

The suggested reading of Dretske is quite an interpretive stretch. Dretske talks about actual behavior (even successful behavior) throughout Dretske 1988 and throughout his replies to critics in McLaughlin 1991. Additionally, operant learning is central to Dretske's exposition of how content is fixed WRONG! THIS APPLIES TO BELIEFS, NOT TO CONTENT (Dretske 1988, pp. 95-107), and as Dretske is well aware, operant learning requires that the learner exhibit actual behavior that is then reinforced. Thus, despite the fact that the suggestion on the table solves Dretske's quasi-behaviorist problems, the counterfactual-based version of Dretske's theory should be rejected as an interpretation of Dretske. In what follows, then, I ignore the counterfactual-based interpretation of Dretske's theory and introduce a further difficulty for the view that a connection between the indicator mechanism and actual, successful behavior is necessary for the creation of determinate content.

b. Occam's razor, cutting away types, and cutting away tokens

Dretske invokes theoretical entities or forces (i.e., contents) to explain actual behaviors. However, Dretske seems to violate standard scientific practices, and the metaphysical assumptions on which those practices are based (and thus, violates Criterion #3), by claiming that the theoretical constructs of his theory fail to exist if they have never played a role in explaining an actual token instance of behavior. At first blush, Dretske's refusal to countenance contents which do not play a role in a

specific explanation may appear to be a reasonable application of Occam's razor. I propose to consider Dretske's view as such, for the moment, but argue that such an application of Occam's razor is misguided.

Occam's razor is often taken to be a guiding principle in scientific theory construction. Roughly stated, the principle of Occam's razor says that if an entity or force is not necessary to explain the phenomena in question (i.e., those which the theory in question is intended to explain), then we should not include such entities or forces in the relevant scientific theory. What is important to note is that Occam's razor has traditionally been applied to types of entities or forces and types of phenomena, and that there is very good metaphysical reason for doing so.

Consider the example of electrons. Physicists and chemists assert that electrons exist as components of nearly all atoms. Furthermore, physicists and chemists assert this even though there are cases where token electrons are not invoked to explain an observed, token phenomenon. We do not use Occam's razor to cut away the token electrons which are not actively playing a role in explanation. Instead we accept the existence of electrons as a type of entity, because they are useful in explaining (and making predictions about) certain types of phenomena.

To simplify matters assume that the only reason physicists and chemists hypothesize the existence of electrons is because the presence of electrons explains the chemical reactions into which atoms of different elements enter. Consider the case of large quantities of atoms of a certain element existing in an homogeneous environment (e.g., at the center of a star.) These atoms do not interact with (and to our knowledge, may never have interacted with) any other chemical elements. In this case, no physicist or chemist would assume that the atoms existing in the homogeneous environment don't have electrons as components just because the atoms in question have not entered into any observable chemical reactions.

In this example, the changes that occur as the result of chemical reactions are the observable phenomena analogous to the behaviors which Dretske wishes to explain by invoking representational content. If he follows the lead of nuclear chemists, Dretske should claim that representational content, as a type of theoretical entity or force, exists independently of whether it's being invoked to explain a token event in need of explanation. In some cases, the only relevant difference between the situations in which Dretske says that content exists and those in which Dretske says that it doesn't exist is the lack of a token event that requires explanation. But if this is the only difference between the two types of situations, then good scientific methodology commits Dretske to the existence of content across the board. He cannot simply say that representational content fails to exist in those token cases where content is not of explanatory value, any

more than the chemists can say that the token electrons don't exist when they're not playing a role in the explanation of a token chemical reaction. (By doing so, Dretske gives a theory that fails to satisfy Criterion #3.)

On what grounds does the nuclear chemist refuse to apply Occam's razor to cut away the token electrons which do not play an active role in the explanation of token phenomena? Why isn't having a role to play in a token explanation the accepted test for the existence of theoretical entities? The problem with accepting such a test is that it leads to a metaphysical mystery of time and causality. The very entities whose causal agency is invoked to explain a phenomenon would have to come into existence at the very instant that they exerted their causal powers (and not a microsecond before). In this way, predicating the existence of a token theoretical entity entirely on its having caused a certain observable phenomenon seems to commit us to a radically different metaphysical picture than the one which normally underlies scientific theory construction. Normally, we assume that for an entity or force to have causal effects (resulting in observable phenomena which need to be explained), that entity or force must exist prior to the observable events which the existence of the entity or force is supposed to explain.

As noted above, Dretske ties the appearance of content not only to a structure's acquiring control over behavior, but to the structure's acquiring control over successful behavior. Metaphysically speaking, this would seem to make matters even worse for Dretske. The success or failure of a behavior is determined by events that occur later in time than the behavior itself. It takes quite a metaphysical imagination to think that whether a structure C has representational content at given time depends on what happens later on C's time line.[iii]

Concerns about time and causality justify restricting our application of Occam's razor to types of entities. Such concerns also demand that we give an independent characterization of theoretical entities which provides the grounds for asserting when token theoretical entities of a certain type exist, whether or not those token entities are being invoked to explain any token phenomenon. Thus, if Dretske is going to invoke a specific content F to explain my behavior B, and Dretske wants to avoid the causal mystery of something's coming into existence and exerting causal agency at the exact same time, then Dretske has to give an independent characterization of content F. In referring to content F when explaining B, Dretske seems committed to giving a characterization of F as a type of entity or force, which characterization is independent of F's control over B. Dretske has to say what it is for someone to possess a structure with

content F, whether or not F is invoked in the explanation of token behaviors of that person. However, if Dretske were to give such an independent characterization of F (or even to admit that one exists), then Dretske would be thereby admitting the possibility of a subject's possessing a structure with content F even though F has never controlled any of that subject's behaviors. This is a possibility that Dretske's theory denies.

c. Indicational indeterminacy

Dretske's earliest theory of content (Dretske 1981) attempts to solve the DP by distinguishing between a learning period and a post-learning period in the acquisition of a concept. Dretske claimed that in the learning period, there is a perfect correlation between the symbol learned and the members of the symbol's extension. This perfect correlation fixes the content of the symbol. After the learning period is over, the subject can go forth, with content determined, to apply or misapply the symbol, as the subject may. This approach to solving the DP has been widely discussed and criticized (cf., Fodor 1990b, pp. 41-42, 61-63, Cummins 1989b, pp. 67-69, and Sterelny 1990, pp. 121-123). In his more recent work, Dretske draws a distinction which plays a similar role to that of the earlier distinction between a learning period and a post-learning period. Unfortunately for Dretske, one of the strongest criticisms of the earlier distinction seems to apply in like fashion to his current theory.

According to the theory of Dretske 1988, there is a distinction between the situations where an indicator merely indicates from those situations where that indication has been utilized to control behavior, giving rise to determinate content. The indeterminacy of indicational content is resolved, according to Dretske, when the indicating device gains control over behavior, and does so because of a specific one of the indicating structure's indicational capacities. Thus, in response to some (but not all) of the criticisms made above, Dretske might claim that content does not need to be invoked to explain behavior in the early stage of content fixation. It is only after the indicator has been assigned a role in the control of behavior that there is any question of the related structure's content explaining behavior.

Regarding Dretske's earlier theory, Fodor complains that the learning period itself contains the indeterminacy which gives rise to the DP. Without endorsing Fodor's criticism of Dretske 1981, we can extend Fodor's criticism to Dretske 1988. This can be done because it would seem that what is indicated at the time at which an indicator is recruited to control behavior is itself indeterminate. Assume that the rodent's indicator C is triggered by an eagle, and C is thus recruited to control evasive action in the rodent. C

now refers to eagles, right? The problem is that C may well have had varying indicational capacities even at the time of the rodent's first perception of an eagle. Perhaps C would have equally been triggered by a high flying airplane, or a sea gull, and in those cases would also have been recruited to control evasive action. It would seem that anything that can cause C in the rodent after content is fixed would have caused C in the first place, i.e., before content was fixed. This implies that in the rodent's first experience with an eagle, C indicated eagle or high-flying airplane or sea gull, etc. And if C indicated the disjunction of all of these things when C was recruited to control behavior, the content that was fixed is disjunctive. The DP remains unsolved. (Thus Dretske's theory fails to satisfy Criterion #1.)

What Dretske needs is a way to isolate the indication of eagles as the indicator function of C in the rodent's first experience with an eagle. Dretske sometimes seems to want to appeal to future success here. If C gains control of evasive behavior, and evasive behavior keeps the rodent alive in the future by helping the rodent escape eagles (not sea gulls or high-flying airplanes), then C refers to eagles. End of story. However, appeals to future success to fix content were rejected above, for good reason. Dretske seems to need some other way to isolate eagle-indicating as the only relevant indicator function of C.

Perhaps Dretske can appeal to history. An eagle caused C on the occasion when C was first recruited to control the rodent's behavior, so that's what C refers to, Dretske might claim. The problem persists, however. Even the actual history is not enough to resolve the indicational indeterminacy. For on the actual occasion in question, C indicated not just the presence of an eagle, but also the presence of a big bird, the existence of something in the universe, the presence of an egg-laying animal, etc. It is difficult to see how Dretske can resolve this indeterminacy without deferring to future success. Given the difficulties for Dretske's theory reviewed here, I propose to leave Dretske's theory and go in search of less problematic theoretical ground.[iv]

C. Fodor

Fodor first attempts to give a theory of intentional content in Psychosemantics (Fodor 1987).[v] There Fodor begins by considering the simple idea that a term of LOT "'A' means A iff all and only A's cause 'A's," (Fodor 1987, p. 110). However, Fodor immediately points out that nobody's perfect at recognizing As and that 'A's are sometimes caused by things which are not As. Therefore, Fodor rightly concludes, there must be a way for 'A' to mean A in the absence of a perfect causal covariation between 'A' and As. Fodor

reacts by qualifying both the 'all' and the 'only' clauses of his content-determining principle. Fodor replaces the demand that all As cause 'A's with the claim that, as a necessary condition for 'A' to mean A, 'A's need only be reliably correlated with As. Furthermore, non-A caused 'A's don't change the meaning of 'A' (it can still refer to As) because of a phenomenon which Fodor calls 'asymmetric dependence'.

In Fodor's more recent work (1990b), Fodor has given up on the idea of a reliable correlation as providing a foundation of a theory of content, and has built his new theory of content entirely on the edifice of asymmetric dependence (AD, hereafter). There are at least two apparent reasons for this. Firstly, AD can well be taken to be an explanation of whatever reliable correlation exists between an LOT term and it reference class. Secondly, Fodor worries that there is not actually as reliable a correlation between As and the tokening of 'A's as one might expect. Fodor is concerned about the phenomenon of semantic robustness, whereby a term's tokening can be, and in the typical case, frequently is caused by items which are not in the term's extension. Horses may sometimes cause the tokening of 'horse'. However, for most people, things other than horses, e.g., thoughts of ranches or visual images of the local racetrack, cause the tokening of 'horse' as often as, if not more often than, actual horses. Thus, Fodor's focus on AD is well-motivated. What remains of the current chapter, then, is a critical examination of Fodor's claim that AD is the theoretical basis of semantic content.

1. Asymmetric dependence

The basic assumption underlying causal theories of content is that a term's content is determined by certain of its causal relations. But because a term's tokening typically has a variety of different causes, it is difficult for a causal theory to give a principle(s) which separates the relevant, content-determining causes of a term from the irrelevant ones. One particular reason such a principle is needed is to solve the DP. Consider again the source of the DP. Horses cause the mental tokening of 'horse', but once in a while (on a dark night, for example), cows also cause the tokening of 'horse'.[vi] If being a cause of a term's tokening is all that's necessary to be in the extension of the term, then cows on dark nights should be in the extension of 'horse' right along with horses. We need a principle to support our conviction that 'horse' refers to horses and not to horses or cows on dark nights.

Fodor summarizes his theory in the following way:

> "X" means X if:
> 1. 'Xs cause "X"s' is a law.

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

Condition 3 is intended to solve the DP. But what exactly does condition 3 say? What is asymmetric dependence? According to Fodor, for one causal connection to be asymmetrically dependent on another, the breaking of the latter connection must result in the breaking of the former, while the converse does not hold (Fodor 1990b, pp. 90-95). In the 'horse' case, Fodor says that if the properties of horses which are responsible for causing 'horse' were causally dissociated from 'horse' (i.e., if they were no longer to cause tokenings of 'horse'), then cows on dark nights would no longer cause 'horse'. However, the reverse does not hold. In the closest possible worlds where cows on dark nights do not cause 'horse', horses still do so.

How are we supposed to understand causal dissociation? According to Fodor, when you break the nomic connection between horses and 'horse', you're breaking the nomic connection between the property of being a horse and the property of being the cause of tokens of 'horse' (Fodor 1990b, p. 102). Fodor claims that he is talking not about something's being the cause of an actual, individual 'horse' token, but rather about the property of being a cause of 'horse' tokenings. The simplest way to understand Fodor here is to think of the property of being the cause of 'horse' tokens as being defined by description. The property of being a cause of 'horse' tokenings is just whatever property is shared by all of the things which cause 'horse' tokenings. In the limited domain of the cow/horse example, this property is a property shared by just horses and cows.

We should be concerned about Fodor's talk of the general property of being a cause of 'horse' tokens. What precisely would it take to remove either horses or cows from the set of things which have the property of being causes of 'horse'? In order to remove either horses or cows from the set of things which share the property of being a cause of 'horse', we would seem to have to dissociate some specific properties of cows or horses from human mind/brains and their tokening of 'horse'. 'Horse' tokens are tokens in the mental language of actual creatures, specifically, human beings. Thus, you can only remove horses from the set of things which cause 'horse' if you can say in virtue of which nomic alterations horses would no longer cause the tokening of 'horse' in humans. After all, people recognize horses and cows by causally interacting with specific properties those animals possess. It is perfectly acceptable for Fodor to talk about the abstract nomic connections between properties. But in order to explain the details of AD, Fodor should discuss the nomic connections between properties that actually matter, i.e., the properties of horses and cows in virtue of which

horses and cows belong to the class of things which can cause the tokening of 'horse'. Thus, in what follows, I interpret AD in terms of the nomic dissociation of the property of being a cause of 'horse' from properties possessed by horses and cows in virtue of which they can sometimes cause the tokening of 'horse'.

An additional restriction should be placed on which possible worlds we consider when evaluating the claim that a specific causal connection is asymmetrically dependent on another. If we are going to consider possible worlds where we sever the nomic connections between having a specific property(ies) and being a cause of a particular term of LOT, we must look at worlds where this connection is severed across the board. So, for example, if we imagine a world where the properties of being big and brown do not alone cause the tokening of 'horse' when people observe these properties in horses, then being big and brown cannot alone cause the tokening of 'horse' when observed in anything else (e.g., in cows).

2. Asymmetric dependence reformulated

In light of the preceding discussion, it seems wise to restate AD in a more precise fashion, so that we have a better idea of how the world is to be nomically altered in order to test for AD. Consider AD*:

> AD*- Bs causing 'A' is asymmetrically dependent on As causing 'A' iff in worlds where we nomically sever all of the observable properties of As from 'A', Bs no longer cause 'A', but not vice versa.

Applied to the cows/horses example, we get:

> Cows causing 'horse' is asymmetrically dependent on horses causing 'horse' iff in worlds where we nomically sever all of the observable properties of horses from 'horse', cows no longer cause 'horse', but not vice versa.

A person might mistake cows on dark nights for horses because cows share certain gross observable properties with horses, properties like being big, being four-legged and tending to hang around in open fields. If we apply AD*, in the worlds under consideration, these properties as well as all other observable properties of horses do not cause the tokening of 'horse'. Therefore, these properties will not cause 'horse' when instantiated in anything--horses, cows or otherwise. Therefore, the first part of the AD condition is met; dissociate all observable horsey properties from 'horse', and cows no longer cause 'horse', right?

Not so fast. It's possible that even in worlds where none of horses' observable properties (perhaps even none of their properties whatsoever) cause 'horse', cows can still cause 'horse' via their initiation of an inferential process. It would seem that no matter what worlds you go to, no matter what nomic connections

you break between horses and 'horse', the possibility will always remain that cows can cause 'horse' by (at the very least) initiating a train of thought which includes a tokening of 'horse' (and the same is true for horses). In response to Louise Antony's and Joseph Levine's criticisms of his views (Antony and Levine 1991), Fodor's own comments suggest that if we break the nomic connections between the observable properties of horses and 'horse', cows can still cause 'horse'.

Here's a little context. Consider the set of all proximal stimuli that can cause 'horse'. Antony and Levine call this set 'P(INF)' (Antony and Levine 1991, p. 13, presumably, because this is an infinite set of proximal stimuli). They express the worry that 'horse' really refers to P(INF), to proximal stimuli, not to actual horses:

> "[N]on-P(INF)-caused H-tokenings ['horse' tokenings] are asymmetrically dependent upon P(INF)-caused H-tokenings, because horses have to effect H-tokenings through P(INF)s. It looks like we have prima facie reason to say that H-tokens mean P(INF). (Antony and Levine 1991, p. 13)

Fodor's not worried. He says that if right now you break the nomic connection between P(INF) and 'horse', there are other ways to cause a 'horse':

> Presumably, I can figure out that there must be a horse behind the bush, thereby achieving a detached tokening of "horse", without tokening any proximal stimulus belonging to P(INF). Close your eyes and think 'horse'. See? Easy. (Loewer and Rey 1991, p. 313)

Here Fodor is denying that horse-caused 'horse' tokens are AD on P(INF) caused 'horse' tokens because breaking the connection between P(INF) and 'horse' does not break the connection between horses and 'horse'. There are still indirect ways for horses to cause a tokening of 'horse'. Fodor's view seems to imply that AD* is an inaccurate interpretation of AD. Fodor seems to be saying that if you nomically sever observable properties of horses from 'horse', you haven't severed the nomic connection between horses and being a cause of 'horse'. However, similar reasoning would lead us to say that, in the same situation, you haven't severed the connection between cows and being a cause of 'horse' either. When we sever the connections between P(INF) and 'horse', the nomic connection which still exists between horses and 'horse' also exists between cows and 'horse'. Somebody's telling me to think of a horse when they see one will cause me to token 'horse', as Fodor suggests. However, I will also token 'horse' in the case where someone sees a cow and, wanting to deceive me, says "There's a horse."

In bringing what Fodor says about P(INF) to bear on the evaluation of AD*, I am assuming that talk about breaking the nomic connections between P(INF) and 'horse' is not importantly different from talk about

breaking the nomic connections between all of the observable properties of horses and 'horse'. This seems reasonable. If you are actually going to sever the nomic connections between 'horse' tokenings and the property of being a cause of 'horse', you have to choose some specific point in the causal chain to make the break. AD* breaks the connections between the proximal stimuli which normally cause 'horse' and the causes of those proximal stimuli (i.e., the observable properties of horses). AD* could have been formulated so as to break the connections between the proximal stimuli which normally cause 'horse' and 'horse' (such that horsey looking patterns on my retina, for example, would just no longer cause me to token 'horse'. This seems to be what Antony and Levine have in mind.)[vii] The breaking of the former connections seems a bit easier to grasp; thus my chosen interpretation. Nothing seems to turn on this choice.

Does Fodor's proposed solution to the P(INF) problem leave room for him to endorse AD*? First let's look more closely at what Fodor has to say about P(INF). If we are to say that 'horse' refers to horses and not to P(INF), P(INF)-caused 'horse' tokenings must be AD on horse-caused 'horse' tokens. This means that if we nomically dissociate P(INF) from 'horse', horses should still cause 'horse'. So Fodor has to give us an example of how horses can still cause 'horse'. The closest he's come is to suggest that someone would tell us that there is a horse present (behind the bush or otherwise). But Fodor's suggestion is inadequate. If someone's telling me that there is a horse present is proximal stimulus sufficient to cause 'horse', then such stimulus will be in P(INF), and it will no longer cause the tokening of 'horse'. Remember P(INF) consists of all proximal stimuli which can cause 'horse'. Thus, any stimulus (including someone's telling me that there is a horse present) that could now cause me to token 'horse' will, by stipulation, not cause me to token 'horse' when P(INF) is nomically severed from 'horse'.

Fodor seems to conflate two questions, both of which are suggested by Antony and Levine's commentary. One question is "Will there be any non-P(INF)-caused 'horse' tokens when we sever P(INF) from 'horse'?" The answer to this question might be 'yes'. It may be that some people can free-associate or in some other way spontaneously token 'horse', i.e., these people may be able to token 'horse' independently of any proximal stimuli. Thus 'horse' may be semantically robust relative to P(INF) in that things other than members of P(INF) can cause 'horse' even after P(INF) is nomically severed from 'horse'. (On the other hand, the criticisms of the preceding paragraph suggest that even seemingly spontaneous 'horse' tokenings are in jeopardy, for any stimulus that might trigger your thinking 'horse' would be in P(INF), and thus, would no longer trigger 'horse'.) If there were truly spontaneous 'horse' tokenings, not caused by any proximal stimulus

at all, this would prove the existence of some non-P(INF)-caused 'horse' tokens, but they would not be enough to prove that 'horse' does not refer to P(INF). The relevant question is "Will there be any horse-caused tokens of 'horse' when we sever P(INF) from 'horse'?" A 'yes' answer here is necessary for Fodor's theory to tell us that 'horse' refers to horses rather than to P(INF). Fodor's example of a person telling me a horse is behind the bush won't do. A person's telling me anything provides an audio stimulus which will be in P(INF), and will not cause 'horse' in the possible worlds under consideration. Fodor seems to recognize the difficulty of the situation in the following passage:

> In fact, given that tokenings of "horse" are often theory mediated, P(INF) probably includes every proximal stimulus since,...the merest ripple in horse infested waters can produce proximal stimuli which cause "horse" tokenings in the mind of a properly [or improperly, natch] informed observer. (Loewer and Rey 1991, p. 256)

In the face of the all encompassing nature of P(INF), Fodor proposes to solve the P(INF) problem by disqualifying P(INF) as the sort of property that could enter into a causal relation on which another causal connection could be AD. Fodor claims that P(INF) does not correspond to a property that can enter into nomic relations (i.e., that the property corresponding to P(INF) is anomic). The problem with P(INF) is that it's description constitutes an open disjunction, and thus does not express any unified or principled property (Loewer and Rey 1991, pp. 256-257). P(INF) can't be the reference of 'horse', then, because the property corresponding to P(INF) cannot be the property on which all other non-P(INF)-caused tokenings of 'horse' are asymmetrically dependent. A nomic connection between properties can only be dependent on another if the second is a bona fide property, which the property corresponding to P(INF) (if there is one at all) is not. Fodor expresses concern, however, at the metaphysical assumptions made by this solution to Antony and Levine's P(INF) problem (Loewer and Rey, p. 257).

Disregarding Fodor's metaphysical scruples for the moment, note that Fodor's solution to the P(INF) problem does nothing to increase the plausibility of AD*. In fact, Fodor's remarks about the enormity of P(INF) only seem to make matters worse for AD*. If any old proximal stimulus is capable of causing the tokening of 'horse', then nomically dissociating the observable properties of horses from 'horse', as AD* does, would not be enough to stop cows (or horses!) from causing 'horse', as it must in order to be an accurate explication of AD.[viii] In order to break the nomic connection between horses and being a cause of 'horse', it seems that we have to make sure that no observable property of anything can cause 'horse'.

Taking the preceding discussion into account, we might try to reformulate AD. Instead of AD*, perhaps we should have AD#:

> AD#- Cows causing 'horse' is AD on horses causing 'horse' iff in worlds where we nomically dissociate all of the observable properties of everything from 'horse', horses no longer cause 'horse'; but not vice versa.

Does this work? If we break the connection between 'horse' and all of the observable properties of everything, then horses won't cause the tokening of 'horse'; true enough. However, AD# fails miserably when we get to the vice versa clause of the AD test. In order to satisfy the vice versa clause of Fodor's AD test, it must be that when we nomically dissociate cows from 'horse', horses still cause 'horse'. But how do we nomically dissociate cows from 'horse'? According to Fodor, the only way to do this is to break the nomic connections between all observable properties of everything and 'horse'. This is the only way to break the nomic connection between cows and 'horse' because we have to ensure that cows won't somehow indirectly cause the tokening of 'horse'. However, if all of the observable properties of everything are nomically severed from 'horse', then horses will no longer cause the tokening of 'horse', and the vice versa clause of AD fails. The cow-to-'horse' connection is not AD on the horse-to-'horse' connection, and thus, the DP goes unsolved. (Fodor's theory fails to satisfy Criterion #1.)

Perhaps I've placed too much emphasis on Fodor's remarks on the enormity of P(INF) and the myriad ways in which an LOT term's tokening can be caused. It's true that inference plays an enormous role in the tokening of LOT terms, but maybe we can downplay the importance of inference by limiting the relevant causal interactions which AD must take into account. Consider an alternative explication of AD, ADD (the second 'D' standing for 'direct'):

> ADD- Cows causing 'horse' is AD on horses causing 'horse' iff in worlds where we nomically sever all of the observable properties of horses from 'horse', cows no longer directly cause 'horse', but not vice versa.

Even though Fodor does not suggest that we distinguish between the direct and indirect causation of a term tokening, doing so may save the AD theory as a theory of content for natural kind terms of LOT.[ix] Formulation ADD tells us to ignore Fodor's remarks about ripples in horse infested waters, and instead focus only on cases where horses or cows directly cause a tokening of 'horse'. Saying that As directly cause the

tokening of t is to be cashed out in something like the following way: An A directly causes the tokening of t in S if S's tokening of t is the result of the A's impinging on S's sensory apparatus, and upon the A's impinging on S's sensory apparatus, S tokened t as a direct output of S's sensory modules.[x] One should have serious concerns about the distinction between direct and indirect causation.[xi] However, for now, let's put aside such concerns and see how far we can go with ADD.

Formulation ADD works fine in the first of our two test cases. When we nomically dissociate 'horse' from all of the observable properties of horses, neither horses nor cows will directly cause 'horse' tokenings. Cows can directly cause 'horse' only when one or more of the observable properties cows share with horses impinges on the subject's sensory apparatus, thus causing 'horse'. However, in the worlds under consideration, these observable properties no longer cause 'horse' when they impinge on the subject's sensory apparatus. All is well so far.

The problem with ADD is that the 'but not vice versa' clause does not seem to be satisfied. It may be true that if you break the nomic connections between 'horse' and all of the observable properties of cows, cows will not directly cause the tokening of 'horse'. But once these properties are nomically dissociated from 'horse', it seems unlikely that horses will still directly cause the tokening of 'horse'. The observable properties of cows which sometimes cause people to token 'horse' would presumably be properties like being large animals, having four legs, and tending to hang around in pastures. Once we dissociate these properties from the property of being a cause of 'horse', what reason do we have to think that horses will still cause the tokening of the same LOT term 'horse'?

Fodor fails to give us a compelling reason to think that, in worlds where the observable properties which cows share with horses (and which sometimes cause the tokening of 'horse'), are causally severed from 'horse', we will still token the exact same LOT term, i.e., 'horse'. If the properties in question are nomically dissociated from 'horse', it is highly unlikely that any conceptual relations will exist between those properties (or the mental representations of them) and the term 'horse'. In other words, in the worlds under consideration, the properties of being large, being an animal, etc., are not part of our concept of horses. This is all the more reason to doubt the claim that in these worlds, we would still token the very same LOT term which we token in the real world when we think about horses.

Let's step back for a moment. Surely I can think about horses without my thoughts being caused by the properties of being big, being four-legged, or tending to hang around in open fields. Probably the most

common causes of 'horse' tokens in my brain are other thoughts, many of which do not explicitly represent the size, number of legs, or typical habitat of horses. However, for humans, the gross observable properties of horses which horses share with cows play an important role in our mastery of the concept of horses; the gross observable properties of horses are central features of our concept of horses. At the very least, Fodor is wanting a theory which provides individuation criteria for LOT terms and describes the conditions under which human mind/brains introduce new LOT terms. With respect to our current concern, such a theory, accompanied by the appropriate supporting arguments, would better inform our intuitions about what actual terms would be tokened in other possible worlds.

Hans-Robert Cram has offered a defense of Fodor's theory against criticisms similar to those raised here (Cram 1992). Cram suggests that Fodor explicate AD in terms of the breaking of just those causal connections to 'horse' that both horses and cows share. Speaking in terms of nomic pathways, Cram says that when we break the connection between cows and 'horse' tokens, we close of some off the nomic pathways from horses to 'horse'. However, there will surely be avenues left open for horses to cause 'horse', even though there will be none left for cows to cause 'horse'. Horses will still cause 'horse' as a result of horses' properties that they do not share with cows (e.g., the property of having that peculiar horsey curvature of the neck possessed only by horses). Thus the vice versa clause in AD is met (Cram 1992, pp. 66-67).

While Cram's explication of AD seems sensible enough, it is, as noted above, difficult to know what the brain will do when central parts of our concept of horse are no longer connected to horses. The question becomes especially pointed when we leave the simple diagram Cram offers (where only one 'horse' causing property is shared by horses and cows) and move to the real world where the shared 'horse' causing properties are numerous. Once all of the observable properties of cows that can cause 'horse' are nomically dissociated from 'horse', quite a great many of the properties which normally cause 'horse'-related-experiences no longer will. If I look at a horse and it doesn't look like a big animal, what reason is there to think that I will token 'horse'? I see none. Thus, Cram's version of AD does not convincingly solve the DP.

In summary, the first two versions of a concrete AD principle suggested above, AD* and AD#, both fail to solve the DP. ADD, as well as Cram's explication of AD, may or may not solve the DP; it's difficult to tell. If they do solve the DP, they do so at great cost. They do so only by assuming a distinction between the direct and indirect causation of LOT terms, or by assuming that a favorable theory of LOT term introduction and individuation is forthcoming.

3. Additional objections to the Asymmetric Dependence Theory

a. Nomically impossible worlds

The plausibility of Fodor's approach rests essentially on intuitions about what would happen to people's mind/brains in nomologically impossible worlds. Insofar as people have strong intuitions here at all (and many people don't, myself included), I see no reason to trust these intuitions. Fodor seems to anticipate this sort of objection. At one point, he compares the counterfactuals on which he relies in the description of the AD theory to the idealizations made by respectable scientists, e.g., those assumed by the ideal gas laws (Fodor 1990b, pp. 94-95). (In terms of the criteria stated at the outset of the chapter, Fodor is claiming that his theory satisfies Criterion #3, regarding the standard practices of scientific theory construction.) Scientists do not need to know what would really happen in a world where the conditions specified by the ideal gas laws were to obtain, Fodor says, in order for these laws to "be in scientific good repute" (Fodor 1990b, p. 95). Similarly, Fodor claims, he does not have to be able to say exactly what would happen in the counterfactual situations which play a central role in his theory.

Fodor's analogy is strained. The ideal gas laws have proven their worth as predictive tools. The motivation behind the scientific community's acceptance of the ideal gas laws is their scientific success.[xii] The closer a system comes to satisfying the ideal conditions, the more similar its behavior is to the behavior of the idealized system. In contrast, Fodor is asking his readers to imagine possible worlds with a nomic structure which is very different from the actual one. Then the reader is asked to make intuitive judgements about what would happen to people's minds/brains in these worlds--no experimental results, no progressively closer approximations, not even a theory of LOT term introduction and individuation to inform our intuitions. Fodor's strategy is not anathema to good scientific theory construction. However, the idealizations Fodor employs are not 'in good scientific repute' in the same way as, or for the same reasons that, the ideal gas laws are. Given the breakdown of Fodor's analogy between the ideal gas laws and the counterfactuals to which he appeal, we are left with no reason to think that his counterfactuals are in good scientific repute. This suggests we give an agnostic vote at best when checking to see whether Fodor's AD theory measures up to the demands of Criterion #3.

To be fair to Fodor, many of the criticisms I've made of his work are given without regard to what Fodor claims to be doing in 1990b. In response to skepticism regarding the applicability of his theory of

content to humans, Fodor says:

> Don't forget, this stuff is supposed to be philosophy. In particular, it's an attempt to solve Brentano's problem by showing that there are naturalistically specifiable, and atomistic, sufficient conditions for a physical state to have an intentional content. (Fodor 1990b, p. 96)

At this point, Fodor is denying that his theory of content should be taken as a theory of content for human thoughts or the constituents thereof. Instead, Fodor claims that he is only trying to show that intentionality could be a natural (i.e., physical) phenomenon (in response to Brentano's claim that intentional phenomena could not possibly reduce to physical phenomena). If this is Fodor's project, then it is illegitimate to object to Fodor's theory on the grounds that he makes unsupported claims regarding what would happen in human mind/brains in other possible worlds.

Fodor can define his project however he wishes. Given the definition of the project he has set, his theory may be above many of the criticisms made here (especially those related to the solution of the DP). On the other hand, I have set for myself a different project than Fodor's. I hope to bring us closer to finding a theory of content which actually applies to humans. Furthermore, one might take my project to be one of great interest, and one which extends Fodor's project in the following way. Fodor has attempted to show that it is possible for intentionality to be part of the natural order. We might then view Fodor's project as a necessary preliminary to the project of showing that the intentionality of human mental life is part of the natural order. However, even on this view it seems reasonable to want to put Fodor's AD theory to the further test of seeing whether it's the correct theory of intentionality for humans.[xiii]

Conciliatory remarks aside, the situation seems worse for Fodor than I've described it as being. BTT should not be seen as a further extension of Fodor's project, carried out only after Fodor has shown that intentionality could be a natural phenomenon. Instead it seems necessary to develop an empirical theory of human intentionality before we can even tell whether Fodor's weaker, philosophical project is a success. Fodor claims that he is only offering sufficient conditions for intentionality, and that as a consequence, his AD theory cannot be refuted by showing that it fails to jibe with what we take to be facts about the meaning of humans' LOT terms (Fodor 1990b).[xiv] But if we cannot test Fodor's theory by seeing whether it applies correctly to human intentionality, then we seem to have no way to test it. Fodor describes asymmetric dependence and seems to say about it "Wouldn't that suffice for intentionality?" However, if we don't base our response to such a question on facts related to human intentionality, we simply don't know whether or not

AD provides a description of a situation sufficient for intentionality to exist. Humans are the only creatures with whom we are acquainted which have rich, intentional lives. This being so, it seems impossible to gain any deep naturalistic understanding of what intentionality is (and whether a particular theory like Fodor's sets appropriate conditions for intentionality) without first understanding how intentionality is instantiated in humans.

b. Atomism and asymmetric dependence

Imagine that there exist two kinds of things, As and Bs, and that these things are observationally indistinguishable to you. These kinds might be fire birds and trans ams, elms and beeches (Putnam 1975, pp. 226-227), or authentic John Lennon autographs and well-forged ones. Imagine also that you have a term, 'A', which, because of your interest in historical or underlying physical properties of As, is supposed to refer only to As. So, for example, you might be an autograph collector who is considering paying a handsome price for a John Lennon autograph. Even though you can't tell an authentic John Lennon autograph from a well-forged one, it is important to you that you get a real John Lennon autograph. In cases like these, we want to be able to say that a subject has an LOT term the extension of which includes the members of only one of the two kinds, even though the subject can't tell members of the two kinds apart.

In a now classic paper entitled 'The Meaning of "Meaning"', Hilary Putnam presents an example much like the ones just described (Putnam 1975, pp. 223-235). Putnam's example involves the existence of two planets, earth and another plant very much like earth called 'twin-earth'. Twin-earth is similar to earth in almost all respects, but there is at least one very important difference. On twin-earth, they do not have water as we know it. On twin-earth, there is no H20. Instead, twin-earthlings drink, swim in, and wash their clothes in a liquid which is indistinguishable from H20 to the lay earthling or the lay twin-earthling. This twin-earth water-like stuff, that seems so much like our water, is composed of the mystery compound XYZ.

Now assume, as Fodor does, that when an earthling has the thought, 'Water is wet', or, 'I would like a drink of water', the earthling is thinking about H20 and not XYZ. Fodor's view runs into trouble here because asymmetric dependencies alone do not seem to assign the correct content to the earth term 'water'. Given that H20 and XYZ are indiscriminable to us, it would seem that if one of the compounds were to no longer cause the tokening of 'water' then neither would the other compound. The properties of H20 which cause us to token `water' (e.g., being wet, being a liquid, quenching thirst) are the very same properties which would make us token 'water' in response to XYZ. Break the connection between those properties (i.e., the

properties of being wet, being a liquid, quenching thirst, etc.) and the tokening of 'water', and as a result, neither H20 nor XYZ will cause the tokening of 'water'. Therefore, it would seem that no asymmetric dependence exists, and 'water' must refer to both H20 and XYZ, according to Fodor's theory, rather than to H20 alone.

Fodor suggests a way out of this problem. According to Fodor, the fact that 'water' is a natural kind term establishes an asymmetric dependence of the right sort to make 'water' refer to H20, but not XYZ. Because 'water' is a natural kind term, speakers and thinkers intend it to refer only to the natural kind which local samples exemplify. (Presumably, similar intentions regarding historical pedigree could be used to separate authentic John Lennon autographs from forgeries.) In Fodor's words, "[T]he intention to use 'water' only of stuff of the same kind as the local samples has the effect of making its applications to XYZ asymmetrically dependent on its applications to H20 ceteris paribus." (Fodor 1990b, p. 115)

Fodor claims that our intentions fix the content of 'water'. However, the idea that the content of 'water' is fixed with the help of some existing contentful intentions does not square with Fodor's claim that his theory of content is atomistic. According to Fodor, if a theory of content is atomistic, then any term could have its meaning fixed independently of any other term. In particular, ""cow"s could be asymmetrically dependent on cows in a world in which no other asymmetric dependencies obtain." (Fodor 1990b, p. 92) In order to properly fix the content of 'water', one has to have certain intentions, according to Fodor. But intentions are mental states with content; therefore, atomism fails. It's not possible for XYZ-caused tokenings of 'water' to be asymmetrically dependent on H20-caused tokenings of 'water' in a world where no other asymmetric dependencies hold. In order for the connection between XYZ and 'water' tokenings to be AD on the connection between H20 and 'water', there must be in effect enough asymmetric dependencies to fix the content of the terms out of which the relevant, content-determining intentions are constructed (e.g., the content-determining intention to use 'water' only to refer to stuff of the same kind as local samples).[xv]

Fodor has another response available. In the summary of his AD theory, Fodor claims that for 'X' to refer to Xs, some 'X's have to be caused by Xs. 'Water' tokens have never been caused by a sample of XYZ. There's no such thing as XYZ. Fodor avails himself of this response at one point, after having moved away from what he calls a 'pure-informational' theory (i.e., one which is based entirely on lawlike relations) to a theory of content which takes into account a subject's actual history when assigning LOT term extensions (Fodor 1990b, p. 120).[xvi] (Of course, this response won't work in the trans am/fire birds example, given

the plausible assumption that our subject has encountered at least one of each of these two models of automobiles.)

Regardless of where Fodor comes down on the twin-earth case, the issue of atomism remains important. In Psychosemantics, Fodor argues that philosophers' current commitments to meaning holism are unfounded (Fodor 1987, Chpt. 3, passim). And Fodor's commitment to atomism in his revised theory of content is intended to keep holism at bay (Fodor 1990b, p. 127). Fodor is wrong, however, to assume that any violation of atomism leads to meaning holism. In fact, the first of Fodor's two solutions to the twin-earth problem suggests a way that one might construct a theory of content which violates extreme atomism without giving in to holism. It may be that one can construct a theory of content where only some terms have their content determined independently of others. A subject could then go on to use those terms whose content was fixed atomistically to fix the content of other, new terms. So long as any term which plays an essential role in the grounding of another term has had its own content previously determined, we avoid holism. I think that this hierarchical approach to building a theory of content is the correct way to go. This is the type of approach I pursue in chapter III.

Notes to Chapter II

---

[i] For readers familiar with Dretske's earlier work on content, the indication relation is very similar to what Dretske calls the relation of 'bearing information about' (Dretske 1981). In Dretske 1988 (pp. 58-59), Dretske acknowledges the connections between his earlier and his later views, and in particular, the similarity between the idea of indicating and the idea of bearing information about.

[ii] Somewhat similar attempts have been made to save Dretske's earlier theory of content (found in Dretske 1981) by Kim Sterelny and Jerry Fodor. Not to give the wrong impression in Fodor's case, his defense of Dretske precedes an all out attack on Dretske's earlier theory. See Sterelny 1990, pp. 121-123, and Fodor 1990b, pp. 62-63.

[iii] Explaining the Bell inequalities is almost as challenging to the imagination, but not quite. Even on the most metaphysically mysterious interpretation of what happens in the Bell experiments, particles change their current properties as a consequence of simultaneous effects from a distance, which effects are somehow determined by the particles' histories (not their futures) (cf. Teller 1988).

[iv] Cummins offers another forceful criticism which applies to both Dretske's old and new theories of content. Both theories allow determinate content only when that content is generated by learning. This is problematic, however, because many of the representations referred to by leading theories in the cognitive sciences are claimed to be innate (Cummins 1989b, p. 68, and Cummins 1991, pp 105-106). Also see Cummins 1996 (pp. 45-47, 54-57) for criticisms of the attempt to ground content on success. BUT IN SOME WAY THIS IS WRONG—DRETSKE DISALLOWS INNATE BELIEFS, NOT INNATE REPRESENTATIONS

[v] Actually, Fodor's first attempt was made in a widely circulated manuscript which was published after Fodor renounced the views expressed therein (Fodor 1990a). I ignore the views of Fodor 1990a in the text, not only because Fodor disavows these views, but also because the theory of Fodor 1990a is not causal/covariational in nature, and thus does not rest the same basic assumptions as BTT.

[vi] You might find it unlikely that you would mistake cows on dark nights for horses (at least you wouldn't that you would do it with regularity and conviction). But this is Fodor's example, and I'll stick with it. Although it may seem a bit far-fetched, Fodor's example actually serves his purposes pretty well. What Fodor wants is an example of an obvious case of misrepresentation which doesn't necessarily occur more than once. What the example lacks in naturalness it makes up for by allowing us to avoid the possible complications of scientific cases, cases where we regularly mistake one thing for another, or cases where the cause of the mistaken tokening of a term is quite a bit like the things which we would intuitively put in the extension of that term.

[vii] A third option would be to talk about breaking the connections between the property of being a horse and the observable properties of horses. On this option, however, AD obviously fails in the horse/cow case. Breaking the connections between the property of being a horse and the observable properties of horses may stop horses from causing 'horse' (and it may not, if we take Fodor's remarks about P(INF) seriously), but it wouldn't stop cows from causing 'horse'. Cows would still have the same old 'horse' causing observable properties. Thus there would be no AD of cow-caused 'horse' tokens on horse-caused 'horse' tokens. Manfredi and Summerfield seem to have this kind of example in mind when they object to Fodor's theory on the grounds that cows could "lose most of the perceptible properties in virtue of which the cow-to-'cow' connection is mediated" (Manfredi and Summerfield 1992, p. 265) and horses could still cause 'cow'. Although the terms 'cow' and 'horse' in Manfredi and Summerfield's example are reversed, it is structurally isomorphic to the one under consideration here. Break the connection between the property of being an X

and X's observable properties, and Xs will no longer cause 'X'. But at the same time some, Ys, which we do not want to be in the extension of 'X', could still cause 'X'. The result is contrary to the one desired, for Ys causing 'X' is not AD on Xs causing X . Fodor could well respond to Manfredi and Summerfield by claiming that they have chosen the wrong interpretation of AD. Regardless of what Fodor might say, Manfredi and Summerfield's example seems to show that we should not interpret the application of AD as requiring us to sever nomic connections between natural kinds and their observable properties.

[viii] The problem persists when we translate AD* into the terms of Antony and Levine's discussion and talk of the connections between proximal stimuli and 'horse'. If we nomically sever all of the proximal stimuli caused by the observable properties of horses from 'horse', there will still be, on Fodor's view, other proximal stimuli which horse can indirectly cause, which can then cause 'horse'. Remember, Fodor thinks that any old stimulus, given the right circumstances, can cause 'horse' (including stimuli caused by cows).

[ix] Note also that distinguishing between direct and indirect causation of LOT terms does not interfere with Fodor's solution of the P(INF) problem. Fodor can dismiss P(INF) as anomic, while at the same time denying the importance of the indirect causation of LOT term tokenings.

[x] The contrast between direct and indirect causation may not be a contrast to which Fodor is unfriendly. Fodor 1983b describes the sensory systems as modules, whose workings are not penetrated by the subject's theory of the world and whose outputs would seem to be LOT terms. Also see Fodor 1983a for an endorsement of the distinction between observation and theory in the spirit of the distinction between the direct and indirect causation of LOT terms. There Fodor clearly supports the distinction between theory-caused LOT terms and observation-caused LOT terms. The question is whether he would support a distinction between A's direct impingement on the subject's senses and A's indirect impingement, relative to the causing of t.

[xi] For many, these concerns are prompted by such skeptical works as Quine 1953, essay II, and Kuhn 1962. Similar concerns about the distinction between observable properties and unobservable properties bear on AD* and variations of it. Perhaps Fodor's AD theory is immune to such concerns. I'm not claiming that AD* is a perfect interpretation of Fodor, but he doesn't describe a better way to apply AD to assign extensions to humans' LOT terms. And I'm running out of ideas.

[xii] It is a legitimate question whether science really accepts these laws as true or, instead, only accepts as true claims about how close these laws come to describing actual physical systems. Ron Giere argues rather convincingly that scientific laws merely define abstract models which are then compared to real systems (Giere 1988, chpt. 3; also see Cartwright 1983, essay 6, for a similar view.) The claims of science which are true or false are claims about the degree to which real systems approximate (in the relevant respects) the abstract models. If Giere is correct, then Fodor is wrong to claim that an ideal gas law is something which "the theory itself tells us is true." (Fodor 1990b, p. 95) The ideal gas laws would be mere tools of approximation, not having any truth-value at all.

[xiii] It seems especially worthwhile to approach Fodor's work from this angle once one recognizes the poverty of well-developed, plausible, naturalistic theories of human intentionality which are also causal/covariational in nature. The causal/covariational nature of Fodor's theory makes his theory, taken as an empirical claim about humans, an appropriate candidate for critical scrutiny here because it is just this type of theory I offer in BTT.

[xiv] Fodor seems ambivalent as to the logical connections between AD and facts about human intentionality. Immediately upon claiming that he is only offering sufficient conditions for intentionality, Fodor attempts to motivate his theory by appealing to the reader's intuitions regarding meaning in natural languages (Fodor 1990b, pp. 96-100). He seems to be saying that his theory need not apply to human intentionality, but that if his theory is at all plausible, it is because the theory explains human intentionality.

[xv] Fodor might claim here that while the intention is needed to fix the content of 'water', the content of the intention is not what is operative in said fixing. The question then would be, "Why is an intention necessary

to fix content of 'water' when that intention's content is playing no essential role in the fixation of the content of 'water'?"

[xvi] See Baker 1991, where she criticizes of Fodor on the grounds that he does not give a consistent construal of AD throughout his presentation of it, and that he can only handle potential counter-examples to the theory by flip-flopping back and forth between differing, inconsistent versions of the theory.

III. THE BEST TEST THEORY OF CONTENT

A. The Basic Principle

1. The principle itself

Humans possess different types of concepts. We have concepts of natural kinds; for example, the concepts of tigers and of oxygen. We have concepts of artifactual kinds, concepts of things like hammers and chairs. We even have what we might think of as concepts of individuals, for example the concept of Bill Clinton or the concept of the Taj Mahal. We might expect there to be something common to the way extensions are fixed for all of these different kinds of concepts (or, more precisely, something common to the way the extensions are fixed for the LOT terms associated with all of these different kinds of concepts). Yet it would also seem reasonable to think that there is something unique to the way content is fixed for each type of concept. The Best Test Theory has the potential to satisfy both intuitions. As a comprehensive theory of content, BTT may shed light on the basic reference fixing mechanisms for all type of LOT terms.

The Best Test Theory distinguishes between two ways in which a natural kind term in LOT can have its extension fixed. Some terms, at some times, have their extensions fixed via the subject's employment of intentions (which themselves must, of course, have previously determined content). In contrast, some terms, at some times, have their extensions fixed independently of a subject's intentions. The latter case is the more basic of the two, and is the subject of BTT's first content-determining principle, BT1:

> BT1- If no other content-determining principle of BTT applies to a natural kind term t of LOT, then the extension of t is the natural kind of actual objects for which the concept associated with the term provides the best test.

At first blush, BT1 appears hopelessly vague. Consequently, the following sections explain in more detail how BT1 is to be interpreted and applied.

2. Application and interpretation

a. Success rates

A concept's being the best test for a natural kind is to be explained in terms of the comparison of success rates. In order to compare the success rates of different natural kinds relative to a given LOT term, we must first have in a hand a success rate for each of the natural kinds relative to the term in question. The success rate of a natural kind relative to a given LOT term is determined by the success rate function, f< K, S, t, m>. This function takes four arguments, one each of the following four types: a natural kind (K), a subject (S), one of that subject's LOT terms (t), and a time (m). When we plug an ordered quadruple from its domain into the success rate function, the function yields an output ranging from 0 to 1 (although this quantity is frequently expressed as a percentage, where, for example, 0.5 = 50%). The output of the success rate function is determined by dividing the number of times members of K have caused a tokening of t in S by the number of times members of K have caused a tokening of any LOT term in S. If, for example, members of K have caused the tokening of some term or other in S's LOT on 100 occasions, and 45 of the terms caused were tokens of t, then the success rate of K relative to S's term t is 45/100 = 0.45. 45% of the times that members of K caused S to token any LOT term at all, that term was t.

In more intuitive terms, the success rate of K relative to t is determined by the rate of success that members of K have in causing the tokening of t. Some kinds may be very efficient in their causing of tokens of t. Other kinds may not be so efficient. It should be clear, however, that, relative to S's LOT term t, many different kinds, at a single given time, can have success rates higher than 0.5. There is no requirement that the sum of the various natural kinds' success rates relative to t must equal 1, for some specific S and m.

The success rate of K relative to t is not to be identified with the percentage of the total number of S's tokenings of t which have been caused by Ks. For example, of all of my tokenings of 'horse', it may be that only 5% of them are actually caused by horses. Most of my tokenings of 'horse' probably result from other causes, like thoughts in a chain of reasoning or other people mentioning horses. This does not mean that the success rate of horses relative to my term 'horse' is 5%. To calculate a success rate, we must look at all of the times the members of a given kind have caused the subject to token any LOT term whatever and ask on what percentage of these occasions the subject tokened 'horse' as opposed to some other LOT term (e.g., 'cow'). Even if actual horses have only caused 5% of my 'horse' tokens, the success rate of horses relative to 'horse' may still be very high, far higher than the success rate of any other natural kind relative to 'horse'. For me, as well as for the typical subject, the success rate of horses relative to 'horse' is probably upwards of 99%. This because when I token an LOT term(s) in response to horses, I token 'horse' almost every time. No other natural kind has a success rate relative to 'horse' which even approaches the success rate of the natural kind

horse. This provides the justification of BTT's assignment of horses, and not the members of any other natural kind, as the extension of 'horse'.

The preceding discussion of 'horse' provides an illustration of BTT's general method of assigning extensions to natural kind terms in LOT. In order to find out whether natural kind K is the kind whose members make up the extension of t for S, we compare the success rate of K relative to t for S to the success rate of all other natural kinds relative to t for S. If K has a higher success rate relative to t than any other natural kind, the members of K constitute the extension of t. It is in this sense that S's concept associated with t provides the best test for Ks.[i]

Herein lies BTT's solution to the DP. Relative to the term 'horse', for example, other natural kinds whose members may sometimes cause a subject to token 'horse' have success rates far lower than the success rate of horses relative to 'horse'. While a cow on a dark night may occasionally cause a subject to token 'horse', cows normally cause the subject to token other LOT terms. Of the total number of occasions on which cows cause a subject to token any term of LOT at all, very few of these will be tokens of 'horse'. For the typical subject, then, the success rate of horses relative to 'horse' is much higher than the success rate of cows (and, presumably, much higher than the success rate of any other natural kind relative to 'horse'). Thus, BT1 implies that for the typical subject, 'horse' refers only to horses.

The Best Test Theory solves the DP by making what I call the 'natural kinds only' assumption (or 'NKO'). By calculating success rates only for homogeneous natural kinds, not disjunctions of natural kinds, we guarantee that in the typical case, reference is not disjunctive. If BTT were to allow the consideration of the success rate for the disjunction horse or cows which are encountered under the specific circumstances which make them look like horses, then BTT would not solve the DP. Chapter V consists of an extended argument in defense of NKO.

b. Concepts

As noted in chapter I, BTT's employment of 'concept' is a bit deviant. However, BTT's concept of a concept is not entirely disconnected from the theories of concepts found in the psychological literature. What follows is an extended discussion of the composition of BTT's concepts, specifically in comparison to analyses of concept composition given by psychological theories of concepts.

In terms of recent psychological work on concepts, the idea of concepts which BTT assumes is most

similar to the 'feature bundle' or probabilistic view (Smith and Medin 1981; also see Lakoff 1987 for a useful, albeit highly polemical, summary of the development of the probabilistic and other, related views.)[ii] Although the view of concepts which I ultimately embrace is distinct from the probabilistic view, it will be helpful to look at the probabilistic view in some detail in order to get an idea of how the two views differ.

According to probabilistic models of concepts, concepts are collections of weighted features. However, in contrast to traditional views about concepts, the list of features which constitutes a particular concept does not provide singly necessary and jointly sufficient conditions for the correct application of the concept.[iii] Instead, a concept consists of a list of features, each of which is assigned a confirmational, or diagnostic, weight. During the categorization process, the subject adds together (subconsciously, in the typical case) the weighted values of all of the features which are perceived. If the sum of these features exceeds a certain threshold, then the concept is applied, and the associated LOT term is tokened. Of crucial importance is the fact that on the probabilistic view, there is typically a variety of possible combinations of features which, when their weights are added together, will cause the positive application of a particular concept. In other words, on the probabilistic view, there is no single combination of features which must be present in order for a given concept to apply.

The probabilistic view of concepts suggests a certain theory of extension, what we might call a 'feature-match' theory, which is similar in some respects to BTT. According to a feature-match theory of extension, a concept applies to a certain set of objects (and the LOT term which is associated with the concept refers to those objects) because those objects match, to a satisfactory degree, the features on the concept's list. To make the feature-match theory of extension more precise, we can construct a feature-match analogue to BT1. On such a view, the extension of a concept (or the associated term in LOT) is determined by testing the concept against the various natural kinds to see which natural kind is best described by the concept's weighted feature list.[iv]

Eleanor Rosch and her associates (whose work was instrumental to the development of the probabilistic view) seem to adopt some version of a feature-match view when discussing the question of the nature of basic categories (e.g., dog, chair, or car).

> A working assumption of the present research is that in the real world information-rich bundles of perceptual and functional attributes occur that form natural discontinuities and that basic cuts in categorization are made at these discontinuities.(Rosch, et. al., 1976, pp. 384-385)

This passage suggests the picture of a world where some kinds of things present distinct bundles of perceptual and functional information to us. And in this world, our concepts (at least our basic concepts) represent, or match, enough of the structure of these bundles of information to secure reference. In terms of the feature-match analogue to BTT, Rosch et. al.'s view suggests that when we look at the degree of match between a given concept and the full complement of natural kinds, the contest between the natural kinds will not be close. One natural kind will stand out among the many as the kind for which the concept in question clearly provides the highest degree of fit. If our concept is a basic level concept (and is also a natural kind concept), it will match one natural kind quite nicely, whereas it will not match other kinds very well.[v]

Ultimately, the semantic nature of the feature-match view limits the usefulness of such an approach. But before criticizing the feature-match approach in detail, I propose to look more closely at the nature of features out of which concepts are supposed to be constructed according to probabilistic theories of concepts. It is important to understand the constitution of BTT's concepts not so much for their role in the application of BT1. At the end of the current section, I will give reasons for thinking that BT1 can, at the risk of expositional convenience, be explained without any reference to concepts and features. However, because the further principles of BTT make essential reference to concepts, it is worth our while to get clear as to how, precisely, these are to be understood according to BTT.

In their exposition of the feature-based, probabilistic view of concepts, Smith and Medin give the following list of features as an example of the weighted features included in the concept Bird: moves, winged, feathered, flies, sings, and small size (Smith and Medin 1981, p. 63). The features which Smith and Medin list are typical of the features cited in discussions of feature-based theories of concepts (cf. Smith, et. al., 1988, Rosch and Mervis 1975, Rosch, et. al., 1976, Armstrong, Gleitman, and Gleitman 1983). These features have certain traits in common. In particular, they are lexical concepts (concepts which are expressed by a single word in a public language), and they are features which are salient at the level of conscious experience (and thus such features combine to make concepts which fit the common description of them as items that people consciously understand or grasp).

Being lexical and easily accessible to consciousness makes certain features obvious candidates for feature-hood, and may make these features easy to work with experimentally. However, Smith and Medin do not invoke lexicality or ease of conscious access to justify their focus on features like the Bird features listed above. Instead, Smith and Medin give three guidelines for identifying the featural elements of concepts. The

first of their proposed constraints on feature identification, constraint #1, is based on the idea that "a property is a useful feature to the extent that it reveals many relations between concepts." (Smith and Medin 1981, p. 15) Ideally, this constraint would yield a set of features which would "exhaust all potential relations between the concepts of interest." (Smith & Medin 1981, p. 15) Constraint #1 encourages us to look for features which draw contrasts between various concepts and also show us what important similarities exist among various concepts. The feature being male, for example, is a useful feature according to the standard set by constraint #1 because the feature male has great discriminatory value in contrasting numerous pairs of concepts, for example, Colt/Filly or Boy/Girl. Being male also shows what numerous subsets of concepts share, for example the sets {Father,Son} and {Boy,Colt} (Smith and Medin 1981, p. 15).

Smith and Medin's second proposed constraint tells us to "seek features with some generality, in the sense that a feature should apply to many concepts within a domain rather than to a few." (Smith and Medin 1981, p. 16) What seems odd about this second constraint is that it either adds nothing to the first, or when it does add something, it contradicts the first constraint. The first constraint tells us to look for patterns of relations between concepts, and clearly one of the most important and common relations is similarity among concepts. Thus, insofar as certain features (the ones that label similarities among concepts) are widely shared, constraint #1 already tells us to look for features which "apply to many of concepts" in our domain of interest. For this reason, constraint #2 seems redundant. On the other hand, if constraint #1 is to yield a set of features which "exhaust all of the potential relations between the concepts of interest", we have to be prepared to admit cases where a property is shared by only a few concepts in a domain, but still seems to offer some discriminatory value in separating the concepts which share the rare feature from all of the others which do not. Here constraint #2 seems to contradict constraint #1. Constraint #2 tells us to look for generality in our choice of features, but does not tell us why we should think that a feature is any less real simply because it is only shared by a small set of concepts. Constraint #1 directs us to find primitive features with respect to a particular set of concepts. If constraint #2 adds anything to constraint #1, constraint #2 seems to suggest that we disregard certain patterns and relations (and the primitive features which would make those relations apparent) without giving any justification for disregarding them.

Both constraint #1 and constraint #2 are supposed to disallow certain features which we might normally consider to be idiosyncratic, relational or for some other reason not legitimate featural elements of concepts. Smith and Medin label these features 'pseudofeatures' and give the following examples of such features: 'saw all of Tuesday Weld's movies in one week' and 'prefers Bloody Marys that contain two parts

Worcestershire sauce for one part Tabasco.' (Smith and Medin 1981, p. 15) Before evaluating Smith and Medin's classification of these features as pseudofeatures, we should examine their third (and final) constraint on the identification of features.

Constraint #3 is "a processing constraint: the features posited should serve as the inputs for categorization processes." (Smith and Medin 1981, p. 16) Of the three constraints, Smith and Medin consider constraint #3 to be the most important.

> Indeed, it must take precedence over other constraints, as we will readily accept a nonprimitive and nongeneral property as a feature if there is convincing evidence that it is used in categorization. (Smith and Medin 1981, p. 16)

Smith's and Medin's ensuing discussion emphasizes the ultimate authority of empirical evidence and the need to identify our features in a way which facilitates the empirically based explanation of people's categorizational abilities. Missing from Smith's and Medin's ensuing discussion of constraint #3 is any argument that constraint #3 rules out the targeted pseudofeatures. (This is of particular interest for my purposes here, because, as Smith and Medin do, BTT honors constraint #3 above all others.) In other words, whatever features are used in the categorization process will be considered elements of concepts. If empirical evidence related to the categorization process supports the use of features like Smith and Medin's pseudofeatures as conceptual elements, then these alleged pseudofeatures are full-fledged features, free to be included as elements of concepts.

There seem to be good reasons to think that the list of features which are required for an adequate, psychological characterization of a given concept will be richer, more detailed, and perhaps more offbeat than such features as are included in Smith and Medin's model of the concept Bird. In what follows, I pursue this line of thought in the following manner. Many of the features which are psychologically real (in virtue of their empirically observable effects on categorization) are not obviously of the same type as Smith and Medin's Bird features (i.e., not lexical and not consciously accessible). While, these features satisfy constraint #3, some of them do not seem to satisfy constraints #1 and #2. In an attempt to make the experimental findings consistent with Smith and Medin's claims regarding constraints on acceptable features, I suggest that we understand these features as beings composed out of features which themselves satisfy constraints #1 and #2. Ultimately, however, there would not seem to be any principled distinction between the empirically justified features under consideration and the features which Smith and Medin identify as

pseudofeatures. Beyond constraint #3, then, there seem to be no substantive constraints on what should count as a feature.

Consider certain perceptual features, such as the silhouette of an object, which may seem unnatural by constraints #1 and #2, but which play a detectable role in categorization. Rosch, et. al., 1976 reports a series of results which highlights the importance of geometrical outlines in the categorization of objects. Experiments 3 and 4 (pp. 398-405) begin with the experimenters choosing easily recognizable outline tracings of objects taken from a variety of categories. The researchers then average these outlines together within basic level categories, within superordinate level categories (e.g., Vehicle, Clothing, Animals, or Furniture) and within subordinate level categories (e.g., Sports Car, Kitchen Chair, or Dress Pants). For example, Rosch, et. al., geometrically average together two outline drawings of two different pairs of pants and also two outline drawings from pictures of two different types of clothing (e.g., one outline drawing of a pair of pants and one of a shirt). As one might imagine, subjects performed much better in categorizing objects by geometric outline when the composite outline is constructed using pictures of objects in the same basic level category or subordinate level category than when the composite in constructed using outline drawings of objects from two different basic level categories which fall under the same superordinate heading. These results seem unobtainable were it not for the psychologically real use of outline information in the categorization process.

The results reported above appear to force Smith and Medin to choose between constraint #2 and constraint #3. The feature having a car-shaped outline is not a feature shared by many concepts, i.e., the feature is not general. Thus the feature 'having a car-shaped outline' is not a legitimate feature according to constraint #2. As is suggested by their discussion of constraint #3, Smith and Medin could respond by saying that so long as people use 'car-shaped' in processing, the feature is legitimate. However, by deferring in this way to constraint #3, we remove the bias against Smith and Medin's pseudofeatures which is supposed to follow from constraints #1 and #2. If psychological reality is all that matters, then we should not pass judgement on the feature 'having seen all of Tuesday Weld's movies in one week' until we've done the relevant experiments.

Perhaps there's a way to make Smith and Medin's set of constraints consistent with respect to their application to geometrical outlines. For example, we might try decomposing outlines into component features (e.g., basic geometrical components) in an attempt to explain Rosch's experimental data in a way that makes such data consistent with all three of Smith and Medin's constraints. If these component features of

outlines are basic geometric components of visual templates, we would expect them to be common to many different outlines (satisfaction of constraint #2) and we would also expect the presence or absence of such features to highlight important contrasts between different outlines (satisfaction of constraint #1). Additionally, because the outlines used in categorization in Rosch's experiments are constructed out of these basic geometrical features, we have indirect evidence that the basic geometrical features are used in categorization (satisfaction of constraint #3).

Two questions arise, however. Firstly, have any plausible theories of geometric components been proposed? The answer to this question is 'yes'. Beginning with basic geometric tools such as various generalized cones and arrangements of vertices and axes, one can characterize the important differences between the typical outlines of members of non-superordinate level categories, e.g., cows, ostriches, cups, and telephones (cf. Biederman 1990, and Marr 1982, chpt. 5).

Secondly, does the existence of such theories give us reason to assert any important difference between the feature of having a particular, prototypical outline and a feature such as 'having seen all of Tuesday Weld's movies in one week'? Features of both types have the potential to be quite idiosyncratic. I assume that very few people have seen all of Tuesday Weld's movies in one week. Similarly, a particular prototypical outline may be very rare, say, because it is the outline typical of members of an endangered species. However, the feature of having a typical outline has the saving grace, which allows it to satisfy constraints #1 and #2, of being constructed out of elements which themselves are parts of a set of features whose members satisfy constraints #1 and #2.[vi] The catch is that the feature 'having seen all of Tuesday Weld's movies in one week' also seems to be composite in the relevant respect. The linguistic item 'has seen all of Tuesday Weld's movies in one week' is clearly constructed out of component parts. And if one assumes the existence of an LOT, there seems to be every reason to think that the mental analogue of 'has seen all of Tuesday Weld's movies in one week' is also constructed out of component parts. When we attribute the feature to an individual movie buff, it would seem that this feature is constructed out of elements which themselves are parts of a system of features which are valid features. (See Jackendoff 1989 for a theory of concept structure which seems general enough to allow all sorts of 'pseudofeatures' to be included as parts of concepts.) For example, the feature 'stars Tuesday Weld' would satisfy constraints #1 and #2 if one were studying the history of film or film criticism. If having a certain prototypical outline can count as a legitimate feature because of its being constructed out of components which satisfy constraints #1 and #2,

then so, it seems, would Smith and Medin's pseudofeatures.

Smith and Medin seem to leave the door open to features like their pseudofeatures even while denying the legitimacy of such features. In their discussion of concept stability, Smith and Medin point out that depending on one's theory of concepts, one may allow a lot of inter, as well as intra, subject variation with respect to a single concept (Smith and Medin 1981, p. 10). In particular, a feature-based, probabilistic view of concepts, together with a standard story about concept acquisition, seems to imply that as a consequence of personal experiences, subjects will vary with respect to which features are included in a given concept or with respect to how much diagnostic weight is assigned to various features. The result is that people may sometimes develop idiosyncratic diagnostic tests for the presence of individuals or members of kinds.

Consider the following case. You are a dogmatic hater of people whom you consider to be politically liberal. You've gotten it into your head that Woody Allen is a disgusting liberal-type, and you figure that only liberals would waste their time watching his movies. 'Has seen numerous Woody Allen films' might then have substantial diagnostic weight in your mind as a featural element of the concept Liberal.[vii] In contrast to you, there may be many other people with the concept Liberal who have never given any thought to a connection between Woody Allen movies and the political views of those who watch these movies. The preceding example seems to take Smith and Medin's idea of inter-subject concept variation to the extreme. However, this kind of variation can easily develop given the extreme diversity of personalities, opinions, and experiences among human subjects.

Smith and Medin suggest a way to minimize the importance of idiosyncratic features by proposing a distinction between the core of a concept and a subject's identification procedure for that concept (Smith and Medin 1981, pp. 20-21). The core of a concept would typically contain a list of abstract features (some of which may even be necessary to members of the kind to which the concept is supposed to correspond). In contrast, a subject's identification procedure would typically include perceptual features which are of practical use in identifying objects to which the concept applies.[viii]

If Smith and Medin were to feel compelled to acknowledge the featurehood of their alleged pseudofeatures, I assume that such features would be relegated mostly to subjects' identification procedures. One potential difficulty here is that it's not clear that Smith and Medin's distinction between a concept's core and its identification procedures is a tenable distinction. However, if we grant Smith and Medin this

distinction, it might be employed to make a feature-match theory of extension more plausible. The relegation of the alleged pseudofeatures to the category of identification procedures would allow the feature-match theory of extension to render Smith and Medin's pseudofeatures impotent vis a vis extensions. A feature-match theory could do so by claiming that the only features which determine extensions for an LOT term are the core features of the concept associated with that term.

Whether or not a feature-match theory of extension restricts its attention to the core features of a concept, a feature-match theory seems unsatisfactory as a theory of LOT term extension. The primary problem with a feature-match theory of extension is that the features out of which a concept is built, and by which reference is determined, are themselves semantically individuated. As defined, a feature-match theory of extension says that an LOT term refers to members of a kind if the concept associated with that term describes the members of that kind with enough accuracy. However, the idea of describing something accurately is a semantic idea. In order for a feature-matching theory of extension to work, the featural elements of the relevant concepts must themselves have their content already determined.

The preceding considerations do not prove that a feature-match theory of extension could not be fruitfully developed. However, the preceding does seem to show that a feature-match theory can not be as general or as independent as we would like. A feature-match theory of extension cannot apply to the features out of which concepts are built unless those feature are themselves constructed from component features. The problem of regress thus arises unless the smallest component features of concepts have their extension (or whatever aspect of their content is relevant) determined according to a different, non-feature-match theory of extension. (Similar problems arise for a feature-match theory of intensions.) A feature-match theory of extension, if at all useful, would be useful only as a dependent theory. It could never determine the content of an LOT term unless the content of some other LOT terms had first been determined in accordance with a different theory of content.

There are important differences between BTT and its feature-matching analogue. According to a feature-match theory, the fundamental mechanism of extension fixation is semantic; the extension of a concept is determined by way of finding the kind whose members are best described by the concept. This is the problematic aspect of the feature-match approach. In contrast, BTT specifies a causal mechanism of extension fixation. Instead of looking for semantic fit between concepts and natural kinds, BTT focuses on the causal connections between members of natural kinds and the natural kind terms in a subject's LOT.

According to BTT, concepts play an essential, intermediate role in the causal chain which runs from members of natural kinds to the subject's tokening of terms of LOT. However, on BTT, extension is ultimately determined by the patterns of causal connections, specifically those expressed by the success rate function. The Best Test Theory does not assume a distinction between a concept's pseudofeatures and its real features. Neither does it assume a distinction between the core features of a concept and identification procedures ancillary to that concept. Nor does BTT assume that the features of concepts are individuated semantically.[ix]

According to BTT, concepts play an important intermediate role in fixing reference. But in order to avoid the semantic dependence of a feature-match theory, so that the door is left for BTT to provide a general theory of extension, the role concepts play in determining reference in accordance with BTT must not depend on those concepts (or any parts thereof) already having semantic properties. Ultimately, then, BTT construes 'feature' very broadly and non-semantically. A featural component of a concept C can be any element of a subject's psychology which mediates the tokening of the LOT term associated with C.[x]

Given the preceding definition of a concept feature, it should be clear how odd BTT's concepts are in comparison to concepts, traditionally conceived of. The set of elements of a subject's psychology which play some role in mediating the tokening of an LOT term t seems like it could be huge. Recall the discussion of the P(INF) problem in chapter II. According to Fodor, any stimulus can cause the tokening of t under the right conditions. If Fodor is right, doesn't this mean that on BTT's definition of concepts, a subject only has one big concept which mediates the tokening of all LOT terms? Shouldn't BTT draw a distinction between a concept core and the related identification procedures (which procedures might involve any stimulus whatever), if for no other reason than to make BTT's concepts manageable in size?

Assume that Fodor is right about the extreme robustness of LOT terms, and that any stimulus can cause the tokening of any LOT term under the right circumstances. By focusing on the way groups of mechanisms work together in these different sets of circumstances to cause the tokening of LOT terms, BTT can yet avoid claiming that each subject only has one big concept consisting of a collection of all elements of the subject's psychology. The Best Test Theory should not define the concept associated with t as a certain collection of mechanisms. Instead it seems that BTT should define the concept associated with t as a set of specifications of the relations between all of the elements of the subject's psychology. Each of the elements in this set is a description of one way all of the elements of the subject's psychology might be related in order

to cause the tokening of t. Thus there is a distinct concept associated with each LOT term, but each of these different concepts mentions the same elements of the subject's psychology (in fact, each concept mentions each element of S's psychology numerous times).

At the level of BT1's application, we must be able to describe the role of concepts in fixing extensions without referring to the content of those concepts (or to the content of the parts of any of those concepts).[xi] Beyond BT1, when we specify a concept, some of the descriptions of concepts and their role in fixing extension make reference to psychological elements individuated according to their content. There we will have to take seriously the idea of concepts as made up of features with semantic content. Why this is so will become apparent during later discussions of extension-fixing principles BT2, BT3, and BT4. For now, I want only to note a certain implication of this distinction between the concepts present at the application of BT1 and those present at the application of BTT's further principles. Where BT1 applies, the content of a concept's elements plays no role in fixing the content of the terms to which BT1 applies. Thus, we can, if we so choose, avoid talk about concepts entirely in the explication of BT1. For the purpose of smooth exposition, it may be convenient to continue talking about concepts as testing for members of natural kinds. However, success rates are only sensitive to causal connections between the members of the natural kinds and tokenings of LOT terms. All reference to concepts in the explication of BT1 can, in principle, be dropped.

c. Levels of application

Above I used the 'horse'/cow example to illustrate how BT1 assigns extensions to natural kind terms in LOT and to illustrate how BT1 solves the DP. However, BT1 does not necessarily apply to the typical subject's LOT term 'horse'. The principle BT1 only applies to an LOT term whose extension is determined without the aid of any extension-fixing intentions. Some subjects, at some times, may have 'horse''s which fit this description, but many subjects, at many times, do not. The difficulty in separating cases where BT1 applies from those in which another extension-fixing principle applies stems from the difficulty in telling when a subject has an intention which is relevant to the fixation of content. Matters are further complicated by the question of exactly how explicit an intention has to be in order for it to be an extension-fixing intention.

Issues concerning the nature and the role of extension-determining intentions are too detailed to be

addressed at this point. However, to appreciate more clearly BT1's true range of application, it may be worthwhile to inquire after terms to which BT1 would apply. One likely candidate is the infant's LOT term 'object'.[xii] In the early development of the infant's object concept, the subject seems to bear no intentions whatever toward the LOT term 'object' that might play a role in fixing the reference of 'object' (where 'object' is the infant's LOT term which controls the his/her successful responses in object recognition and manipulation tasks).[xiii] The use of 'object' by infants seems reflexive in that the infant applies 'object' without any conscious thought regarding what the term's extension should be.[xiv]

This example raises questions, however, about cases where cognitive processing takes place largely at the sub-conscious level, where the existence of subconscious extension-fixing intentions may preclude the application of BT1. Even though, for example, the infant has no conscious intention to treat 'object' in a specific way, there may well be subconscious, yet explicit, representations guiding the infant's tokening of 'object'. The neural detection of the edge of a solid object via the detection of a zero-crossing segment in the firing of retinal cells (Marr 1981, pp. 64-66) is an example of a subconscious mental representation that might guide the infant's tokening of 'object'. Does a zero-crossing segment have its extension fixed by BT1? Do zero-crossing segments (or any other so-called 'lower-level' representations) have extensions at all?[xv]

A reasonable, general answer to the question of levels is to advert to psychology. Representations with extensions exist at any level at which the assumption of their existence facilitates the construction of good psychological explanations. If psychology gains empirical power from talk of representations at a certain level, then they exist at that level. To reach a conclusion regarding where representation first appears in visual processing, we should ask how far up we can proceed from the lowest level of visual processing (i.e., stimulation of individual retinal cells) before it becomes empirically expedient to posit representations. This is a difficult question to resolve. Marr claims that the bars, lines and blobs of his raw primal sketch are 'symbolic representations' which are 'physically meaningful' (Marr 1981, pp. 70-72). However, psychological explanation may get along just fine without positing representational states until one reaches the stage of the fully spatial, viewer centered representation which Marr calls the '2 1/2-D sketch' (Marr 1981, pp. 149-150, 278-279). It is at this stage, according to Marr, that the visual system first makes available, to other cognitive subsystems, a coherent picture of what is being seen. It may thus seem natural to assume that the elements of the 2 1/2-D sketch are the lowest-level 'representations' which influence behavior in a way

which compels the attribution of extension-based content.[xvi]

Two comments are in order. Firstly, whether or not one needs content for theoretical purposes below the level of the 2 1/2-D sketch is an empirical matter which cannot, and need not, be settled here. The same holds true for similar questions which arise in the debate over the nature of representation in connectionist models of cognition. Whether or not representation occurs at (and BT1 applies at) the level of 'microfeatures' is an as yet unanswered empirical question which is tangential to the development of BTT.

Secondly, so long as one gets referential content where it is theoretically useful (most likely at the higher levels of cognition), it may not make much difference what we say about content at lower levels of processing. If one can find no principled reason to limit the application of BTT to structures above a certain level, one may be free to apply BTT across the board without any great consequence. (However, whether such free application of BTT leads to conflict or not depends on the details of the relevant empirical theories.)

d. Causal connections

In order to calculate the success rate of natural kind K relative to an LOT term t, we divide the number of the subject's tokenings of t caused by members of K by the total number of times the subject has tokened any term at all in response to Ks. Lacking from the explanation of success rates thus far is the explanation of what it is for a member of a kind to cause a tokening of an LOT term. While it is beyond the scope of the present work to analyze the concept of a causal interaction, it is important to specify which causal interactions are relevant to the determination of success rates. For the purposes of understanding and applying BTT's principles, I propose the following criterion of causal relevance:

> CR1- A member k of kind K (or the group k1...kn of members of kind K) caused t in S iff had k (or the group k1...kn) not had causal effects on S, S would not have tokened t on the occasion in question.[xvii]

Criterion CR1 seems too liberal, for, CR1 seems to allow causal dependencies to be relevant to content determination even when they appear distant or irrelevant. Consider your heart. It seems that, according to CR1, your heart makes a perfectly legitimate causal contribution to the tokening of each and every LOT term you token. Had your heart not had some causal effect on your brain, you would never have tokened any LOT term.

As strange as it might sound, the heart example does not constitute an objection to CR1. Because your heart makes a causal contribution to every tokening of an LOT term, the success rate of hearts will be very low with respect to any particular LOT term. Choose an LOT term, 'object', for example. A heart will contribute causally to every tokening of 'object'. However, this does not mean that the success rate of hearts relative to 'object' is 100%. To calculate the success rate of hearts relative to 'object', one must divide the number of times heart causally contributed to the tokening of 'object' by the number of times hearts causally contributed to the tokening of any LOT term whatsoever. The former number will be but a minute percentage of the latter. Thus, the success rate of hearts relative to 'object' is bound to be much lower than the success rate of objects relative to 'object'.

The indirect causal effect which hearts have on the tokening of every LOT term does not form the basis of a legitimate objection to CR1. However, there may be other types of examples which show CR1 to be too liberal. Consider a case where a cat kills a mouse in your garage and the mouse rots there without your knowing it. Eventually you are confronted by the smell of the rotting mouse corpse, and you think 'Gross!'.[xviii] In this case, CR1 would seem to imply that the tokening of your LOT term 'gross' was caused, in the relevant sense, by a member of the natural kind cat. If there had not been a causal chain leading from the cat to you, you would not have tokened 'gross'. Such a determination of relevance would, of course, lower the success rate of cats relative to your tokening of 'cat', perhaps even to the point of making 'cat' refer to something other than cats.

The cat/'gross' example seems to illustrate something unsatisfactory about CR1, if for no other reason than that it makes you aware of the sheer number of caused events relevant to the calculation of any given success rate. As far as it has been developed, the cat/'gross' example does not prove the insufficiency of CR1, but it may prompt us to try to revamp CR1 so as to exclude the consideration of indirect causal relations (such as the relation that holds between the cat and the tokening of 'gross'). Thus, as a possible alternative to CR1, I propose the following criterion CR2:

> C̄R2- A member k of natural kind K (or the group k1...kn of members of K) caused the tokening of t in S iff had k (or the group k1...kn) not had a direct causal effect on S, S would not have tokened t on the occasion in question.

I have two concerns about CR2. As stated, CR2 lacks any explanation of the difference between direct and indirect causal effects. In chapter II, we sketched a distinction between a direct and an indirect

cause of the tokening of an LOT term, and we might defer to that distinction here. The downside of this approach is that the distinction between direct and indirect causation is notoriously difficult to make in a way that stands up to counterexamples.

Secondly, you may wonder whether CR2 is, in contrast to CR1, too restrictive. It would seem not to allow people to have LOT terms which refer to the members of natural kinds with whose members they have never directly interacted. Think of how many concepts children acquire on the basis of pictures. If BT1 is going to rely only on the causal history of the subject to fix extension (more on this momentarily), then we have to allow that the pictures of horses that the child sees were caused by horses, and that the child's ensuing tokenings of 'horse' were caused by horses. In light of these two difficulties for CR2, my inclination is to endorse CR1, pending any demonstration that allowing the relevance of indirect causal dependencies actually causes problems for BT1.

e. Causal history and counterfactuals

To this point, I have assumed that the success rate of K relative to a subject S's LOT term t is based solely on S's actual history. There is, however, another option which seems worthy of consideration. Rather than actual causal history, we might try formulating a counterfactual-based success rate function. For example, we might identify the extension of t as the natural kind which would, under typical circumstances, have the highest success rate relative to t. The primary virtue of the counterfactual-based approach would be to free BTT from objections based on peculiar or idiosyncratic subject histories.

As we saw in chapter II, Fodor's AD theory relies heavily on claims about what our minds/brains would do in counterfactual situations in which the nomic structure of the universe is altered. Fodor's commitment to claims about what would happen to human minds in such worlds renders his theory difficult to evaluate as a theory of content for LOT terms. In contrast, the counterfactual-based interpretation of BT1 directs us to consider counterfactual situations which are relatively unproblematic; these counterfactuals do not involve changes in the nomic structure of the universe. We need only to consider counterfactual cases in which subjects have typical causal interactions with members of the natural kinds in question. For example, we would ask ourselves how likely S would be to token 'cow' were a cow to walk right past S in broad daylight.

In chapter VI, I develop a detailed counterfactual-based version of BTT in response to a specific type

of objection to the actual history version of BTT. There, by the way of developing a counterfactual version of BTT in detail, I answer some of the obvious questions regarding how a counterfactual version of BT1 might work. I also give the reader some idea as to how I think that the relevant problems for BTT might be solved without resorting to counterfactuals. Which approach to solving these problems is more satisfactory is left to the reader to decide.[xix] In the remainder of this section, I merely attempt to convey the flavor of the advantages and disadvantages of each approach.

The counterfactual-based calculation of success rates has the virtue of protecting BTT against objections of the following type. Consider a subject S who encounters but one cow and acquires what seems like a perfectly good concept of a cow (i.e., a concept of a cow which you would expect to be a pretty good test for the presence of cows). Assume also that at the time at which S acquired her cow concept, S had never been causally affected by a horse. Now add to our story the stipulation that, against all odds, immediately upon acquiring the concept cow, S goes out on a dark night and encounters two horses, one after another. In response to both horses, S mistakenly tokens 'cow'. Immediately following, S encounters an oddball cow and tokens an LOT term other than 'cow'. At this point, the success rate of cows relative to S's LOT term 'cow' is 1 divided by 2, i.e., 0.5. In contrast, the success rate of horse relative to S's LOT term 'cow' is 2 divided by 2, i.e., 1. Ergo, S's LOT term 'cow' refers to horses, even though S's LOT term 'cow' is associated with a concept which seems to be a perfectly good concept of a cow.

All of this trouble is averted if we embrace the counterfactual approach. If we were imagine running a bunch of cows and horses by S under the typical conditions of S's observation of large animals, the success rates would seem to come out right. Given that the concept in question was a legitimate cow concept from the time it was acquired, we can assume that from the very beginning, the relevant counterfactuals would lead to a higher counterfactual success rate for cows than for any other kind, including horses.

What can be said in favor of the actual history approach? First off, notice that the example described above is incredibly unrealistic. And because we're imagining in the example that S's errors were the result of pure chance, and that S's concept of a cow is a good one, the success rates should fall into line before long. Furthermore, in contrast to the consideration of counterfactuals, considering only a subject's actual causal history is comparatively simple (metaphysically and epistemologically). On the actual history view, the extension of t for S is determined by past causal effects of members of the various natural kinds on S. If there are individual quirks in S's process of acquiring and beginning to use a concept and the associated term,

then the proponent of the actual history interpretation of BT1 might have to bite the bullet and say that terms sometimes have unexpected extensions in these quirky cases (further, more palatable options are explored in chapter VI). However, since such cases are quirky to begin with, there may be reason to claim that we shouldn't worry too much about these cases anyway. We should not be disturbed by funny sounding results when we're dealing with funny sounding cases.[xx]

B. Extension-Fixing Intentions

1. The second principle

The Best Test Theory consists of a hierarchy of extension-fixing principles. The extension-fixing principle BT1 defines basic reference, reference in cases where subjects have no relevant intentions directed toward the LOT terms in question.[xxi] Cases where any extension-determining intentions are present require the application of a principle beyond BT1. Furthermore, such intentions must be built up out of terms whose extensions are fixed by BT1.[xxii] Thus, we extend BTT by adding more content-determining principles, to explain how reference is fixed in cases which involve various types of extension-fixing intentions.[xxiii]

Principle BT2 applies only in a particular kind of case, the case where a subject S intends that a specific LOT term refer to the natural kind exemplified by some sample item S has at hand (cf. Putnam 1975, chapter 12). Imagine there is some natural kind K* whose success rate relative to t is equal to or higher than the success rate of K relative to t. By procuring samples of K (say, from an expert), and intending that t refer only to the members of the same natural kind as K, the subject can insure that the members of K* will not be the extension of the t.[xxiv] Put into the form of a content-determining principle, we have:

> B̄T2- If S has had kind-minded intentions toward a natural kind term t in S's LOT, and S has had such intentions in relation to actual samples with which S has causally interacted, then the extension of t is the set of members of the natural kind which (1) is among the natural kinds exemplified by sample individuals toward which the subject has had kind-minded intentions, and (2) among those kinds toward which the subject has kind-minded intentions, is the set of members of the natural kind for which the concept associated with t provides the best test.

The preceding refers to kind-minded intentions. This category of intentions includes S's intention that t be used as a natural kind term and S's intention that t apply to all members of the natural kind exemplified

by certain samples with which S has causally interacted. Typically, kind-minded intentions also describe members of the type to which t is supposed to refer. Such a description of the sample identifies which aspects of the sample are relevant to fixing the extension of the term in question. In many cases, it may be necessary that S have certain descriptive intentions in order to achieve a determinate fixation of reference. Intending only that t refer to all of the members of the same natural kind as that exemplified by a sample often leads to an indeterminacy of reference because the sample(s) in question exemplifies more than one natural kind. By including specific information as to what properties of the sample are relevant, S can identify exactly which natural kind is the kind whose members are included in the extension of t.[xxv]

The principle BT2 allows the subject to limit the class of candidate kinds for the extension of a given term in a way that solves what Devitt and Sterelny call the 'qua problem' (Devitt and Sterelny 1987, pp. 63-65, 72-79). Devitt and Sterelny raise the problem for causal theories of reference for a public language, but the problem can also arise for a causal theory of reference for LOT terms. The qua problem results from the fact that virtually any sample with which a subject causally interacts is a sample of more than one natural kind. For example, any sample of the natural kind gold is also a sample of the natural kind chemical element, and any sample of kangaroo is also a sample of the natural kind mammal. To solve the qua problem, the subject must be able to pick out the relevant natural kind among the various natural kinds to which a given sample might belong. Devitt and Sterelny claim that in order to solve the qua problem, the speaker applies a description to the sample which the subject faces (Devitt and Sterelny 1987, p. 75). Unless such descriptions are spoken aloud (which hardly seems plausible), these descriptions would have to take the form of descriptive intentions on the part of the speaker. However, any theory which says that contentful descriptive intentions are used to fix the reference of a term must explain how these intentions, these mental states, got their content in the first place.

Enter BT1. The principle BT1 explains how the LOT terms which constitute descriptive intentions get their content in the first place.[xxvi] The principle BT2 then allows the subject to solve the qua problem by allowing the subject to direct specific intentions, already imbued with content by BT1, toward a new LOT term and the sample of the kind toward which it applies. To do so, the subject focuses on certain aspects of the sample which the subject can already represent. Say, for example, subject S wants 'kangaroo' to apply to a sample kangaroo, qua kangaroo. To narrow down the range of kinds which might thus be fixed as the extension of 'kangaroo', the subject can focus specifically on the sample kangaroo's brown coat or on the size

of the sample kangaroo's feet. So long as the subject is capable of representing the relevant characteristics of the sample, the qua problem is solved. 'Kangaroo' refers to kangaroos rather than mammals because among the natural kinds to which the sample kangaroo belongs, the subject's explicitly represented concept of the kangaroo provides the best test for kangaroos. This does not imply that being brown and having big feet are necessary for something to be in the extension of the subject's LOT term 'kangaroo'. The point here is only that as part of the subject's extension-fixing intentions, the subject explicitly represents those elements of S's concept associated with 'kangaroo' that are relevant to determining the extension of 'kangaroo'. Furthermore, the kind kangaroo has a much higher success rate relative to those elements, 'brown with big feet', than does the kind mammal.[xxvii]

The principle BT2 does not offer a specific formula for calculating success rates of composite descriptions from the success rates of the constituent terms. One approach would be to use a version of the feature-match theory of extension. Recall that the primary drawback of feature-match theories of extension is that they help themselves to the semantic content of features without explaining how the content of features is fixed. We don't face the same problem here. The contents of the relevant features have, by hypothesis, been fixed in accordance with BT1. Thus we're free to defer to the content of the features 'brown' and 'big-footed' to fix the extension of S's term 'kangaroo'. 'Brown, with big feet' is a better description of kangaroos than it is of mammals in general, and thus BT2 says that S's LOT term 'kangaroo' refers to kangaroos, not to mammals in general.

But how do we explain the idea of 'being a better description of' within the framework of BTT? The most straightforward way to do so may be to defer to the way the features in question are distributed across existing mammals and kangaroos. Kangaroos are much more likely to instantiate the featural properties in question than are mammals in general. While this formula may sound like it relies too heavily on what happens to exist in the real world, this approach should yield the right answers in all cases of central importance. If mammals were just as likely to have all of the properties listed in S's kind-minded intention (including 'big', 'brown-footed', 'having that kangaroo-ish outline', etc.), we would seem to have no grounds to claim that S's LOT term 'kangaroo' refers to kangaroos and not mammals. Imagine that S's first kangaroo was an atypical kangaroo, but which had features, f1, f2 and f3, very common to other, non-kangaroo mammals. Now imagine that S's kind-minded intentions directed toward the atypical kangaroo describe the kangaroo in terms of f1, f2, and f3. According to the proposed formula for determining extension from

conceptual components, S's term 'kangaroo' would refer to mammals and not to kangaroos. However, as the case was described, this hardly seems objectionable. There was nothing distinctively kangaroo-ish that S had in mind when grounding 'kangaroo'.

The situation gets more complicated when the relevant set of features f1, f2,...fn are features of two different types, some of which are atypical of kangaroos, but typical of mammals, and some of which are distinctive kangaroo properties. What do we say when S intends that 'kangaroo' refer to all the members of natural kind K, but then describes K, via a sample, in a way that is neither distinctive of mammals (relative to other natural kinds) nor distinctive of kangaroos (relative to other natural kinds?) I will not discuss such cases further. Because of the idiosyncratic nature of these cases, there is no clearly correct extension, by dint of either intuition or psychology, which must be assigned.[xxviii]

2. The third principle

Principle BT2 applies in cases where subjects have kind-minded intentions directed toward samples of a natural kind. Putnam focuses primarily on this kind of case in presenting his theory of reference for natural kind terms in a public language (see also Kripke 1980, lecture III). However, we might want to add a third principle to BTT, BT3, to address cases where S intends that a natural kind term of LOT apply to a certain natural kind, without S's directing this intention toward any extant sample.

> BT3- If S bears kind-minded intentions toward a natural kind term t in S's LOT, but does not direct those intentions toward any particular sample, then the extension of t consists of the members of the natural kind for which the concept associated with t provides the best test.

The principle BT3 typically applies in cases where a scientist identifies the traits which she believes K to possess, intends that an LOT term refer to the members of K, but at the same time has no samples on hand which she believes to exemplify K.[xxix]

3. Incidental intentions, standing intentions, and implicit intentions

An intention is an incidental intention if it is explicitly (though not necessarily consciously) represented by S in the very circumstances in which the intention plays a content-determining role. The

preceding discussion of BT2 and BT3 may give the impression that these principles apply only when subjects have incidental intentions of the right sort. However, it's quite possible that the content of some terms is fixed via S's use of non-incidental intentions which we might call 'standing intentions'. Consider the budding scientist who is taking classes and continually acquiring new concepts and new LOT terms which are supposed to refer to natural kinds. With respect to each new LOT term that is coined, the student does not, in all probability, think to herself, "I intend that this term apply to a natural kind."

In order for BT2 or BT3 to apply in the absence of incidental intentions, we need the notion of a standing intention, an intention that is not explicitly tokened each time its extension-fixing influence is felt, but is psychologically real all the same. The introduction of standing intentions at this point is not an ad hoc affair. Standing intentions help explain a number of other phenomena. An example of such a phenomenon is the routine performance of self-protective actions. Agents continuously perform actions the effects of which are to reduce the risk of harm to themselves. For example, whenever I'm working in the kitchen and I see a large knife near the edge of a counter, I move the knife to a safer location. In most cases, I do not possess a conscious awareness of any intention to protect myself from bodily harm. Yet such an intention seems operative. In fact, it was just such an intention that initiated the first instance of my intentionally moving a knife away from the edge of the counter.

Similarly, it would be reasonable to consider numerous actions you perform in the course of driving an automobile to be actions motivated by standing intentions to protect yourself. For example, every time you check your mirrors, you seem to be acting to protect yourself from harm. But you do not typically think to yourself, each time you check the mirrors, "I don't want to get hurt; I had better check my mirrors." An obvious way to explain your behavior is by referring to a standing intention, i.e., your intention to protect yourself from harm.

We may be able to explain away the appearance of standing intentions by substituting in their place subconscious, incidental intentions. In order to avoid a subject having too many standing intentions for them all to be explicitly represented at the subconscious level in an ongoing fashion, we might propose a scheme whereby such intentions as are relevant to a given situation are reconstructed at the subconscious level.

The idea of a standing intentions has some appeal, however, especially when one thinks of cases where the subject is not likely to have ever consciously tokened the intention in question. In such cases (as well as in others), philosophers sometimes appeal to the existence of implicit states. In chapter IV, I take up a particularly relevant example, that of children's classification of LOT terms as natural kind terms, which

raises questions about the nature of implicit intentions. There I review specific empirical data which support the claim that children are disposed (perhaps innately) to treat many of their concepts as concepts of natural kinds. These data raise difficult questions about the role of what appear to be implicit intentions in fixing extension in accordance with BT2 and BT3. Consequently, much of chapter IV consists of an effort to identify the role of content-fixing intentions which appear to be implicit, together with an effort to understand the range of application of BTT's various principles in the context of the child's implicitly intending to treat certain LOT terms as natural kind terms.

C. Extending the Best Test Theory

Before moving to chapter IV, it is worth stopping to briefly explore the open-endedness of BTT. BT1, BT2, and BT3 only apply to natural kind terms of LOT. However, if we assume that the most basic type of reference is to natural kinds, BTT can be extended by adding principles which apply to other types of kind terms in LOT.

One case to which BTT might be extended is that of artifactual kind terms. As an example of an artifactual kind concept, consider the concept hammer. Hammers have a typical shape and size. They also have a characteristic function. Consequently, the concept hammer would seem likely to be a complex specification of a hammer's typical physical characteristics and of a hammer's typical function. The way these factors interact to yield determinate reference is a matter which deserves more attention than I give it here. However, in order to specify the function of a hammer or identify the typical physical traits of a hammer, the subject must possess the cognitive tools to describe those functions or traits. By assuming that reference begins with reference to natural kinds, we assume that the terms that constitute the subject's descriptive intentions directed toward hammers and 'hammer' are themselves natural kind terms (or if not, that their reference was fixed via intentions built out of natural kind terms, and so on). Thus, any extension of BTT to artifactual kind terms would assume the previous application of BT1, BT2, or BT3 (or BT4, see chapter VI) to the terms composing the intentions used to ground the reference of artifactual kind terms (or such principles would have to have been applied to the terms which compose the intentions directed toward the terms which make up the intentions directed toward the artifactual kind terms, and so on). In this way, the extension of BTT to artifactual kind terms would be an extension of BTT, not an independent treatment of artifactual kind terms.[xxx]

Notes to Chapter III

---

[i] When we say that concept C provides the best test for Ks, we do not necessarily mean that C provides anything like an ideal test for Ks, or that C provides a test for Ks which is better than any other possible test for Ks. What we mean is that out of all the natural kinds the presence of members of which C can be a causal indication, C provides a better test for members of K than for the members of any other natural kind.

[ii] The probabilistic view described in the text is, technically speaking, only one variation on the probabilistic approach, the one Smith and Medin call the 'featural approach' (Smith and Medin 1981, chapter 4). The featural approach is the most commonly discussed, and the most general, of the various versions of the probabilistic approach. This justifies the emphasis in the text on the feature-based approach.

[iii] Rey 1983 (footnote 2, pp. 256-257) argues that this characterization of classical concepts, as lists of singly necessary and jointly sufficient conditions, is incorrect. Rey gives the example of natural number as a concept which is supposed to be classical, but which does not consist of a list of singly necessary and jointly sufficient features. According to Rey the concept of a natural number is disjunctive, and thus defies the requirement that it provide singly necessary and jointly sufficient conditions for category membership. This is because the concept of a natural number is the concept of something which is 0 or is a number which is the successor of a natural number. While Rey is correct in so far as he has given a mathematically adequate definition of 'natural number', it is doubtful that the definition Rey gives is the concept that the average person understands when she thinks or talks about natural numbers. The concept of something which is a non-negative, whole number seems to be a classical concept of natural number which is much closer to the average person's understanding of natural numbers. This is in spite of the fact that there may be some sense in which Rey's disjunctive definition is as good as, or even better than, the average person's concept for the purposes of systematizing mathematics. The disagreement here about what makes a concept classical may result from the gap between the view of concepts as metaphysical entities and the view which instead focuses on what is psychologically real, or active, in the subject.

[iv] The feature-match analogue to BTT, and the feature-match approach in general, are presented in contrast to BTT. They are not intended to be a part of BTT. Later in present section, I explain why we should not accept a feature-match theory as a theory of extension for natural kind terms of LOT.

[v] I am not here endorsing Rosch's views about basic-level concepts and categorization, nor am I claiming that Rosch explicitly held a feature-match theory of extension. Rosch's comments are merely suggestive of the way in which many discussions of categorization seem to assume that a concept represents a natural kind by matching the features of the members of the kind which the concept represents.

[vi] It is interesting to note that such features as are basic in theories like Marr's or Biederman's are not like the features included in Smith and Medin's concept of a Bird. Being similar in shape to a certain collection of generalized cones would not seem to be a feature which is lexical or consciously accessible to the average subject.

[vii] In the interest of balance, you might instead imagine yourself to be a dogmatic hater of conservatives. Now consider movies starring Chuck Norris. The feature 'has seen numerous movies starring Chuck Norris' may have diagnostic value for you as an element of your concept Conservative.

[viii] Slote 1966 develops what he calls the 'Theory of Important Criteria' which draws a distinction similar to the distinction between a concept's core and the related identification procedures. Applying his theory to natural language terms, Slote distinguishes between two aspects of a term's meaning, those aspects which express something about the members of the extension of the term which is important to these things being

what they are, and those aspects of a term's meaning which are not important in this way. Also relevant here is Putnam's discussion of cluster concepts in Putnam 1975, chapter 2. See Quine 1953, essay II, for arguments regarding the nature of meaning for natural language terms which would, if applied to concepts, seem to challenge to the tenability of an absolute distinction between the core features and identification procedures of concepts.

[ix] This is true at the most basic level at which extensions are fixed, the level at which extensions are fixed atomically. At higher levels, there may exist conceptual elements which we might consider semantically individuated, for the following reason. Once some natural kind terms of LOT have had their extension fixed at the atomic level, these terms can be employed by the subject to fix the extension of further natural kind terms of LOT. The natural kind terms of LOT which are used by the subject to fix the extension of further LOT terms could thus be considered semantically individuated features of the concept associated with these further LOT terms whose extension is fixed non-atomically. The question of levels will be taken up in more detail below.

[x] To students of Jerry Fodor's work, the viewpoint presented in the text may seem familiar. Although not explicitly addressing the issue of concept structure or constitution, Fodor takes a similar viewpoint with respect to concepts in presenting his causal theory of content. As Fodor puts it, "It's the existence of reliable mind/world correlation that counts, not the mechanisms by which that correlation is effected." (Fodor 1987, p. 122) My view is, of course, less extreme than his. I claim that if mental states with content play a certain type of role in fixing content (i.e., one in which the content of the state helps fix or revise the content of an LOT term), then we can't dismiss the presence of these states as mere mindless mechanisms at work.

[xi] At the level of BT1's application, some concept descriptions may be easier to give than Fodor's remarks on robustness would lead one to believe. For example, early in cognitive development, there is much less inference involved in categorization than there is in adulthood. The collection of mechanisms and the relations between them which causally mediate the tokening of a given LOT term in the infant or the young child may thus be much more manageable in their size and complexity than is the collection of such mechanisms and their relations in the adult.

[xii] There may be some disagreement as to whether terms such as 'object' (and 'brown' and 'big-footed', to be discussed below) are natural kind terms (and as to whether their extensions are natural kinds). However, for the purposes of developing a semantics of natural kind terms, there seems to be little reason to separate natural properties (such as 'being an object') from natural kinds. (This seems to be Kripke's view. See Kripke 1980, p. 134.) Arguably, to be a member of a natural kind is just to have a certain natural property, and vice versa. Consequently, I treat terms denoting natural properties as natural kind terms.

[xiii] It may be that at later stages in the development of the object concept, BT1 no longer applies. For example, when, as an adult, a physics student studies phase transitions, the student is quite likely to bear detailed intentions toward 'object'. Such intentions may refine or otherwise change the reference of 'object'. At the very least, the presence of these intentions would imply that BT1 no longer is the relevant extension-fixing principle for 'object'.

[xiv] See Spelke 1990, 1991, and Bower 1989 for descriptions experimental results which reveal the infant's possession of a strikingly rich object concept. In light of such results, it is reasonable to think that if BT1 applies to the infant's LOT term 'object', then BT1 assigns the class of objects as the extension of the infant's LOT term 'object'.

[xv] Many of the mental structures in which cognitive scientists are interested, e.g., representations of rules in phonology, appear at the subconscious level. Ray Jackendoff is an example of a cognitive scientist who studies such structures, but who does not think that it is theoretically fruitful to attribute extensions to them (Jackendoff 1989, p. 76).

[xvi] For an argument that the lower-level structures in Marr's theory of vision, perhaps even the elements of the 2 1/2-D sketch, are not representations with determinate extensions, see Montgomery 1989. For opposing views, see Shapiro 1993 and Hatfield 1991.

[xvii] Note that CR1 allows for group causation of the tokening of an LOT term. For the purpose of calculating success rates, when a group of members of kind K cause the tokening of t, this is treated as an instance of a member of K causing the tokening of t.

[xviii] Granted it would sound odd to say that 'gross' is a natural kind or natural property term. This is of no consequence here, however, for the point of the example is one about indirect causal chains. Nothing turns on 'gross''s being or not being a natural kind term.

[xix] I leave the decision to the reader not out of pure laziness. Instead I believe it is difficult to decide now what type of solution, in service of prediction and explanation, is actually required for the problems detailed in chapter VI. Both the counterfactual and the actual history version of BTT appropriately fix extensions in typical cases. In contrast, atypical cases cause the problems discussed in chapter VI; and it is difficult to gauge the success of either version of BTT in these cases precisely because of their atypical nature.

[xx] Perhaps we could protect the actual history interpretation of BT1 from counterintuitive results in quirky cases by modifying the actual history interpretation slightly, in order to discount cases where breakdowns in a subject's physical systems (e.g., sensory systems) are involved. Of course this works only when such breakdowns can be specified non-circularly.

[xxi] In the text I assume that we have in hand a theoretical characterization of intentions. Such a description is necessary in order to give some meaning to the distinction between BT1 and BT2. We might advert to a state's possession (or lack of possession) of certain functional characteristics in order to define what it is to be an intention. For example, intentions may have a characteristic role in the causation of actions. (Fodor takes this approach to defining mental state types [cf. Fodor 1987, p. 69].) The Best Test Theory could appeal to this type of functional description of what it is to be an intention when distinguishing between LOT terms which are the objects of extension-determining intentions and those which are not. However, such a functional characterization of intentions should not include the requirement that for a state to be an intention, it must play a role in fixing the content of an LOT term, because such a definition of intentions would seem to lead to circularity in characterizing the distinction between situations in which BT1 applies and those in which BT2 applies. Once an independent characterization is given of what it is to be an intention, extension-fixing intentions emerge as intentions with certain extension-fixing contents; in order for the state to play a content-fixing role, it must already be an intention. Thus, we don't want the question of whether or not a particular state is an intention to turn on whether it has actually played or is currently playing a content fixing role. Nevertheless, we may still admit a general characterization of intentions which says that they must be capable of playing a certain type of extension-fixing role. Still, the definition of this role should not rest on any distinction between BT1's and BT2's application.

[xxii] If an extension-determining intention is not built entirely out of terms whose content is fixed in accordance with BT1, then the terms whose content is not fixed by BT1 must themselves have had their content fixed by intentions whose component terms had their content fixed by BT1. And so on. There is a further possibility that the terms appearing in an extension-fixing intention would have intensional, but not extensional, content, and that this intensional content would be operative in fixing the extension of the term toward which the intention in question is directed. Such a possibility is considered in chapter IV in a slightly different context.

[xxiii] Throughout this section, I assume the accuracy of some combinatorial theory of semantics for LOT. I'm not committed to the accuracy of any particular combinatorial story, but the current discussion assumes that terms which have their content determined by BT1 can somehow be combined to construct intentions of the proper sort.

[xxiv] It should be noted that the phrase 'members of kind K' is meant to be interpreted loosely so that it applies to natural kinds like gold which don't have 'members' as they're normally conceived of. The members of the natural kind gold should be thought of as all of the existing samples of gold. There may be an important theoretical difference between biological natural kinds and other natural kinds like gold or electron. The question of the relevance of such a distinction to the plausibility of BTT's principles will, however, be put off until chapter VI.

[xxv] In this way, the kind-minded intentions of BT2 play a role in grounding natural kind terms of LOT that is similar to role played by Timothy Maudlin's explanandum profiles (Maudlin 1986, pp. 48-50) in grounding natural kind terms in a public language. According to Maudlin, a speaker uses an explanandum profile to ground a natural kind term by specifying which properties of a sample are to be explained by membership in the natural kind to which the speaker wants the term in question to refer. Maudlin's view is incomplete in that "the explanandum profile clearly does depend on the beliefs, intensions, observational capacities, &c., of the linguistic community." (Maudlin 1986, p. 49) The current project can thus be seen as an attempt to fill out Maudlin's story by (partly) explaining how the beliefs of the members of the linguistic community get their contents in the first place; for it seems that these beliefs must have their content already fixed in order to specify an explanandum profile for a public language term.

[xxvi] In chapter VI, it is shown how BT1 alone solves the qua problem for the terms to which BT1 applies. I am at pains here in the text only to show how the contents of extension-fixing intentions are determinate to solve the qua problem for those terms to which BT2 applies. This is necessary because once intentions take over the extension-fixing process, we have to make sure that the qua problem can be avoided.

[xxvii] Sterelny 1990 (pp. 134-140) suggests a solution to the qua problem that is in some ways similar to the solution offered in the text. Sterelny's solution is similar to BTT's in that the subject uses non-defining features of the sample to effect determinate reference in some cases. Sterelny's approach is different from mine in that he claims that in the case of basic reference, teleology, rather than a principle such as BT1, fixes reference. Another way to solve the qua problem is to allow the speaker or thinker to make some reference to experts who have greater discriminatory powers than the speakers or subjects have themselves. The deferential use of experts to fix reference, an idea originally due to Putnam (Putnam 1975, pp. 227-229), is examined in more detail in chapter VI.

[xxviii] Another approach is to simply advert to S's history. For simplicity's sake, we could take an average of the success rates of mammal relative to 'brown' (for S) and mammal relative to 'big-footed' (for S). Presumably this success rate would be much lower than the average of the success rate of kangaroos relative to 'brown' (for S) and the success rate of kangaroos relative to 'big-footed' (for S). (In the name of accuracy, we may want to complicate matters by weighting the various features in the calculation of the average. We could do so according to weightings expressed by S's intentions, if there are such weightings expressed, or the weightings could be determined in proportion to the number of causal interactions on which each of the component success rates of the average are based.) While this approach seems promising, difficulties may result from its heavy reliance on S's particular history. Think again of the idiosyncratic learning history which motivated the consideration of a counterfactual-based version of BT1 in section 2.e of the present chapter. (More on such matters in chapter VI.)

[xxix] For a case of theoretical specification taken from the history of natural languages, see Hacking 1983, 87-90, where Hacking describes the circumstances surrounding the introduction of the term 'meson' into the terminology of physics. 'Meson' took the meaning of "whatever it is that satisfies Yukawa's conjecture" (Hacking 1983, p. 90), where Yukawa's conjecture is, in essence, a description of a particle unobserved at the time of his conjecture.

[xxx] This brief discussion of artifactual kind terms raises an interesting question regarding the persistence and flexibility of concepts. Concepts have been shown to be highly context sensitive in some regards (see Goschke and Koppelberg 1991, pp. 140-145, for a brief review). It may be that LOT terms have, in some sense, a flexible reference determined by the intentions of a subject **at the time of use**. This would seem to be especially likely in the case of artifactual kind terms. Thus, it may be that on some occasions 'hammer' would refer broadly to anything which would serve a particular hammering purpose, but on other occasions would refer only to items which fit the description of a typical hammer. Much connectionist theorizing is intended to explain just how such flexibility of meaning is possible (cf. Clark 1989, pp. 113, 191). If, however, the explanation of this flexibility requires that the subject have some intentions toward the concepts or terms involved, then these intentions must first have their content determined in some way. Thus might the fixation of meaning via BT1, BT2, or BT3 precede, and be required for, the flexible, context-sensitive determination of meaning for artifactual kind terms of LOT.

IV. THE BEST TEST THEORY AND NATURAL KIND TERMS

The Best Test Theory rests on the assumption that there is a well-defined class of natural kind terms in LOT. In this chapter, I argue that this assumption is justified based on data collected by developmental psychologists. To this end, the first part of the chapter reviews some of the relevant developmental data. The discussion of developmental data also provides an opportunity to clarify the range of application of BT1 and BT2. Bearing the developmental data in mind, I close the chapter with an extended examination of the idea of an implicit intention (or state or rule) in an attempt to make more precise how an intention with content can help to fix the extension for a new natural kind term in LOT.

A. Natural Kinds and Categorization in Children

As part of a general attempt to understand the development of children's abilities to represent categories, Frank Keil has produced some striking results regarding the way children categorize objects using natural kind terms. In what follows, I review two of Keil's studies, both of which seem to show that from an early age, children treat natural kind terms differently than they treat some other types of terms. This is striking in that the differential treatment of natural kind terms occurs at an age where children typically have not acquired any detailed scientific knowledge and do not have any sophisticated understanding of the underlying characteristics which give natural kinds their unity.

The first study, the discovery study, is an investigation of children's categorization judgements in cases where items which appear to be As are discovered to have the internal structure of Bs (e.g., Bs' blood or Bs' bones). Keil told groups of 5-, 7-, and 10-year-olds (as well as a group of adult subjects) sets of stories like the following:

> There are animals that live on a farm. They go "neigh" and people put saddles on their backs and ride them, and these animals like to eat oats and hay and everybody calls them horses. But some scientists went up to this farm and decided to study them really carefully. They did blood tests and X-rays and looked way deep inside with microscopes and found out these animals weren't like most horses. These animals had the inside parts of cows. They had the blood of cows, the bones of cows; and when they looked to see where they came from, they found out their parents were cows. And, when they had

> babies, their babies were cows. What do you think these animals really are: horses or cows? (Keil 1989, p. 162)

For contrast to the case of natural kinds, Keil mixed in similar discovery stories about pairs of artifactual kinds (e.g., key/penny or boot/sail).

At all ages, the subjects judged the scientists' discoveries to be irrelevant in the case of artifacts. The situation was quite different when the stories were about natural kinds. As age increased, the subjects showed a greater and greater tendency to judge the discoveries about the internal properties of natural kind members to be relevant to the determination of kind membership. Kindergartners showed only a slightly greater tendency to judge the discoveries relevant in the case of natural kinds. But by second grade, the subjects were far more likely to judge the discoveries to be relevant in the case of natural kinds than they were in the case of artifactual kinds. This provides preliminary evidence that young subjects treat natural kind concepts differently than artifactual kind concepts. Furthermore, assuming that LOT terms are distinct from the concepts with which they are associated, we can infer that children treat the LOT terms associated with natural kind concepts differently than they treat LOT terms associated with artifactual kind concepts.

In a follow-up study, Keil tested the effects of surface transformations on subjects' judgements about kind membership (Keil 1989, pp. 183-193). Keil told stories to his subjects (groups of kindergartners, second-graders, and fourth-graders) which involved surgeons/scientists starting out with something which was identified as an A, and then through surface doctoring, transforming the item into something with the appearance of a B. Stories about both natural kinds and artifacts were included in the stimulus set.

The results of the transformation study were similar to the results of the discovery study. All subject groups regularly responded that an external transformation was enough to make an artifact change types. Kindergartners were only moderately resistant to the idea that an external transformation changed the kind of a natural kind member. But by fourth grade, the subjects almost always judged that the surface transformation was not sufficient to change the kind type of a natural kind member. Again Keil's results seem to show that from a fairly young age, ca. 7 or 8 years of age, subjects conceive of natural kinds differently than artifactual kinds. In turn, this provides some grounds for thinking that the LOT terms associated with natural kind concepts form a well-defined, identifiable category, distinct from other types of LOT terms.[i]

While Keil's work is provocative in many ways, his results would be more pertinent to our concerns

had he shown that his younger subjects (the five-year-olds) treat natural kind concepts as an homogeneous class of LOT terms. It seems reasonable to think that the extensions of a child's natural kind terms play some explanatory role, and thus should be fixed, before age 7 or 8 (i.e., before the age at which Keil's subjects consistently treat natural kind terms differently than artifactual kind terms in his experimental paradigms). If BTT is going to fix the content of the natural kind terms for these younger children, we should want some evidence that the younger children treat natural kind terms of LOT in a distinctive way.

Susan Gelman and Ellen Markman provide some evidence that children treat natural kind terms of LOT distinctively as early as age 4 or 5. Gelman and Markman's experiments test the way children make inferences when given information regarding category membership. They found that, even at preschool age, their subjects seemed to understand the special implications of shared category membership for kinds labeled by natural kind terms.

In the first experiment from Gelman and Markman 1986 (as reported in Markman 1989), Gelman and Markman investigated the degree to which preschoolers would base inferences about the possession of unknown properties on category membership. Gelman and Markman presented preschool-aged subjects (four-year-olds with a mean age of 4:5) with sets of pictures, three different items to each set. Two of the items were given category names and a piece of contrasting information was given to the subjects regarding these two items. The third item in each set of pictures looked like only one of the first two items pictured but shared category membership with the other item (i.e., the third item shared category membership with the item which the third item did not resemble perceptually). The experimenter then asked the subject to decide which of the two other items the third item was similar to with respect to the property originally used to contrast the first two items. For example, a tropical fish was called a fish and the subject was told that the fish breathes underwater. Additionally a dolphin was called a dolphin and the subject was told that the dolphin has to rise to the surface to breathe. Along with the tropical fish and the dolphin, the subject was presented with a picture of a shark. The subject was then asked whether the shark (referred to as a 'fish') breathes underwater or whether it has to come to the surface to breathe (Markman 1989, pp. 95-100).

Sixty-eight percent of the time, Gelman and Markman's subjects used common category membership as the basis of their decisions, rather than relying on perceptual similarity. So, for example, with respect to the tropical fish/dolphin/shark stimulus set, the children were substantially more likely to say that the shark breathes underwater. This is despite the fact that the shark resembles the dolphin, rather than the tropical fish, at the perceptual level.

The same pattern of responses held when Gelman and Markman controlled for the influence of linguistic repetition on the subjects. Using synonyms to name the category type of the two items which shared a category (e.g., rabbit/bunny), Gelman and Markman ran the same experiment with similar results. In this follow-up study, children based their decision regarding the attribution of a new property to a third item based on shared category membership (as indicated using synonyms), rather than perceptual similarity, 63% of the time, still significantly above a chance response (Markman 1989, pp. 100-101).

Gelman and Markman 1986 focuses specifically on how children draw inferences when natural kind categories are involved. These results are important in that they provide evidence that young children are treating natural kind categories in the way you would expect adults to. The results of these studies would be even more important, for present purposes, were they to contrast the children's inferential behavior for natural kinds with that for other kinds. Such a comparison could determine whether children are relying on common category membership alone as a general strategy for drawing inferences. Inference regarding artifactual kinds provides one type of case where we would expect children to draw inferences differently than they do in the case of natural kinds. In the case of artifactual kinds, we would expect the children to rely on perceptual similarity, rather than on information about common category membership, to guide their inferences.

Gelman 1984 (as reported in Markman 1989, p. 108) provides evidence that children aren't focusing solely on common category membership as a general strategy for drawing inferences regarding newly encountered items. To groups of four-year-olds and groups of seven-year-olds, Gelman taught new facts regarding natural and artifactual kinds. The subjects were then asked whether the fact they were taught about the member of the natural or artifactual kind in question is also true of various other items. Some of these items bore a perceptual appearance to the teaching item, some did not. In addition some were of the same natural kind and resembled the teaching item, while some were of the same natural kind (e.g., a superordinate kind) but did not resemble the teaching item. In this experiment, the seven-year-olds drew significantly more inferences (claiming that the new fact would also be true of a new item) in the cases where the teaching item was a member of a natural kind than when the item was a member of an artifactual kind. In the cases where the teaching item was identified as the member of a natural kind, the seven-year-olds were more likely to extend their application of the new fact to members of the same natural kind as the teaching item even when the new items did not bear a high degree of perceptual similarity to the teaching item. Regarding the members of artifactual kinds, the seven-year-olds were likely to limit the

application of the new fact only to items which shared the appearance of the teaching item. While these results were not achieved with the four-year-olds, the overall results strongly suggest that by seven-years-old, children believe that natural kind categories have an underlying coherence that artifactual kind categories do not have.

The results from Gelman and Markman 1986 and Gelman 1984 provide substantial support, as Keil's results do, for the claim that children treat natural kind concepts in a distinctive manner in cognition. At the very least the natural kind concepts are treated differently than one of the other important types of concepts with which young children are likely to be familiar, artifactual kind concepts. This strongly suggests that there is a distinct, cohesive category of LOT terms, natural kind terms. Neither Keil's work nor that of Gelman and Markman demonstrates the child's consistent, distinctive treatment of natural kind terms at as young an age as we might hope it would. However, there is reason to believe that certain, very general, natural kind terms are treated as such from infancy. In section IV.C, we'll review evidence that infants begin treat 'object' as referring to kinds with an underlying unity characteristic of natural kinds. Such data suggest that the category 'natural kind term' exists in LOT from a very early age, and is extended to more and more terms as the child develops and acquires new terms.

B. Types of Terms in the Language of Thought: An Interlude

1. The strong Language of Thought Hypothesis

In the preceding, it was assumed that if a set of concepts are, pre-theoretically, treated as a distinct, coherent kind of concepts, then that set of concepts corresponds to a distinct, coherent category of LOT terms. This assumption raises important questions regarding the nature and structure of LOT term categories. In what follows, I attempt to clarify the nature of the LOT term category Natural Kind in contrast to other LOT term categories identified in the literature.

BTT assumes that there are certain terms in LOT which have natural kinds as their semantic value (or reference). But what precisely must be true about LOT, and the cognitive processes in which LOT terms play a role, in order for us to identify natural kind terms of LOT as a distinct, cohesive category?

A number of theses are associated with the LOT hypothesis (cf. Fodor and Pylyshyn 1988, pp. 12-14). Among them are the following:

LOT1. LOT is divided syntactic categories roughly analogous to those employed by standard grammars for

natural languages.

LOT2. LOT contains recursive formation rules which define what it is for an LOT formula to be well-formed by reference to the syntactic categories of LOT.

LOT3. We can best understand cognitive processes by seeing them as formal processes which exploit LOT's combinatorial structure, as described by the recursive formation rules. (These formal processes consist of sets of algorithms which operate partly by treating terms from the various syntactic categories of LOT differently.)

LOT4. The cognitive processes described in LOT3 are not sensitive to the semantic content of the symbols over which the computations are carried out. The processes in LOT3 have their status as cognitive processes because the formal rules which define the processes cause the processes to respect or mimic certain semantic relations.

The distinction between LOT terms which refer to natural kinds and those which do not is a semantic distinction. In contrast, LOT1-4 suggest that we need distinguish between LOT term categories on a syntactic basis only. This seems to put the LOT hypothesis as it's standardly construed at odds with the idea that natural kind terms of LOT form a class of terms treated distinctively in cognitive processing. To resolve this apparent conflict between BTT and the Strong LOT Hypothesis, I begin by briefly reviewing some of the virtues of the syntactic approach expressed in LOT1-LOT4.

Consider first the role syntactic categories play in linguistics. In linguistics it is assumed that the category to which a term belongs constrains the possible combinations in which terms can legitimately appear. Linguists codify these constraints for particular languages by stating formation rules for the language. For instance, English contains the formation rule which says that a one-word adjectival phrase precedes the noun it modifies. Such rules have the function of delimiting the category of word strings which count as legitimate sentences of the language. Without the separation of the lexicon of a language into syntactic categories, we cannot specify the rules defining well-formedness for the language.[ii]

The linguist's assignment of terms to syntactic categories serves to explain public language users'

intuitions about well-formedness and users' general abilities to communicate effectively. It is not written in stone anywhere that each complete English sentence must contain a verb. However, English speakers have great difficulty communicating with each other using verbless sentences, and they will most likely reject verbless sentences as ill-formed (especially when given without a context). By containing syntactic categories and rules for building sentences that are stated in terms of these syntactic categories, a language provides guidelines to insure that users will utter and write sentences which have meaning to other members in the linguistic community. The determinate semantic values of sentences depend on the existence of well-defined syntactic categories which can be referred to in the statement of rules for forming legitimate sentences and in the statement of rules for interpreting those sentences (cf. Pylyshyn 1984, p. 69, Dowty, Wall, and Peters 1981, p. 16).

In the study of psycholinguistics, one finds the syntactic categories of public language reflected in the categories of terms in LOT. In order to explain the typical speaker's linguistic skills, psycholinguists postulate a generative (or transformational) grammar. This generative grammar works at the cognitive level to generate, and facilitate the understanding of, public language sentences (cf. Devitt and Sterelny 1987, chapter 6, Stillings, et. al. 1987, pp. 239-261). One of the primary functions of the generative grammar is to translate what is known as the deep structure of a statement into its concrete expression as a well-formed linguistic entity, and one of the primary theoretical tools in generative grammar for explaining how these transformations take place is the phrase-structure tree. These trees are used to graphically represent the rule-governed transformations that lead from a deep structure to the surface structure of a sentence (the surface structure being how the sentence would appear were it spoken or written out). Phrase-structure trees branch out from nodes labeled with the names of syntactic categories of standard grammar for natural languages. At the bottom of each branch of a tree appears a morpheme belonging to one such syntactic category. Thus, in order for generative grammar to explain language processing in the human mind, it is assumed that LOT divides into categories roughly analogous to the syntactic categories of public language.[iii]

Work in other areas of cognitive science assumes that LOT divides into categories roughly analogous to the syntactic categories of public language. Consider a typical strategy for the use of artificial intelligence (or AI) to explain human behavior. Attempting to specify the psychological processes which underlie human behavior, AI theorists write programs in computer languages which simulate the outputs of

the processes in question, so that a computer running the program computes the same function as does the subject whose behavior we're attempting to explain (Pylyshyn 1984, chpts. 3 and 4).[iv] If a computer program is taken as a literal simulation of a cognitive process, then we must assume that LOT contains any distinctions between categories of terms which are assumed by the program as it's written (This approach is known as the 'strong AI' approach [Searle 1980].) Such programs are sometimes written in such a way that the programs' categorizations of possible inputs, and the methods for generating outputs as responses to inputs, depend on the categorization of input symbols into categories roughly analogous to syntactic categories of natural languages. However, whether or not a program makes distinctions which reflect distinctions between natural language syntactic categories depends partly on what the program is supposed to do. For example, a chess playing program may, in order to separate meaningful inputs from those which are not, treat names for pieces as a category of symbols distinct from names for other objects in the chess playing environment, e.g., those that identify squares on the chess board. When the program is written in this way, it can easily recognize commands like 'five queen pawn to six queen bishop' as meaningless. The program can do this because it separates the input lexicon into categories and, further, states rules defining meaningful input strings in terms of those categories. At least one of these categories, the category chess piece name, is a category whose members share not a distinctive syntactic role (they are nouns just like 'square' is a noun), but share a semantic function. They all refer to types of playing pieces.

An analogous point holds with respect to the natural kind terms of LOT. Whether or not a cognitive system is taken to distinguish between two different types of LOT terms depends almost entirely on what sort of behavior you are trying to explain. If the best explanation of Keil's and Gelman and Markman's data is that children apply different rules of reasoning to natural kind terms than they do to artifactual kind terms, then we should assume that the child's LOT is divided into the appropriate categories. In computational terms, we can think of the child as having one set of functions which has natural kind terms as its domain and another, partially intersecting set of functions which has artifactual kind terms as its domain. In order to apply these functions differentially, the child must have the natural kind terms and the artifactual kind terms separated into two categories.

The distinction between natural kind terms and artifactual kind terms of LOT is consistent with theses LOT1-4 so long as we make one systematic change in our statement of these theses. We need only give up the idea that all of the primitive distinctions between LOT terms types are those marked in standard

grammatical theories of natural languages. Such a move is reasonable and independently motivated (think again of the chess program example). Thus, instead of talking in LOT1 about 'syntactic categories analogous to those given in standard grammars for natural languages', we should substitute 'whatever syntactically primitive term categories are useful in explaining the behavior of the cognitive system'.[v]

2. The Best Test Theory, the language of thought, and connectionist models

Cognitive processing appears to be sensitive to distinctions between categories of LOT terms, which distinctions are analogous to those between different syntactic categories in public languages. However, some connectionist models of cognition challenge this assumption. Some such models seem to leave little place for the idea of an LOT divided into syntactic categories, where, for example, the well-formedness of a sentence in LOT is defined by reference to these syntactic categories, or where the meaning of a sentence in LOT is computed from the meanings of its component terms according to the type of formal rules described by LOT3.[vi] On the other hand, it has been suggested that if we apply the proper mathematical techniques, we may find in connectionist models at least some of the syntactic structure which the LOT hypothesis claims to exist (Davies, 1991, p. 254, Clark 1989, pp. 191-194).

In the explication of BTT, it would be best to leave open questions regarding how connectionist symbols are individuated and what precisely their roles are in cognitive processing. BTT assumes that LOT term content contributes something to the content of the states in which those terms appear as components. However, BTT is not committed to any particular story about conceptual combination. The biggest connectionist challenge to BTT (which I do not attempt to resolve here) comes in the form of connectionist claims that connectionist representations are radically context-sensitive. If this is true, then it may make little sense to talk, as BTT does, about the tokening of a given LOT term on various occasions.

C. The Developmental Data and the Best Test Theory

1. Keil on the development of causal theories

The data reviewed in section IV.A are developmental data in the sense that the experiments involved testing the way subjects categorize and reason with natural kind concepts as they progress through their childhood years. This data seems to show that natural kind terms constitute a well-defined, functionally homogeneous class of LOT terms. At this point, I would like to turn to a discussion of how the

developmental data might be useful in helping us clarify the range of application of BT1 and of BT2.

The principle BT1 is distinctive among the BTT's set of content-determining principles in that in the cases where BT1 applies, the content of natural kind terms in LOT is determined independently of the subject's possession of explicit, contentful intentions directed toward those terms. The best test theory assumes that BT1 has some actual range of application, i.e., that there are times when at least some of a subject's meaningful natural kind terms in LOT had their content fixed independently of the subject's explicit intentions. In a typical subject's development, this would seem most likely to occur during childhood. As people grow older, it is likely that more and more of their natural kind terms have their content influenced in some way by explicit beliefs about those kinds. Therefore, in trying to show that there is a class of natural kind terms to which BT1 applies, it seems natural to look at the developmental data. We should not, however, oversimplify our interpretation of the developmental literature. In particular, we should be as clear as possible as to when explicit intentions toward natural kind terms appear in the child. We should also get as clear a picture as possible regarding what, if any, relevant activity takes place in the child's mind before explicit intentions appear.

A recurring theme in Keil 1989 is that the child's treatment of some terms as natural kind terms develops in a domain-specific fashion (Keil 1989, pp. 20, 23, 83, 267). In claiming this, Keil speaks against a long tradition of viewing changes in the child's representational capacities as global changes, i.e., changes in the child's general representational capacities. Many historically important psychologists, such as Werner, Vygotsky, and Piaget and his associates, claimed that children go through a (or a series of) global shifts in the way they organize knowledge or use concepts (Carey 1990, Keil 1989, chapter 2). In contrast, Keil's data seem to show that children shift to a more adult perspective one conceptual domain at a time as knowledge in that domain increases. In the case of natural kinds, Keil claims that the development of knowledge in each domain amounts to the construction of a causal theory in that domain. As the child builds a causal theory for a given natural domain, the child begins to treat her concepts related to that aspect of the world as theoretical concepts, rather than concepts which group objects together purely on the basis of perceptual similarity.[vii]

As Keil describes the child's emerging theories (Keil 1989, pp. 272-275), the essential element of these theories is the existence of a pattern of causal beliefs. Whence these patterns of causal beliefs, and how do they guide the development of full-blown theories? Part of what it is for a child to treat a concept

as a natural kind concept is for the child to have developed a pattern of associations between concepts in a domain which is not a mere function of perceptual similarity. The existence of such patterns in children is evinced by the data collected by Keil and Gelman and Markman, and provides the empirical basis for Keil's attribution of theoretical beliefs to young subjects. According to Keil, children develop these patterns of associations as a result of "domain-specific biases acting as the precursors of specific theories." (Keil 1989, p. 273) Keil does not elaborate on the nature of these domain-specific biases, but such biases sound a lot like what we might call 'implicit beliefs' or 'implicit intentions', beliefs or intentions which we may conveniently attribute to a subject on the basis of his/her behavior, without the content of these states being explicitly represented in the cognitive system.[viii]

In what follows I focus narrowly on the issue of the child's general treatment of terms as natural kind terms. This is in contrast to Keil's emphasis on how the child's specific causal theories develop. My choice of emphasis is due partly to the fact that Keil's and Gelman and Markman's data provide stronger support for the claim that children treat natural kind terms as a functionally homogeneous group than they provide for Keil's claim that specific causal theories develop from innately given domain-specific biases. The development of BTT also motivates the focus on the child's categorization of LOT terms as natural kind terms; for BTT relies on the assumption that there is a well-defined set of natural kind terms in LOT, and this seems to require that there be some distinctive method of acquiring these terms which results in their being put into the category of natural kind terms.

Limiting our discussion to the question of how natural kind terms come to be treated as such, we seem to face a choice between two contrasting pictures: (1) From early on in development, children treat natural kind terms of LOT as such because children have explicit intentions to do so. If this picture is accurate, BT2 would seem to apply to the child's natural kind terms toward which these intentions are directed. There are two difficulties with this option. Firstly, it seems to rob BT1 of its work. BT1 becomes inert because at the time of the appearance of the child's first natural kind terms, BT2 already applies. Secondly, we would need another theory, independent of BTT, as to how content is determined for the terms which make up the child's explicit intentions to treat some terms as natural kind terms. (2) According to this second picture, children treat natural kind terms of LOT as such because of a (possibly innate) processing bias which is not explicitly represented, but which causes the child to treat LOT terms differently, depending, for example, on the nature of the stimulus which first triggers the tokening of the

term. Such biases are something like implicit mental states in the child. On this view, BT1 would apply to the child's everyday natural kind terms until which time the child (or young adult) begins directing explicit intentions toward these terms.

It is quite implausible that the child formulates conscious intentions to treat certain LOT terms as natural kind terms. However, the lack of the relevant conscious intentions does not rule out the presence of explicit, subconscious intentions with the requisite content to decide the issue between Options #1 and #2. In what follows then, I ignore the possibility that the child might consciously intend to treat certain LOT terms as natural kind terms. Instead, I attempt to understand whether, and how much, explicit representation is required in order for the child to effect the categorization of some LOT terms as natural kind terms.

2. Theory development and implicit intentions

a. Implicit rules and implicit intentions

Given the shortcomings of Option #1 described above, Option #2 seems more attractive. However, by identifying the child's relevant states as implicit, Option #2 puts us in an explanatory bind. The intentions and beliefs implicated by Keil's and Gelman and Markman's data (and, for that matter, Macnamara's data, see note #8 to this chapter) are supposed to be causally efficacious states. But if these states are only implicit, how can they be causally efficacious?

On one plausible theory of implicit states, the dispositional theory, an implicit state is not causally responsible in a strict sense for the subject's behavior at the time the state is attributed. According to the dispositional view, implicit states are mental states which have no current causal effect on the subject in question, but which the subject would likely enter under certain circumstances. Your average educated adult has, for example, a dispositional belief that 2,459,456 is greater than 2,459,455. This belief is dispositional for the average subject because there is no need to invoke the belief in the explanation of any of the average subject's behavior, and furthermore, its explicit representation is not implied by a general theory of learning or belief acquisition.[ix] We say that the subject has an implicit belief about arithmetic in this case because the subject has, we assume, an explicit desire to believe arithmetical truths and also explicit means for identifying such truths. (See Dennett 1987, pp. 216-218, for a discussion which makes the implicit/explicit distinction in roughly this way [although Dennett is wary of applying the distinction to

humans' mental states in the way that I suggest]. Also see Fodor 1987, pp. 21-26, and 1990b, pp. 23-24.)

The problem with the dispositional view is that it does not seem to allow implicit states causal efficacy. As I've described it, the dispositional view sees implicit states as potential states, i.e., states which we assume the subject would enter were certain circumstances to obtain, but which are not currently playing (and perhaps have never played) a role in the causation of the subject's behavior. Matters change when we consider dispositional rules, as opposed to states, of computation that a cognitive system follows (Fodor 1987, p. 22). It is the question of implicit rule following that seems relevant for our purposes; for in the case at hand, the child seems to have an intention which expresses a rule. To be (slightly) more specific, the child's intention expresses a rule of the form, 'Terms acquired under certain types of circumstances should be treated one way in processing and terms acquired under different circumstances should be treated differently in processing.' (Call this rule-schema 'R1'.[x]) In order to make R1 respectable for our purposes, we need to explain how R1 is encoded and how it causally affects the child's behavior.

Fodor offers one explanation of how an implicit rule can be causally efficacious in a cognitive system. According to Fodor, a system's behavior can be explained by reference to an implicit rule so long as two conditions are met, (1) the rule being followed must be "emergent...out of explicitly represented procedures of implementation, or out of hardware structures, or both," (1987, p. 25) and (2) the entities over which the rule quantifies must themselves be explicit representations.[xi]

Does R1 meet Fodor's two conditions? With respect to his first criterion, it is simplest to assume that R1 is hardwired, i.e., not emergent from other explicitly represented rules. If we claim that R1 is emergent from other explicitly represented rules, we create for ourselves the task of explaining how the explicit representations which make up these other rules have their content determined. While the assumption of hardwiring facilitates discussion of other aspects of R1, we should bear in mind that this assumption has the mere status of a plausible assumption.

With respect to Fodor's second criterion, we face a dilemma. The application of R1 is supposed to cause the natural kind terms of LOT to achieve their status as natural kind terms. Once the terms have such status, but not before, BTT tells us how their content is determined. Thus, R1 must be applied to LOT terms before the extensional contents of those terms are fixed. The rule R1 quantifies over entities (newly coined LOT terms) which, in the first instance at least, would not seem to be explicit representations. Put to our purposes, R1 violates Fodor's second criterion for being an implicit rule.[xii]

In his discussion of the issue of implicit versus explicit rules, Fodor is not concerned specifically with learning. When Fodor does discuss the learning of word meanings (cf. Fodor 1975, chapter 2, and 1981, chapter 10), he assumes that learning occurs by hypothesis confirmation. According to his view, "you cannot learn a language whose terms express semantic properties not expressed by the terms of some language you are already able to use." (Fodor 1975, p. 61) Fodor's solution to the dilemma presented above would appear to be to say that terms to which R1 applies, i.e., the natural kind terms of LOT, have their extensions fixed innately, before R1 is ever applied to them. This would mean that certain LOT terms, which eventually have the status of natural kind terms, would have content before they're categorized in the cognitive system as natural kind terms. On this view, R1 does not play a role in fixing the original content of natural kind terms. Instead, if there is any role at all for a rule like R1, it would be to help the cognitive system refine its treatment of natural kind terms of LOT, e.g., to treat them more appropriately in inferential processes.

We should be leery of Fodor's view. First of all, it is not consistent with BTT's approach. The Best Test Theory takes seriously the idea that if your theory of extension is supposed to apply only to natural kind terms, then you have to provide evidence that such terms make up a distinct, well-defined class of LOT terms. Fodor's view requires that you have an independent theory of content that assigns extensions to LOT terms before such terms are categorized as natural kind terms. Such an independent theory is something we are trying to do without (see note #12 to this chapter).

Furthermore, Fodor's radical concept nativism has struck many as implausible (cf. Katz 1995, pp. 501-502, Devitt and Sterelny 1987, p. 156, Jackendoff 1989, p. 98). Fodor's extreme nativism is just so...extreme (thus, radically counterintuitive). Additionally, if you are at all tempted to think that causal history plays a role in the determination of content, as Fodor himself sometimes is (Fodor 1990b, p. 121)[xiii], then the view that the semantic content of all natural kind terms of LOT is fixed independently of experience with the entities which are in the extensions of those terms seems that much less plausible. At the very least you can't have both the nativist view and a causal history view of LOT semantics.[xiv] It seems worth our while, then, to locate another way to resolve our dilemma.[xv]

Assume that R1 is encoded in the hardware, and that the LOT terms to which R1 applies do not have their representational content fixed at the time of R1's first application to them. On this view, R1 is not an implicit rule according to Fodor's criteria, for it fails to satisfy the second criterion. Perhaps, then, we

should reconsider Fodor's conditions on being an acceptable implicit rule.

Fodor's conditions are supposed to be satisfied for an implicit rule to be psychologically explanatory in a way that is consistent with what Fodor calls (and what many others have called) the 'representational theory of mind' (RTM). According to RTM, the representational (or for our purposes, 'referential') content of mental states (in part) explains the behavior of human subjects. As an advocate of RTM, Fodor wants to avoid explaining the behavior of subjects solely, or even largely, in terms of rule-following in the absence of representations. However, it may be that we can have our representational cake and eat it too, with frosting. We may be able to hold on to RTM in general, while denying that the LOT terms to which R1 applies (at least in its early applications to a given term) have representational content, and while still claiming that R1 is an implicit rule. Consider the following.

One of Fodor's primary concerns is to wed the computational theory of the mind to a representational view. The computational view of the mind says that cognition is the result of the application of rules (either explicitly represented or implicitly followed as the result of hardware structure) to internal states. Fodor weds representationalism to the computational view by claiming that the states on which internal computations are carried out are representations; they have extensions. Without denying the overall accuracy of Fodor's view, it seems perfectly reasonable to think that given the rich computational resources with which humans are endowed, humans may perform some computations on states other than representations. In order to explain how I add twenty-seven and one hundred and forty-two in my head, RTM would do best to appeal to computational rules defined over LOT terms which represent numbers. However, in other cases, perhaps in the application of R1, a computation may be defined over classes of mental structures which do not represent anything. This may even be an essential part of the process which takes the child to the point where she can compute functions over representations of natural kinds. Thus, I propose to give a broad definition of implicit intentions. According to the broad definition, an implicit intention is identifiable with the state which results from the application of a rule which is not itself explicitly represented and which, parting ways with Fodor, does not take explicit representations as inputs. Forthwith, I refer to Fodor's implicit rules as 'narrow' implicit rules (or beliefs or intentions, where this seems appropriate). I will also refer to the states constructed by the application of such rules as 'narrow' implicit intentions (beliefs, etc.)

How would a system work if it were to embody R1 as a broad implicit intention? Firstly, the system would have to have hard-wired computational routes leading from certain types of sensory stimuli to the

mechanism which introduces new LOT terms. For example, it may be that the child reaches a point in maturation such that whenever a sensory stimulus of any new cohesive object acting independently of human control leads to the tokening of a new term in LOT, that term is treated henceforth as a natural kind term.[xvi] The term need not have an extension in order to be so treated, but once it is treated systematically as a natural kind term, we are then free to say that the term acquires a determinate extension in accordance with BT1. This approach seems promising; the child's bootstrapping her way to semantic content seems an attractive alternative to Fodor's extreme nativism.

A complication arises when you begin to fill in the details of the bootstrapping process. We need to know how types of stimuli are identified by the child. For example, how does the child recognize a stimulus as a stimulus of the type presented by a cohesive physical object acting independently of human control? If this is done without any mitigation by representational states, then the child turns out to be doing a lot more bootstrapping than we might have first imagined. There would seem to be a lot of computation involved which is carried out over mental structures which do not represent. The cognitivist cry of 'No computation without representation' is placed in serious jeopardy. Above, R1 was introduced as a special case of computation without representation. However, we now seem to be at risk of eliminating cognitive states altogether. If the child can carry out such sophisticated procedures as the application of R1 without representing anything, many would be tempted to think that the mind can do all (or much) of its other work without the use of representations.

We might avoid the specter of the elimination of intentional states by claiming that even though the natural kind terms to which R1 first applies do not have content, other terms in the relevant computations must. The states resulting from R1's application would thus be intentions of a mixed nature. Such intentions would be composed of LOT terms some of which are representations and some of which are not. This seems reasonable in that in order to apply R1 in the way outlined above, the subject would seem to need to token a representation of a cohesive object acting independently of human control. Because R1 has some representations in its computational domain, it seems to be a narrow implicit rule after all, its applications resulting in narrow implicit intentions to use certain LOT terms as natural kind terms.[xvii]

While this approach has its merits, e.g., we stave off eliminativism, the picture presented so far is misleading in some respects. Firstly, looking ahead a bit, we may wonder how such representations as are in R1's domain get their content. If these representations have content before R1 applies to them, doesn't

that mean that for the purposes of applying BTT, we should consider the intentions that result from R1's applications explicit, even if the rule R1 is itself implicit? The newly coined LOT term to which R1 applies may not be an explicit representation at the time of R1's first application to it, but the stimulus to which the new term is applied is represented explicitly; it is a part of the intention which is constructed in the application of R1.

Secondly, 'object acting independently of human control' contains at least two natural kind terms, 'object' and 'human'. One of our primary goals is to explain how the child can treat natural kind terms in a special way in order to justify the claim that there are special extension-fixing principles, BTT's principles, that apply to them. But it seems that in order to get new LOT terms into the category of LOT terms, some terms (e.g., 'object' and 'human') must already be in that category. How does the LOT term 'object' get into the category of a natural kind term of LOT in the first place? A certain amount of nativism seems required. At the very least, terms like 'human' and 'object' would have to be classified as natural kind terms on some other basis than the child's early experience.

This is not an entirely implausible view. Much evidence supports the nativist view of the innateness of the object concept (Spelke 1990, 1991, and Bower 1989). While the infant's object concept is not identical to the adult's, key aspects of the adult object concept seem to be present in the child from a very early age, an age much younger than the age at which Keil's and Gelman and Markman's subjects begin to consistently treat natural kind terms in the adult fashion. For example, in Spelke 1991 (pp. 139-147), it is shown that infants as young as 2 and 1/2 months old treat objects as solid and temporally continuous.

b. Implicit intentions and the application of extension-fixing principles

A coherent picture of the role of implicit intentions is emerging. Certain natural kind terms (e.g., 'object') are innately treated as such.[xviii] If it can be said that humans have any intentions at all to treat these terms as natural kind terms, such intentions are broad implicit intentions. Beyond the LOT terms innately determined to be natural kind terms, the categorization of new LOT terms as natural kind terms is then effected by narrow intentions (and while these were called 'narrow implicit intentions' above, it should now be clear why that is a misnomer). These intentions are products of the application of an implicit rule, R1, which is (most likely) built into the child's hardware, and which operates partly on innately determined natural kind terms.

The question is left open, however, as to when BT1 applies to the relevant LOT terms and when the applicable principle is BT2 instead. One reasonable approach is to rule broadly implicit intentions out of the category of intentions referred to in BT2, but to claim that the presence of any narrow implicit intentions toward a term t automatically triggers the application of BT2 to t. On this view, when BT2 mentions intentions, BT2 refers only to intentions which are at least partly constituted of explicit representations. Call this the 'broad-exclusive' approach, 'broad' because we define implicit intentions broadly, and exclusive because we exclude implicit intentions, so defined, from the province of BT2. Furthermore, assume that the broad-exclusive approach is also narrow-inclusive, i.e., it includes narrow implicit intentions (any intention composed even partly of representations) in the domain of BT2.

On the broad-exclusive interpretation of BT2, BT1 applies to LOT terms which are treated as natural kind terms as the result of innate specification, so long as these terms achieve their status as natural kind terms independently of the use of any explicit representations of natural kinds. On the other hand, BT2 would apply to the terms of LOT whose status as natural kind terms is established by applications of R1. The principle BT2 would apply to such terms because applications of R1 are assumed to require as inputs at least some terms which are already categorized as natural kind terms and which already have content.

Even though we call broad implicit intentions 'intentions', it is important that BT2 not be triggered by their mere presence. Broadly defined, an implicit intention corresponds to (among other things) any state of processing which results from a hardwired processing bias where the processes in question don't take any representations as input or yield representations as output. Broad implicit intentions are everywhere. If we allow BT2 to apply to t whenever the subject has broad implicit intentions toward t, BT1 never applies. Take for example the innately determined treatment of term 'object' as a natural kind term. Assume that the infant begins treating objects as a natural kind as early as 2 and 1/2 months, and that over the first year the child's treatment of objects in the appropriate ways becomes more and more pronounced. Assume also that the infant's object concept is innately specified in the developing brain.[xix] The infant has a hardwired computational route from a certain range of stimuli to the tokening of 'object', and to the treatment of 'object' as a natural kind term. On the assumption that no representations come into play during this process, the operation of the mechanisms which connect certain stimuli to 'object' gives rise to a broad implicit intention to treat 'object' as a natural kind term of LOT. The result, then, of allowing BT2 to cover broad implicit intentions is that BT1 would seem to have no range of application. For example, at the time

of the appearance of what seems like one of the infant's first natural kind terms, BT2, not BT1, would apply.[xx]

As we've just seen, excluding broad implicit intentions from the purview of BT2 is well-motivated. In contrast, it's not entirely clear why we should include narrow implicit intentions as intentions whose presence triggers the application of BT2. As a first pass, we might say that BT2 applies because of the mere presence of representations in a narrow implicit intention to categorize a new term as a natural kind term. Rule R1 may truly be an implicit (i.e., not explicitly represented) standing intention to treat certain types of terms in a certain way. However, the intentions in which we are interested, those which result from specific applications of R1, seem to be explicit in nature. These intentions are specific states or events, made up at least partly of representations, and directly responsible for the categorization of new terms as natural kind terms. If these so-called narrow implicit intentions are representationally explicit, isn't that enough reason to say that BT2 applies to the LOT terms which are the objects of such intentions? Perhaps, but our guide here should be a consideration of those purposes for which we included BT2 as part of BTT in the first place. To a review of those purposes I now turn.

One function of BT2 is to solve the qua problem for natural kind terms of LOT to which BT1 does not apply. Often a sample of one natural kind is also a member of another natural kind (for example, a piece of gold is also a chemical element). Thus, intending that a term refer to a natural kind of which you have a sample is not enough to fix the extension of the term as a single natural kind. BT2 solves the qua problem by deferring to the concept in the mind of the speaker or thinker when she is grounding a term. When grounding a term t, the subject may confront a sample which belongs to more than one natural kind. However, the subject's concept at the time of grounding t is typically a much better test for one of these natural kinds than for the others. The success rate of one of these natural kinds relative to the concept which the subject has in mind when grounding t is typically much higher than the success rate for the other candidate natural kinds relative to that same concept.[xxi]

When the seven-year old child uses the term 'tiger', he is no more referring to mammals than is the adult who uses 'tiger'.[xxii] Thus, BTT should explain how the child, as well as the adult, solves the qua problem. The key question we face is whether the intentions which the child uses to categorize a term as a natural kind term are the same intentions the child uses to solve the qua problem. Simply knowing whether or not the child's narrow implicit intentions to categorize 'tiger' as a natural kind term are actually explicit in

nature does not decide the issue. We need to know whether the child's intention to treat a given LOT term t as a natural kind term is applied in the absence of intentions that would limit the extension of t per BT2, or whether the intention to treat t as a natural kind term and the intention to limit t's extension in certain ways are somehow combined into a single intention or mental act.

It seems quite plausible that in the very act of classifying a term as a natural kind term, the child would tie the term's use to the sample which prompted the categorization of the term. It is an empirical question as to how tight such a connection really is. However, it seems unlikely that a child would put an LOT term on her list of natural kind terms in LOT without associating any concept with that term's use. It is this association of the concept with the term, together with some demonstrative act toward the sample, which limits the application of t in accordance with BT2, thus solving the qua problem.[xxiii] While far from being conclusive, the preceding considerations suggest that as early as the child begins treating the kind terms to which R1 applies as natural kind terms, BT2, rather than BT1, determines the extension of these terms.

Since we have little knowledge of the details of the child's subconscious intentions when she classifies LOT terms as natural kind terms, we may wonder whether we could apply BT1 to the terms which the child categorizes as natural kind terms using R1. If we were to apply BT1 liberally, along these lines, what role would be left for BT2? We might consider retaining BT2 as an explanation of how people's application of conscious intentions can restrict the extension of an LOT term in the relevant ways. According to such a view, BT1 would apply to LOT terms toward which the subject has subconscious intentions containing representations. This can be allowed in keeping with the desire to avoid holism, so long as it's done properly. The key is to make sure that content accumulates in the proper building block fashion. This restriction implies that if the content of an intention toward a term t has any effect on how t gets its content fixed, then the intention itself must have its content fixed prior to its effects on the fixation of t's content. This view does not imply that any intention to treat a term as a natural kind term automatically triggers the application of BT2 to t. So long as the intention was not a conscious intention to limit the extension of the term in the ways specified by BT2, and so long as the intention has its content independently fixed, BT1 might still apply. I am not inclined to take the route just described. However, given the fuzziness of the notion of an implicit intention, and also given the fact that the extensions assigned will not differ in most cases regardless of whether BT1 or BT2 applies, I will not attempt to

completely resolve issues regarding BT2's range of application here.

Let us return now to the question of how extension is fixed for innately determined natural kind terms of LOT, bearing in mind the question of how the whole process of fixing extensions for natural kind terms gets started. Assume that the innate processes determining some LOT terms to be natural kind terms have their required effects in the absence of representations. Are there mechanisms which we can describe non-intentionally by which an innately determined function can connect types of stimuli to the tokening of a term like 'object', thus causing the term to be treated as a natural kind term? Lest the reader worry that we fall into the hopeless trap of concept empiricism, notice that we need not specify anything like a reductive definition of 'object' in terms of sense data. We need only to locate properties that can be defined in terms of basic physical structures and which can provide the child with a reliable indication of the presence of the kind of objects in question. Such structures would most likely be structures of the perceptual apparatus, e.g., elements of Marr's primal sketch (Marr 1982, pp. 91-96) which itself is reducible to physical patterns of retinal cell firings. There may be other ways to cash out such perceptual properties. For example, Bower argues that the properties to which the infant is sensitive are amodal properties of stimuli, i.e., patterns of relations among the parts of a stimulus which could be shared by stimuli coming from two or more different sensory modes (Bower 1989, pp. 25-39).

On the other hand, there may be reason to doubt that the non-intentional specification of perceptual stimuli related to the infant's tokening of 'object' is forthcoming. Bower 1989 (pp. 20-24) reports results which seem to indicate that very young infants treat sets of distal stimuli as equivalent rather than treating sets of proximal stimuli as equivalent. In Bower's experiments, infants typically responded in an equivalent manner when distal variables were held constant. In contrast, the infants treated as different situations where various distal arrangements were used to create that same proximal stimuli (e.g., patterns of the same size and shape on the retina). Bower's results suggest that the infant has representations of the external world which influence cognitive processing in the development of even the most basic concepts like that of an object. Whether these representations are representations of natural kinds is open to question. If not, then BTT as stated is a dependent theory of extension for natural kind terms, i.e., it must be supplemented by some theory of how representations at the lowest levels of cognition have their extensions fixed.

To resolve this question of dependence, we might notice that the infants in Bower's experiments focus on the consistency of the distal properties of size and shape. Since size and shape may well be taken to be objective properties of objects in the natural world, the child's LOT terms for particular sizes and shapes

would seem to be natural kind terms. Thus, such perceptual features (modal or amodal) may be the original natural kind terms in the cognitive system. In such a case, these would be the terms to which BT1 most clearly applies and to which BT1 must apply in order to get the extension-fixing ball rolling if BTT is to be independent.

To be confident in our application of BT1 to 'size' and 'shape', we should like to have data which shows that the child treats 'size' and 'shape' as natural kind terms.[xxiv] This data may be hard to come by. At the very least, it would not seem to come from experiments along the lines of Keil's and Gelman and Markman's. We can't expect to see a contrast between the way the infant treats the perceptual features of a square with the way the child thinks of the square's underlying nature. In the case of a square, such a contrast between perceptual features and underlying nature seems to make no sense. The square seems to wear its underlying nature right out where everyone can see it. Furthermore, it's impossible to get responses from infants on many of the relevant questions. We can't expect, for example, the infant to tell us whether a square is still a square when it's used to hammer a nail. Still, we should not underestimate the cleverness of experimenters. Perhaps the developmental data in which we're interested would come from ingeniously designed dishabituation experiments with infants. I leave the issue here, then, hoping that future data will resolve the question whether the infant treats 'size' and 'shape' as natural kind terms.[xxv]

We seem to have three options in identifying the province of BT1:

Option A: At the level at which representation of natural kinds first appears, the relevant terms are innately determined to be natural kind terms. The categorization of these terms as natural kind terms is effected without the aid of any representations. Principle BT1 applies to these innately determined LOT terms, and perhaps to further natural kind terms classified as such partly by the subject's use of innately determined natural kind terms.

Option B: The earliest treatment of LOT terms as natural kind terms term requires the subject's use of non-natural kind terms which have extensions. According to this option, at the level at which natural kind terms first appear, even if these terms are innately determined, the subject must already possess other representations which have their extensions fixed in accordance with principles other than those offered by

BTT.

Option C: Same as Option B, but with BTT scope expanded so that its principles tell us how extensions are fixed for these basic representational terms, even if the subject does not treat such terms distinctly, as natural kind terms.

So long as we are limiting the scope of BTT to natural kind terms, Option A seems to be the neatest of the three. According to Option A, the subject uses terms innately classified as natural kind terms to bootstrap her way to the fixation of extension for a whole range of natural kind terms. This picture allows BTT a degree of independence that seems desirable. Although lacking the neatness of Option A, Options B and C may seem more realistic. Which option we choose depends on when in the child's development good psychological explanations of the child's behavior begin to make reference to extensions (or to representations which are taxonomized according to types of things to which the representations refer). Bower's results, as interpreted above, suggest that this occurs in infancy. His results thus push us away from Option A, toward either Option B or Option C. The choice between the two depends partly on the answers to two other open questions (both of which will remain so here). One is the question whether the infant's terms like 'size' and 'shape' are natural kind terms of LOT. The other is the question as to whether the scope of BTT should be revised and extended beyond natural kind terms.

D. Why a Special Content-Determining Principle?

There is fundamental question of justification lurking behind the discussion of LOT term categories. Even if there are different LOT term categories, why should there be different content-determining principles for the different types of terms? To answer such a question, we might claim that BT1 defines something like basic reference which, in turn, provides the basis for all content (except perhaps the content of logical terms [cf., Fodor 1990b, pp. 93-95, 110-111]). This is a very sweeping claim, and although I suspect it is true, I do not argue for it here.

An alternative, and less grand, strategy, would be to think of the development of a semantic theory for LOT as a huge project which should be tackled a little bit at a time. Adopting this less ambitious strategy, we can see BTT as part of a divide and conquer approach to LOT semantic theory. Taking this approach, we attempt to identify the distinct types of LOT terms and give a reasonable semantic theory for each type of LOT terms. If we succeed in giving a reasonable theory for just one of these types of terms,

then according to this less ambitious view, we will have accomplished something important. Then, having developed a reasonable semantic theory for one of these groups, particularly if it is a group which seems fairly central, we may be in a position to extend at least some of the insights of this theory to the other types of LOT terms, one type of terms at a time. Ultimately, a unified theory of content for LOT terms may appear. But in the absence of a convincing unified theory, it seems justifiable to pursue the piecemeal approach, while maintaining hopes for theoretical unification at some later date.[xxvi]

Notes to Chapter IV

---

[i] Among the natural kinds themselves, there was a significant difference in how terms for animals, plants, and minerals were treated, with the animal terms being treated

most distinctively as natural kind terms and the minerals least distinctively as such (Keil 1989, pp. 186-187).

[ii] Syntactic categories play a similar role in the statement of formation rules which hold across all languages. The study of such rules (or classes of possible rules)

comprise what is known as the study of 'universal grammar' (Stillings, et al. 1987, pp. 378-382).

[iii] It is not precisely clear what implications linguistics and psycholinguistics have for theories of cognition in general. On one hand, it is claimed that language processing is

a modular affair (Fodor 1983b, passim, but especially pp. 46-92), and from this, one might infer that the investigation of such capacities does not tell us much about how LOT is used in cognitive processing in general (cf. Fodor's remarks about central systems [Fodor 1983b, pp. 101-119]). On the other hand, it has been claimed that language processing is non-modular, and that, on this ground, linguistic data reveal the general nature of cognition (Lakoff 1987, p. 67).

[iv] Robert Cummins has developed a theory of content, interpretational semantics, based on the observation of the way a function computed by a cognitive system can

mirror a mathematical function as it's instantiated in nature, or more broadly, in the external world (Cummins 1989b).

[v] Landau and Gleitman 1985, pp. 126-173, argue convincingly that the blind child (as well as, presumably, the sighted child) can use detailed information contained in the

syntactic structures of sentences to help her to identify the meanings of some terms. Landau and Gleitman do not argue that all semantic distinctions marked in LOT are reducible to the syntactic roles of the relevant LOT terms. However, their work does suggest that some semantic distinctions marked in LOT could be constructed entirely out of the syntactic roles of the terms involved, which roles Landau and Gleitman call 'subcategorization frames'. If this were true for all semantic distinctions marked in LOT, including the distinction between natural kind and artifactual kind terms, then perhaps we would be able to stick with LOT1 as originally stated. However, such a program of reduction is quite ambitious, and I do not assume that it can be successfully carried out.

[vi] Van Gelder 1990 provides a general description, together with some examples, of connectionist models that do away with traditional, recursively-defined (over

syntactic categories) formation rules. Also see Goschke and Koppelberg 1991 (pp. 137-157) for criticisms of the combinatorial approach to LOT semantics, and for the outline of a connectionist alternative to a strict combinatorial understanding of how we form complex concepts.

[vii] Similarly, Susan Carey interprets the results of her developmental studies of children's biological kind terms as showing that the children's concepts change as the result

of the domain-specific accumulation of knowledge (Carey 1985). Carey's view of concept development is substantially different than Keil's, however, in that Carey is inclined to think of conceptual change as something like a miniature, domain-specific, Kuhnian scientific revolution (Kuhn 1962). Carey (and Kuhn)

seems to think that theoretical relations entirely define a concept. In contrast, Keil claims that at least some portion of a concept can have atheoretical content and that conceptual change may result in the way the theoretical and atheoretical portions of a concept are weighted with respect to one another (Keil 1989, pp. 20-23). On the issue of how concepts are embedded in theories, see Murphy and Medin 1985 and Smith 1989. Both works argue that understanding concepts as embedded in theories is of great importance to our understanding of a variety of psychological phenomena (e.g., the comprehension of complex concepts [Murphy and Medin 1985, pp. 305-306]).

[viii] Macnamara 1982 also presents developmental data which raise questions regarding the role of implicit states in the cognitive system. Macnamara argues that in the

process of learning the rules of grammar, two to three year-old children construct the relevant syntactic categories by first employing a gross semantic taxonomy of the world around them. According to Macnamara, this taxonomy consists of general categories such as object, attribute, and action (Macnamara 1982, chapters 7 and 8). Macnamara's work raises challenging questions deeply related to those addressed in the text. Is the child's performance reported by Macnamara the result of the child's explicitly representing the relevant semantic distinctions, or is the differential treatment of terms by which the child constructs syntactic categories due to a processing bias built into the child's cognitive system? Furthermore, does the degree of explicit representation of the relevant semantic distinctions have any bearing on how the contents of the child's early words are fixed?

[ix] In the vernacular, we might say that the average subject has just never thought about this particular arithmetical comparison. I do not put matters this way in the text

because explicit intentions are not the same as conscious intention. All conscious intentions are explicit, but not vice versa.

[x] Granted R1 is a rule-schema. However, throughout the remainder of the chapter, I talk as if R1 is a specific rule with all the requisite details filled in. Nothing turns on

this use of 'rule' in the place of 'rule-schema', so long as the reader bears in mind that I have not provided such details as would transform R1 from a schema into a specific rule.

[xi] Fodor doesn't claim that these conditions must be met in order for any implicit rule to be causally efficacious. He makes the weaker claim that the two stated conditions

must be met by any implicit rules of an intentional psychology, i.e., a psychology which assumes the theoretical importance of talk about representations (Fodor 1987, p. 25). The view that the extensions of some LOT terms play a role in good psychological explanation was assumed at the outset of the present work. Thus, Fodor's additional qualification is omitted in the text.

[xii] One way around this problem is to claim that natural kind terms of LOT have referential content before they become natural kind terms. Such content would have to

have been fixed in accordance with principles of content other than those of BTT. However, it seems best in developing BTT to avoid deference to other content-determining theories. To defer to another theory in this way has the appearance of offering an ad hoc solution to a legitimate problem, and it undermines the optimism expressed earlier that we may be able to extend the scope of BTT to cover (virtually) all LOT terms.

Other options, which will not be pursued here, are (a) to adopt an expansive view of natural kind terms such that all LOT terms are, in the first instance, natural kind terms (thus all terms would be subject to BTT, R1 or no R1), and (b) to identify a process other than the application of R1 by which the child separates natural kind terms of LOT from other LOT terms.

[xiii] If Fodor assumes the pure-informational version of his asymmetric dependence theory of extension (see Baker 1991, p. 19, for a discussion of the different versions of

Fodor's AD theory), then the causal history of the subject's interaction with the entities in the extension of a given term is irrelevant to the fixation of that extension. Instead content is fixed by a pattern of nomological relations alone, which pattern may exist long before R1 is ever applied. As Baker points out, however, Fodor is not entirely consistent in his characterization of the role of a subject's causal history in determining the content of her LOT terms.

[xiv] A qualification is in order here. According to some causal-history based theories of content for LOT terms, content is determined by the causal history of the species,

not the individual organism. Such a view is consistent with the nativist conception of content insofar as the nativist view is applied to current members of the species (as opposed to the ancestors whose history led to the original determination of content). Fodor rejects such a view, however (Fodor 1990b, chapter 3).

Ruth Garrett Millikan is an example of someone who holds the view that content is fixed at least partly by species history (Millikan 1984). However, Millikan's theory of content is much more subtle and detailed than a simple evolutionary view, and thus I am hesitant to describe her view as a pure species-history view. One reason for my concern can be found in Millikan's discussion of how new words are introduced into a public language (Millikan 1984, pp. 81-82). According to Millikan, a new word can be introduced into a public language and acquire determinate content as early as the second use of the new word. According to Millikan's view, species history contributes partly to the fixing of the content of the new word, but so also does the history of the organisms who are still alive as the term acquires determinate content. Those organisms whose history was integral to the fixation of the new term's content may in fact be the very organisms who go on to use the new word to express determinate content. If such a view is to be applied to LOT, nativism will fail as a theory of extension for at least some terms and at least some subjects.

[xv] In fairness to Fodor, he may well have a way of making his concept nativism consistent with the demand that R1 play a role in the fixation of content for terms which

did not previously have content. At many points in his work, Fodor separates the semantics of LOT from its syntax. For example, one of Fodor's primary goals in Fodor 1994 is to explain how purely syntactic mental processes can respect the semantic relations that hold between the things to which LOT terms refer (cf. Fodor 1994, p. 14, and passim). As Fodor says, "*semantics isn't part of psychology*." (Fodor 1994, p. 38) And since I'm talking about whether or not the natural kind terms of LOT have certain semantic values (i.e., extensions) innately, Fodor could simply say that his innateness thesis is psychology, and that since I'm asking a semantic question, I'm misapplying his innateness thesis. In other words, Fodor could claim that his innateness thesis is limited only to the syntactic resources needed to learn language. Fodor could then agree that R1 is the guiding rule in the categorization of some LOT terms as natural kind terms where this categorization does not change the essential syntactic properties of the terms in question (although one must wonder what these innate, syntactic properties are). This view is consistent with the view of R1 that I've presented in the text in that (1) the terms to which R1 applies would have no representational content when R1 is first applied to them, and (2) these terms would then, after the relevant application(s) of R1, could have their extensions fixed in accordance with the principles of BTT. I'm not sure that this approach will work, or that Fodor would want to take it, but it is a possibility that is strongly suggested by some of the views he has taken.

[xvi] While this story stands a chance of being approximately correct in the case of biological kinds, it clearly does not apply to chemical or mineral kinds.

[xvii] It's not entirely clear that this categorization is correct. I am assuming that so long as the computation of R1 is carried out over some representations, Fodor would

count an application of R1 as an implicit intention. However, Fodor does not make it clear how much representation is required to make a rule's application explanatorily relevant within the framework of RTM.

[xviii] It may not be entirely clear what the innateness amounts to here. The idea is that for some natural kind terms, e.g., 'object', connections are built into the child's

hardware between certain types of stimulus and the specific LOT terms. These terms are innately wedded to a certain range of stimulus, none of which will occasion the tokening of a new LOT term. In contrast, when R1 is applied, whether or not a new LOT term is coined depends partly on the subject's learning history. In this way, R1 provides innate guidance to the child, but only with respect to what types of stimulus are indicative of the presence of a member of some natural kind or other.

[xix] Such an assumption probably oversimplifies matters, in that the mind/brain seems at odds with itself in developing the object concept. For example, Diamond 1991

provides striking evidence that the way the brain matures during the first year of life can hinder the child's appropriate treatment of objects. Diamond's work suggests that the child attempts to treat 'object' as a natural kind term even before the child has the necessary neurological tools to do so effectively and consistently.

[xx] 'Object' is used only as an example here. The infant may acquire numerous natural kind terms, perhaps as the result of innately determined processes, before the infant

begins to token 'object'. For such terms, we could construct arguments analogous to the one give in the text with respect to 'object' to point out the difficulties of allowing BT2's application to be triggered by the presence of broad implicit intentions.

[xxi] In chapter VI, I take up the question of what happens when two of these kinds have the same or nearly the same success rates relative to the term and the subject in

question.

[xxii] In light of Inhelder and Piaget's work on children's representation of class inclusion, one may think that children are in a different position here than that of adults. (See

Markman 1989, pp. 140-145, and Macnamara 1982, pp. 56-68, for summaries of the relevant results along with critical discussions of Inhelder and Piaget 1964.) Inhelder and Piaget seem to have shown that until children are 7-8 years old, they do not represent class-inclusion accurately. For example, in one experiment, the children had trouble classifying a given rose as a rose and a flower at the same time. Very quickly, the experiment runs as follows. The experimenter shows the child six flowers, three roses and three daisies. The experimenter asks the child, "How many roses?", and the child says, "Three." When the experimenter then asks the child, "How many flowers?", the child says, pointing to the daisies, "Three." Once the child has classified the roses as roses, the child seems unable to turn around and classify these same items as flowers.

Taking Inhelder and Piaget's results at face value, one may question whether the qua problem is actually solved by children. However, much of the critical discussion of Inhelder and Piaget's results in this area suggests that children's poor performance in class-inclusion experiments is due to task demands and other factors. Regardless of who is right in their interpretation of Inhelder and Piaget's experimental results, insofar as children sometimes react to a rose qua rose and sometimes react to a rose qua flower, we would seem to want attribute to children the necessary apparatus for fixing the extension of both 'rose' and 'flower' appropriately.

[xxiii] By going outside the set of natural kind terms, and bringing demonstratives into the picture, we leave the realm of BTT's application. However, any account of human

cognition or interaction with the environment must explain the role of humans' ability to focus their attention on specific inputs or observations. Thus, I assume an LOT device of ostension without giving an account of its semantics or its syntax.

[xxiv] It may be more accurate to talk about terms for specific sizes and shapes. However, for ease of expression, I ignore this complication in the text and speak simply of

the LOT terms 'size' and 'shape'.

[xxv] If it turns out that the mental structures which guide the early tokening of natural kind terms in LOT are not representations at all, BTT will be dependent in the

following sense. In addition to BTT, we will need a theory of the child's bootstrapping process which is driven, at the lowest level, by the tokening of something like phenomenal feature-structures. If these structures have any representational content, it would seem to come in the form of reference to types of phenomenal or proximal stimuli. Our bootstrapping theory would then have to explain how it is that these phenomenal feature-structures are used to classify terms as natural kind terms. Furthermore, the categorization of terms as natural kind terms will have to be done in a way that will facilitate the fixation of the extensions of the natural kind terms without requiring that these phenomenal feature-structures refer to natural kinds or properties.

[xxvi] There have been recent attempts to give a general semantics for LOT (Jackendoff 1989, Lakoff 1987, chapter 17). Disregarding questions about the adequacy of

such theories as theories of concept meaning, note that such theories as Jackendoff's and Lakoff's are primarily theories of intension and do not explain in any principled way how the extensions of concepts (or the associated LOT terms) are fixed.

V. NATURAL KINDS ONLY

According to BT1, a natural kind term t of LOT refers to the members of the natural kind for which the concept associated with the t provides the best test. To determine which natural kind is the kind for which the concept associated with t provides the best test, we have to compare the success rates of all natural kinds relative to t. What is striking about this picture is that success rates are not calculated for anything other than natural kinds. This does not imply that a natural kind term of LOT can not have a disjunctive extension (for examples of allowable disjunctive extensions, see chapter VI). However, by calculating success rates only for natural kinds, BTT rules out the possibility of a natural kind term having as a reference class the set of horses together with cows on dark nights. This narrowing of the field of candidate extensions for natural kind terms is key to BTT's solution of the DP. Thus, the primary aim of the current chapter is to justify the assumption that natural kind terms in LOT refer to causally homogeneous groups and not to the disjunctive sets which seem to cause the DP (call this assumption 'NKO', for 'natural kinds only').

A. A Methodological Argument for the Natural Kinds Only Assumption

As presented in Chapter 2, Criterion #1 sets the solution of the problem of misrepresentation (in particular, the solution of the DP) as a constraint on any naturalistic theory of extension. The following argument provides one possible justification of Criterion #1.

Argument #1-

> Premise 1. If we don't solve the DP and reference is allowed to be disjunctive in the way that the DP describes it, then psychology loses an appreciable amount of explanatory/predictive power.

> Premise 2. We don't want psychology to be void of any of its potential explanatory or predictive power.
>
> Therefore, a theory of content which is sufficient for psychology's purposes will have to solve

DP.

If this argument is sound, then for the purposes of solving DP, we can assume from the outset that the reference of natural kind terms in LOT is to natural kinds. A crude causal theory of extension would say that LOT terms refer to such classes as that of horses and cows on dark nights. Argument #1 claims that psychology is empirically weakened by allowing such classes as reference classes. If this is correct, then we are justified, as a matter of good scientific methodology, in modifying our theory of extension to do away with such disjunctive extensions in a principled fashion. In particular, BTT is justified in calculating success rates only for natural kinds.

Call any theory that makes the NKO an 'NK theory'. Further, we'll use the name 'DR theory' for any theory that endorses extensions that are disjunctive in the sense which causes the DP. Translating premise #1 of Argument #1, it says that NK theories are explanatorily/predictively superior to DR theories. In such a case, if you are only interested in giving a naturalistic theory of content, your philosophical theory which solves DP does not need to explicitly to rule out the view of content assumed by DR theories. If NK theories are empirically superior to DR theories, then we don't need a philosophical theory which tells us, for example, that 'horse' refers to horses and not to horses or cows on dark nights. We only need a theory of extension for natural kind terms that tells us which of the various natural kinds 'horse' refers to, i.e., the theory need only tell us why 'horse' refers to horses rather than cows. This, of course, is what BTT does.

Before we flesh out the details of Argument #1 any farther, we should address a question that may seem natural to many philosophers. Aren't there other reasons, one might ask, besides our desire to underwrite all or part of psychology, for wanting a philosophical theory of extension which solves the DP for LOT terms? One reason for wanting to solve the DP, a reason which may have little to do with the theoretical foundations of psychology, is that DP is a purely philosophical puzzle which arises as a result of our attempt to straighten out our pretheoretical concept of representation. Philosophers' puzzles which result from conceptual tensions have traditionally been thought to be worth solving independently of any concern for the theoretical foundations of the empirical sciences.

There is nothing particularly wrong with trying to solve philosophers' puzzles. However, given the stated goals of the present work, I give little attention to this aspect of the DP. The only philosophical motivation I can offer for taking an NK approach is a general assumption of naturalism. Whatever the extensions of natural kind terms, the assumption of naturalism tells us that these extensions are determined

by the natural order. And since the natural order consists of relationships between types of things (i.e., regularity in the interactions between types of things), it seems justified to build a theory of reference for LOT terms on the relations between types of things. In the case of BTT, this amounts to the calculation of success rates for natural kinds only. In this way, BTT focuses on the relation between one type of thing (a natural kind) and another type of thing (a type of term in a subject's LOT).

Premise #1 of Argument #1 says that, generally speaking, NK theories in psychology are more powerful than DR theories. The remainder of the chapter is an attempt to support this claim.

B. Natural Kind-Based versus Disjunctive Reference-Based Psychological Theories

1. Parameters

There are two primary requirements that examples cases must meet in order to be relevant the present debate. The first requirement is that the case in question must be a case in which BT1's application yields a determinate answer. Extension-fixing principle BT1 states that the reference of t is the natural kind which has a success rate higher than that of any other natural kind relative to t. It is clear that 'higher than' is a graded relation. Once one has determined which natural kind has a higher success rate than all other natural kinds, the further question remains as to the relevance of the degree to which one kind's success rate is higher than other kinds' success rates. How do we decide when there is a clear winner? Does there have to be a natural kind whose success rate is some minimum percentage higher than the nearest competitor in order for reference to be determinate?

In chapter VI, I consider cases where success rates for two natural kinds relative to a single LOT term are equal or approximately equal. However, in order to simplify matters, we will restrict the current discussion to cases where there is a clear winner in the comparison of success rates, i.e., we will restrict the discussion to cases where one natural kind has a substantially higher success rate than the natural kind with the next highest success rate relative to the term in question.

The second condition on example cases is that they must be cases in which having any external reference at all is explanatorily useful or important. Some philosophers claim that there are no such cases (cf. Stich 1983). As a basic assumption of the current work, I reject such skepticism outright (or, at least, I hold it at arm's length). This assumption granted, there may still be specific cases where talk about extensional content doesn't seem to do any explanatory work. For example, for some low-level data structures, it may not seem that there is any theoretical benefit, vis a vis the explanation of

behavior, which results from assigning reference classes (of either the NK or the DR variety) to these structures. Accordingly, in what follows, I only consider examples involving LOT terms which correspond to what people normally think of as consciously held concepts. I assume that the terms corresponding to these consciously held concepts have external reference and that this reference, disjunctive or otherwise, is explanatorily important to psychology.

While it is convenient to set NK psychological theories against DR psychological theories in a general competition, such talk oversimplifies matters. It's possible be that there are some cases in which NK theories are more powerful than DR theories, and that there is a different range of cases in which the reverse is true. If the distinction is clear between the cases in which a theorist is supposed to apply an NK theory and those cases in which it is better to apply a DR theory, the two types of theories may be able to coexist in the scientific community.

Accordingly, we should be modest in our goals in this chapter. We need not show that NK is, without a doubt, the superior theoretical structure in all of psychology. In order for this chapter to succeed, we need only to show that there is an important range of cases in which NK theories are likely to be empirically superior. If such a range of cases exists, and is well-defined, then NK theories (and philosophical theories which serve them [e.g., BTT]) have a secure theoretical place in psychology.

2. Comparison of the theories

A simple causal theory of extension implies that LOT terms have a kind of disjunctive reference. 'A' refers to all of the different kinds of things that cause 'A', including all of the kinds of things that, speaking pretheoretically, are sometimes 'mistaken' for As. This result seems to cause deep trouble for DR theories. In what follows, I run through a number of objections to DR theories, and consider possible DR responses to these objections. Ultimately, I conclude that there is no reason to prefer DR theories to NK theories. To the contrary, NK theories seem to have a number of advantages over DR theories.

a. Predictive inadequacy

Assume that a subject S possesses the LOT terms 'A' and 'B' and has what we would normally describe as a good command of the concepts associated with 'A' and 'B'. Subject S rarely mistakes an A for a B and vice versa. However, assume that S has on at least one occasion mistaken an A for a B and has on at least one occasion mistaken a B for an A. According to a DR theory, both As and Bs are in the

extension of both 'A' and 'B' for S. This situation causes difficulty from the viewpoint of someone who is trying to predict S's behavior using a DR theory.

Assume that a DR theorist knows that S will soon come into contact with a specific A (call it 'a1'). According to a simple DR theory, a1 is in both the extensions of 'A' and 'B'. But considering only the extensions of 'A' and 'B' for S, the DR predictor will not have any grounds on which to predict A-recognition behavior over B-recognition behavior. By hypothesis, S reliably recognizes As (as well as Bs). Thus, A-recognition behavior is, in fact, much more likely to occur in response to the encountering of a1 than B-recognition behavior. And in contrast to a DR theory, an NK theory will provide solid ground for predicting this A-recognition behavior instead of B-recognition behavior.

The type of case under consideration is common in the realm of folk psychology. Folk continuously make reasonably accurate predictions about how the people around them will react to newly encountered items. The overall accuracy of such predictions is not affected by the fact that humans occasionally make mistakes when categorizing newly encountered individuals.[i]

What can be said in defense of DR theories? Predictions in many sciences are probabilistic, especially the social and behavioral sciences. Predictions made by DR theories may also have a probabilistic nature, ignored by the preceding criticism. To facilitate accurate prediction, the DR theorist might attach weights to the various natural kinds which make up the extension of an LOT term. (The weights might, for example, be determined by the causal history of the individual subject.) A DR theorist could then give a probabilistic prediction of S's behavior where the probability of S's exhibiting a certain behavior, say, A-recognition behavior, in response to a1 would be equal to the likelihood (as determined by S's causal history) that S will token 'A' in response to any given A. Assume that in 95% of the past cases where S has encountered an A, S categorized it as an A. Assume that only two percent of the time in the past when S has encountered a B, S has categorized the B as an A. Based on these facts about S's history, the DR theorist can say that there is a 95% chance that S will engage in A-recognition behavior and only a 2% chance that S will engage in B-recognition behavior in response to a1. In terms of probabilistic weights, in the extension of 'A', As have the weight 0.95 and in the extension of 'B', As have the weight 0.02. Based on these weights, the DR theorist can make the correct prediction that in all likelihood, when faced with an A, S will exhibit A-recognition behavior.

One might think that by taking into account the odds of S's making a mistake, a DR theory which

builds probabilistic weights into its LOT extensions (call such a theory a 'DRP' theory) would be predictively superior to an NK theory. However, any advantage for DRP would seem to be illusory. When predicting behavior, the NK theorist must also take into account the likelihood of errors in categorization. Presumably, NK theories will follow the folk here, and make affordances for error by considering a subject's past history together with a description of the situation at hand.

The DR theorist could claim that she has a more powerful theory of extensional content than the NK theorist on the grounds that more information is built into the extensions themselves in a DR theory. However, the importance of such information would seem to be overstated. As described thus far, the weightings attached to the different natural kinds found in the extension of S's term 'A' are determined solely by the proportion of the number of times in the past S has tokened 'A' in response to As to the number of times S has tokened 'A' in response to Bs. However, this is not the crucial piece of information necessary to accurately predict errors in categorization. Instead, it is the taking into account of an accurate description of the circumstances surrounding the categorization act that is crucial to the accurate prediction of errors in categorization. In many cases, knowledge of the circumstances surrounding a particular act of categorization would seem to be override any information provided by DRP's weightings. For example, in a case where S categorizes a1 under ideal conditions (no technical meaning intended), there would seem to be no grounds for saying that there is a two percent chance that S will categorize a1 as a B, even if S has categorized Bs as As in two percent of S's past encounters with Bs.

The weightings assigned by DRP theory don't seem to give the DRP theorist any advantage over the NK theorist. If the DRP theorist wants to make the most accurate predictions possible, the DRP theorist would need to consider the same factors as the NK theorist, i.e., past history together with a description of the situation at hand. Even though the DRP theory builds potentially useful information about past errors into the extensions themselves, this information is often misleading and must be discarded in favor of information regarding the circumstances of categorization.

b The mistaken assignment of false beliefs

It seems that DR theories have to do without the very useful predictive strategy of assuming that people have a lot of true beliefs regarding personal matters (e.g., regarding their own likes, dislikes, and past experiences). Instead DR theories imply that such beliefs are, for the most part, false.

Consider the following case. John plans a trip to a dude ranch. John has two beliefs which he

expresses prior to his departure. He believes that horses are fun to ride, and he believes that Brahma bulls are not fun to ride. Also assume that John has, on at least one past occasion, mistaken a Brahma bull for a horse (say, when driving past a rodeo at dusk). Now imagine that John goes to the dude ranch. Having ridden some horses and enjoyed it, he decides one night to go out for a ride. In the dark, John mistakes a Brahma bull for a horse, carelessly jumps on, gets thrown, knocked unconscious, and winds up in the hospital. (In other words, he doesn't have any fun.)

An NK theory predicts the outcome of John's midnight ride based on the beliefs John professed before the ride. An NK theorist assumes that John's beliefs in this case are probably true and uses them to guide her predictions. The NK theorist endorses such conditionals as the following: "If John rides a horse this week-end, he will have fun," and "If John rides a Brahma bull this week-end, he will probably not have fun." Taken together with the fact that John does ride a Brahma bull on the week-end in question, the latter conditional implies the outcome of John's ride. By assuming that John has true beliefs regarding such mundane matters as his own tastes in recreation, beliefs based on his past experiences, the NK theorist is likely to accurately predict the outcome of John's ride.

DR theories don't seem to fare as well. Because a Brahma bull caused a tokening of 'horse' in John prior to his dude ranch vacation, Brahma bulls are automatically in the extension of John's LOT term 'horse', according to a DR theory. If the DR theorist assumes that John's beliefs, expressed prior to departure, are true, then she won't know whether to predict that John will have fun on a Brahma bull or not. According to a DR theory, when John thinks, "Riding horses would be fun," John is really thinking a thought which, roughly speaking, has the extensional content **riding anything which has ever caused me to token 'horse' would be fun.**

The DR theorist cannot productively use the common strategy of assuming that a subject has reasonably accurate beliefs about personal matters, such as his own taste in recreation. This is not just another example of the predictive inadequacy of the DR approach. What distinguishes objection #2 is that the DR theory seems to imply that John's original belief "riding a horse would be fun for me" was false. Given the DR truth conditions of this belief, riding a Brahma bull should have been fun for John. Since it was not, the DR theory seems committed to the implausible claim that John was wrong in his evaluation of his own recreational tastes.[ii]

A DR theory might get around this difficulty by introducing a theory of partial truth. When a

particular belief predicates a property of an extension class, a DR theorist could claim that the predication is likely to be true of part of the class and likely to be false of other parts. A method similar to that used in the calculation of probabilistic reference in a DRP theory could be used here to separate reference classes into the relevant parts. For example, the DR theorist could define a portion of the reference class (e.g., the horses portion of the DR extension of 'horse') to be the primary reference of the term in question. Then the DR theorist could assume, for the purposes of making predictions, that a subject's belief about a matter of personal taste is likely to be true when the LOT terms which comprise the belief in question are evaluated only according to their primary reference.

Here again, DR theories have to go through theoretical contortions in order to stay on a predictive/explanatory par with NK theories. This is a strike against such theories. This will be especially clear when we take up questions of practicality and ease of use below.

c. Second-order problems

If a DR theory is true, then it's difficult to see how real, living DR theorists can refine their theories in response objections #1 and #2. In order to develop and apply a probabilistic version of DR, or to develop a DR theory of partial truth, the DR theorist has to have some way to represent the different natural kinds which constitute a given DR reference class. The DR theorist may be able to give a general theory of primary reference, because at the general level, the DR theorist may be able to do so without herself representing specific natural kinds. However, as soon as the DR theorist attempts to apply the theory to any specific case, she will encounter difficulty.

Imagine that a DR theorist is trying to identify the class of horses as having a 0.95 weighting in the extension of 'horse' for a given subject S (so, e.g., the theorist can predict horse-recognition behavior to be more likely than cow-recognition behavior when S encounters a horse). In order to do so, the DR theorist would have to be able to represent the class of horses. However, assuming that DR theories are true, the DR theorist cannot do so. So long as a DR theory of reference is assumed, the DR theorist cannot represent the natural kind horses. When the DR theorist tries to represent horses, say, by tokening 'horse', all of the other kinds of things which have ever caused the tokening of 'horse' will be included, at least to some extent, in the extension of the DR theorist's LOT term 'horse'. The DR theorist cannot designate horses as privileged in the extension of her term 'horse', for she almost certainly has no term in her LOT which has horses and only horses as its extension.

There is a sense in which it doesn't matter to DR theorists whether psychologists are able to represent natural kinds or not. In the grand scheme of things, a DR theory could still be true, regardless of the difficulties humans might have in representing the details of DR theories. However, we should be impressed again by the extreme practical difficulties faced by those who wish to develop DR theories, difficulties the suffering of which offers no apparent reward.

d. Difficulty of use

A striking fact about DR theories is that nobody seems to be offering them. One is hard put to find psychological work which combines the two following assumptions: (1) talk about the reference of natural kind terms is an important part of some good psychological explanations, and (2) a natural kind term refers to all of the types of things that have ever caused its tokening. Put a little differently, it's hard to find psychological theories which make claims of the form "whenever a subject encounters anything which has ever caused the tokening of 'A', the subject will X," where X can be thought of as a description of A-recognition behavior. This suggests that there is something empirically inferior about DR theories.

Consider a case where a researcher wants to know how subjects react when they think "There's an A." The dominant experimental paradigm of the day directs the researcher to design the experiment so that the conditions of categorization insure that the subject will not mistake the experimental As for something else. Theories of the DRP variety seem to recommend a different strategy. If the important information for predicting errors is built into the DRP extensions, and researchers look only to these extensions to guide them in constructing experiments, the resulting strategy for preventing errors in the experimental condition would be to find subjects with the right history (i.e., subjects who have, as nearly as possible, in the past categorized all As as As and only As as As). This strategy is, so far as I can tell, never pursued. Another experimental strategy suggested by the DR approach is to sift through all of the DR extensions of a subject's LOT terms to find out how often S mistakes A's for other things, proportionally discounting experimental exposures to actual A's in order to account for the fact that the A's in the experimental situation might be mistaken for these other things. While it is somewhat common for experimenters to take error rates into account in an experimental paradigm, the calculation of such rates is normally based on pilot runs of the specific experiment in question, not on the results of a search through the experimental subjects' DR extensions for

their LOT terms.[iii]

One explanation of the fact that cognitive psychologists do not actively develop DR theories, and the accompanying experimental paradigms, is that they simply do not want to go through all of the difficulty of developing DR theories when these theories offer no apparent advantage over currently held NK views. If DR theories are to be satisfactory, they must be constructed (at great pains) to do everything standard NK theories do. However, no advantage over NK theories seems to result from devising, for example, a theory of partial truth or a system of probabilistic weighting of the natural kinds in DR reference classes.

Another reason why DR theories may not be developed in cognitive psychology is because people find it unnatural to think of extensions as DR reference classes. In chapter IV, I reviewed evidence that people are inclined, from an early age, to treat the reference classes of natural kind terms as causally homogeneous, in the sense that the members of a given reference class are expected to share important theoretical properties. To adopt the DR approach would be to struggle against such an inclination. You would have to give up the assumption that members of a kind can be expected to exhibit similar behaviors, causal influences, etc., and you would lose the advantages that go along with the human ability to categorize items quickly and to efficiently draw, on the basis of category membership, inferences about how the items will behave or affect the other items around them.[iv]

The DR theorist might respond by claiming that the human tendency to think of categories as causally homogeneous is simply that, a human tendency. If, instead, people were to think of reference classes as weighted collections of causally homogeneous sub-classes, people could survive just as well. Only the structure of the inferences would be changed, the DR theorist might claim, not the outcome of the inferences.

The DR response is a non-sequitur. It is irrelevant whether or not it would be possible for humans to reconceive of the extension of their natural kind terms of LOT in accordance with DR theory. The present claim is merely that humans have a strong tendency to think in NK terms. And if humans are naturally inclined to think of their terms as referring to natural kinds, then NK theories should be pursued rigorously. As a general rule, when faced with cases where two notations or sets of concepts would seem to be equally good for predictive and explanatory purposes, humans should develop theories based on the notation or conceptual structure which is the more accessible of the two (i.e., more accessible of the two to humans). Humans would seem to get more truth, more quickly, by employing the concepts and terminology which

come naturally (rather than struggling with a difficult theoretical structure), so long as there is no reason to think that the alternative, more complex terminologies or sets of concepts would offer greater theoretical power. In so far as humans do get more truth more quickly using the more accessible theoretical structure, there is a clear sense in which the resulting theories are empirically superior to those developed on the foundation of a difficult or opaque theoretical structure. Even if the two theories are analytically equivalent, the simpler (vis a vis human capacities) theory is still empirically superior in the real world because it leads to more accurate predictions and explanations, and more of them.

In this chapter, I hope to have made the prospects for DR theories seem rather dim. The types of adjustments one is likely to have to make in our normal ways of thinking in order to pursue a DR theory come at too great a cost and offer no apparent benefit. In closing, however, we might consider a worst-case scenario. What would become of BTT if DR theories were to win out on empirical grounds? The Best Test Theory as stated implies the truth of NKO. If we have reason to think that NKO is false (say, because DR theories are empirically superior to NK theories), then we have reason to think that BTT is false. But rather than discarding BTT part and parcel, there are rather trivial adjustments which we could make to BTT which could give BTT a central role as part of the conceptual foundations of DR theorizing. The clearest case of such an adjustment would be to construct BTT as a theory of primary reference for natural kind terms of LOT rather than a theory of reference for natural kind terms of LOT. Since a DR theory would need such a theory of primary reference, BTT could be reformulated to say, in a nutshell, that the primary reference of a natural kind term of LOT is the natural kind for which the concept associated with that term provides the best test. Although we have every reason to think that NK theories are superior to DR theories, BTT has available to it the fallback position that, if DR theories were to somehow emerge as the empirical favorites, the central ideas behind BTT would yet be well worth developing.

Notes to Chapter V

---

[i] In the text, I focus only on the advantage of NK theories over DR theories in predicting behavior. However, NK theories would seem to hold a similar advantage over DR theories in the realm of explanation. Imagine that the DR theorist tries to explain, after the fact, why S exhibited A-recognition rather than B-recognition behavior in response to a1. Since a1 is in the extension of both 'A' and 'B', the DR theorist would have no reason to think that a1's being an A explains S's A-recognition behavior. A DR theory seems to imply that the chance of S's exhibiting A-recognition behavior in response to a1 is equal to the chance of S's exhibiting B-recognition behavior in response to a1.

[ii] There is a fairly well known range of cases in which people do not accurately report their own mental states (Nisbett and Wilson 1977, Stich 1983, pp. 228-242). However, in contrast to the relatively artificial situations constructed by Nisbett and Wilson for their experiments, I have described a rather mundane example in the text having specifically to do with the accuracy of people's beliefs about their own tastes. There is no reason to assume that in such situations, people do not accurately report their mental states. Furthermore, I assume that even if people are sometimes wrong in these reports, we get better predictive results by making some use of these reports than we would get if we were to ignore first-person reports. When we accept these reports as prima facie accurate, we seem to make better predictions regarding subjects' behavior than we could have made had we not based our predictions, in any way, on subjects' first-person reports.

[iii] A third DR strategy would be to figure out what items the subjects mistake for A's, and in what frequency, and include a proportional number of such other 'A' causing items in the set of test items. This strategy is, however, no more popular than the two suggested in the text.

[iv] This suggests another argument in favor of NKO, one based on evolutionary considerations. Assume, as we have been assuming, that extensional content matters to good psychological theorizing. Also assume that people take their natural kind terms to have causally homogeneous kinds as their extensions (see chapter IV). Then, given that people's inferences using natural kind terms are frequently successful, chances are that people's terms have the types of extensions people think they have. It seems unlikely that (1) the extensions of these terms really matter psychologically, (2) our survival depends on the successful use of these terms, while at the same time (3) we are wrong about the nature of the extensions of these terms. These three claims could all be true at the same time. Much has been made of the possibility of false beliefs having survival value, e.g., believing that all causes of a certain term t are predators may be biologically useful even if some of the causes of t are not predators (cf. Cummins 1996, pp. 45-48). However, the operative question here is how often it happens that false beliefs (especially false beliefs of such an enormous scale as the belief that natural kind terms refer to causally homogeneous classes) are evolutionarily useful.

VI. OBJECTIONS AND REPLIES

A. Objection Number One

1. Extensions shift arbitrarily

If BTT assigns extensions on the basis of a subject's actual history, extension is subject to seemingly arbitrary changes. Consider the following unusual case. A subject S has encountered numerous tigers and has almost always recognized them as such. Say the success rate of tigers relative to S's term LOT 'tiger' is 0.95. Imagine now that S encounters a hyena for the first time in her life and mistakenly tokens 'tiger'. Based on this one encounter with a hyena, the success rate of hyenas relative to S's term 'tiger' is 1/1 = 1. Thus, it seems that after this one encounter with a hyena, the extension of S's term 'tiger' shifts from tigers to hyenas because hyenas have a higher success rate relative to 'tiger' (= 1) than tigers have (= 0.95). This result runs contrary to widely felt intuitions, and thus constitutes a forceful objection to the actual history version of BTT.

2. Discussion and replies

Before considering how we might modify BTT in response to the tiger/hyena example, we should be sure that there is a need to endeavor such modification. We should have in mind some concrete problem for psychological prediction or explanation caused by the tiger/hyena example. Take prediction. We would like to be able to predict what S will do the next time S sees a hyena. Assume that S is afraid of tigers, but that S's past responses to members of previously unencountered animal species has been a display of cautious curiosity. Given this, we would like to be able to predict, on the basis of BTT, that there is a reasonable chance that S's next encounter with a hyena will cause S to behave in the manner of cautious curiosity, rather than fear (although the precise outcome of the meeting will depend on surrounding circumstances). The problem with BTT, actual history version, is that given the fact that hyenas are now in the extension of 'tiger',

BTT implies that the next meeting with a hyena will result in fear behavior. All, or nearly all, of S's past tokenings of 'tiger' which resulted from S's being in close proximity to a member of the extension of 'tiger'

were accompanied by such fear behavior. Thus, given that 'tiger'-tokening has in the past been accompanied by fear behavior, and given that hyenas are now the extension of 'tiger', BTT would seem to predict that S will engage in fear behavior upon meeting another hyena.[i] [ii]

The simplest way to defend BTT against the tiger/hyena objection is to advert to the composition of S's tiger concept. If S is a typical subject, it seems likely that BT2 fixes the extension of 'tiger' for S. And since we've assumed that S has a concept of tigers which is a much better test for tigers than it is for hyenas, BT2 assigns tigers as the extension of S's term 'tiger', based on the composition of S's tiger concept.

Unfortunately, there's a catch. One might worry that problems structurally similar to the tiger/hyena problem could arise for terms to which BT1 applies. Imagine that S meets the member of a natural kind heretofore unexperienced by S, and misapplies a natural kind term t in S's LOT to which BT1 applies. The extension would seem to shift in the way the extension of 'tiger' seemed to shift above, but in contrast to the tiger/hyena case, S has no concept associated with t which itself has fixed extensional content. Without such a concept to fall back on, we seem to have no grounds for claiming that objection #1 is refuted, and that t's reference does not shift in a seemingly arbitrary way.[iii]

Cases where BT1 applies are the cases where Objection #1 has actual force. However, rather than attempting to identify a term to which BT1 unquestionably applies, I continue to call the problem at hand the 'tiger/hyena' problem, and I continue to address the problem as a problem caused by S's interactions with tigers and hyenas. Nevertheless, we must bear in mind that the problem is, at root, a problem with the application of BT1, and that we can not appeal to the constituents of S's tiger concept to solve the tiger/hyena problem.

a. An actual history-based response

In defense of an actual-history based version of BTT, we might appeal instead to principles of sampling to explain more fully how extensions might be relevant to psychological explanation and prediction. The Best Test Theory assigns extensions on the basis of statistical information. Just like other statistically-based generalizations, BTT's principles work best in the service of explanation and prediction when the application of the principles is based on a large sample size, in this case, a large number

of interactions between the members of the relevant natural kinds and the subjects in question. Thus, we should base predictions of S's behavior when meeting the next hyena on our knowledge of the extensions of S's terms, where these extensions are determined by as large of a sample size as possible. On this approach, the prediction that S will exhibit fear behavior upon seeing another hyena seems unwarranted. Hyenas became the extension of 'tiger' on the basis of just one interaction with S.[iv] In contrast, hyenas are likely to have been in the extension of other of S's LOT terms (e.g., 'unfamiliar species'[v]) whose extensions are much more stable. We can assume that the extension of 'unfamiliar species' is assigned on the basis of numerous interactions between S and the members of the extension of 'unfamiliar species'. Therefore, of the extensions to which a given hyena belongs (i.e., the extensions of 'tiger' and 'unfamiliar species'), the extension of 'unfamiliar species' is by far the more stable of the two. It is more widely tested and confirmed, we might say. Thus, when we use BTT-assigned extensions to make predictions about S's reaction to the next hyena, we should base these predictions not on the fact that hyenas are in the extension of 'tiger' but instead on the fact that hyenas are in the extension of 'unfamiliar species'.[vi]

Basing our predictions on the most highly confirmed extensions seems to be an effective way to predict S's behavior in response to the next hyena. Will this approach work when attempting to predict S's future reaction to tigers? Statistical confirmation is a process which takes place over time, over numerous sample events. Hyenas became the extension of S's LOT term 'tiger' after S had only one experience with hyenas, but S has, presumably, had numerous interactions with tigers. Thus, the extension of 'tiger' before the reference shift to hyenas was much more stable and sound than the extension of 'tiger' after the shift (i.e., after S encountered her first hyena). Therefore, when making predictions about S's future behavior in response to meeting a tiger, we should base our prediction on the stability and longevity of tigers' tenure as the extension of S's term 'tiger'. In comparison to the stability of tigers as the extension of 'tiger', hyenas have the lowly status of only a short-lived and poorly confirmed extension of 'tiger' for S.[vii]

b. A counterfactual-based response

This statistics-inspired maneuver seems to defuse any objection to BTT based on examples like the tiger/hyena example. However, concern lingers. What really seems to be causing the difficulty for BTT in the tiger/hyena case is the fact that regardless of a subject's actual history, it is the nature of the concept associated with an LOT term which seems to determine that term's extension. In the

tiger/hyena case, the tigers are well-entrenched as the extension of S's term 'tiger' at the time of S's first meeting with a hyena. This fact is essential to the success of the statistics-minded solution to the tiger/hyena problem. But what if S were to instead have fairly little past causal experience with tigers, yet still have what we would normally think of as a very good concept of tigers? What seems to be needed is some consideration of counterfactual cases in the calculation of success rates. When S has had little past experience with tigers, it seems we need to appeal to claims regarding which LOT terms S would token under various possible circumstances in order to justify the assignment of tigers as the extension of S's LOT term 'tiger'. In what follows, then, I present an alternative version of BTT which directs us to consider counterfactual situations in the assignment of extensions to natural kind terms in LOT.

Before describing the details of the counterfactual-based version of BTT, a disclaimer or sorts is in order. Only one type of case is driving the interest in counterfactual claims. We are worried only about cases where a natural kind term t of S's LOT meets three conditions: (1) t is subject to BT1,[viii] (2) t has been caused very few times by members of what we take to be t's appropriate extension, and (3) S misapplies t to the member of a kind whose members have never before caused S to token any LOT term. There may well be very few terms which meet all three of these conditions simultaneously. From the naturalistic point of view, the rarity of such cases is relevant. In the physical sciences, if a theory provides accurate answers across a wide-range of centrally important cases, this is often enough to garner the adherence of the scientific community. Therefore, as we proceed through the development of the counterfactual version of BTT, we should keep in mind how very limited the benefit of doing so seems to be.

According to the counterfactual version of BTT,[ix] the extension of a natural kind term of LOT continues to be the natural kind for which the concept associated with the term in question provides the best test. Which natural kind this will be continues to be determined by a comparison of success rates. However, on the modified version of BTT, the success rates being compared are counterfactual success rates, i.e., success rates based on what term the subject in question would token under certain circumstances. The first four arguments for the success rate function remain the same, a natural kind, a subject, a natural kind term in the subject's LOT, and a time. But a fifth argument is added, a specification of relevant circumstances. The success rate is now given by dividing the number of times members of K would cause any tokening of an

LOT term in S under the relevant circumstances into the number of times members of K would cause S to token t in the relevant circumstances given S's constitution at m.

In order to apply the counterfactual version of BTT (BTT-C, hereafter), we must clearly specify the circumstances relevant to the calculation of a given kind's success rate.

> CF1- A set of circumstances is relevant to the calculation of the success rate of K for term t and subject S at time m if and only if those circumstances are circumstances in which S has, in the past, tokened any LOT term which a member of K has ever caused S to token. (If no member of K has ever caused the tokening of any term in LOT, then the success rate of K relative to t is 0.)

These circumstances have the special feature of being drawn from S's past history. By basing the relevant circumstances on S's actual history as BTT-C does, we avoid having to specify anything like ideal conditions for the observation of members of a certain kind, and we also avoid any worries about what would happen if S were in an unfamiliar setting, on Mars, for example. We also avoid having to specify what would happen in Fodor's nomologically altered worlds. (See chapter II for a critical evaluation of Fodor's reliance on counterfactual situations in which the universe has been nomically altered.)

When calculating the success rate of hyenas relative to S's term 'tiger' in the tiger/hyena case, the relevant circumstances would stack up as follows. We would have to consider what would happen in circumstances just like those present in S's lone meeting with a hyena. We would also have to consider all of the circumstances in which S has tokened 'tiger' in the past. As a result, BTT-C keeps hyenas out of the extension of S's LOT term 'tiger'. Consider what terms S would have tokened if a hyena had been present in past cases where S tokened 'tiger'. Assuming that in many of these past situations where S tokened 'tiger' there was good light, S was not drunk, etc., hyenas, had they been present, would have caused the tokening of some other term besides 'tiger'. Thus, the counterfactual-based success rate of hyenas relative to S's LOT term 'tiger' is fairly low.

In contrast, the counterfactual-based success rate of tigers relative to S's LOT term 'tiger' should remain basically the same as it was on the actual-history based approach. The counterfactual situations to be considered are those situations where tigers caused the tokening of 'tiger' in S's past, where tigers caused the tokening of some other LOT term, all those circumstances under which S tokened any LOT term which had ever been mistakenly applied to tigers, and those circumstances where S had tokened 'tiger' in response to a non-tiger. What one winds up with is a set of counterfactual tests run under a wide range of circumstances drawn from S's past. Humans have the ability to make correct categorizations in a wide range of

circumstances. Thus, under such a wide range of testing conditions, there is every reason to believe that the success rate of tigers relative to S's LOT term 'tiger' will be much higher than that of hyenas.

According to CF1, a number of counterfactual situations are relevant to the determination of success rates for the various natural kinds relative to S's LOT term 'tiger' at a specific time in S's history. However, CF1 does not determine the relative importance of these different situations. Do any of the past circumstances in which a tiger has caused the tokening of 'tiger' carry more weight than other such circumstances in determining the success rate of tigers relative to 'tiger'? I will take the seemingly ecumenical approach, and give equal weight to each of the counterfactual situations identified by CF1. However, one of the virtues of this approach is that it is not as egalitarian as it may sound. Many of the situations identified as relevant by CF1 have elements in common. Thus, such elements are given more weight than other elements as environmental factors which are counterfactually relevant. For example, we can be fairly sure that most of the situations in which S tokened 'tiger' in the past are situations in which S was awake. Because of its preponderance as an element of the various situations identified by CF1, 'being awake' is thus favored as an environmental factor, over a factor like 'hanging upside down from a tree', which, we will assume, S has rarely been doing while tokening 'tiger'.

We should want the counterfactuals identified by CF1 to have as determinate a character as possible. In this regard, we may worry that CF1 does not tell us in detail how actual past situations are to be transformed into relevant counterfactual situations. Consider situations in which actual tigers have caused S to token 'tiger'. Isolate one of these, and call it TC. In order to make TC into a relevant counterfactual situation, we might simply try extracting the tiger from TC and inserting a hyena in the tiger's place, leaving everything else the same. This seems like a good strategy to pursue. If we want to know whether a hyena would have caused the tokening of 'tiger' in TC, it seems sensible to substitute a hyena for the object that caused 'tiger', that object being the tiger in TC.

Problem: How do we identify the cause of the tokening of a specific LOT term? Why remove the tiger from TC (putting the hyena in its place) rather than some other element of TC? Weren't there other things besides the tiger which causally contributed to the tokening of 'tiger' in TC? In chapter III, I identified two notions of relevant causality which might be employed in spelling out the details of BTT, those expressed by CR1 and CR2. The latter made a distinction between direct and indirect causation which may seem to be of use here. For example, we might say that when constructing the relevant counterfactuals, we substitute a hyena for the direct cause of 'tiger'. If 'tiger' was the output of one (or more) of S's sensory

modules in TC, then 'tiger' was, in the terminology suggested in chapter II, observationally-caused. To identify the tiger in TC as the direct cause of 'tiger', however, we have to claim that the tiger was the unique item which impinged on S's sensory apparatus resulting in the sensory module's output of 'tiger'. This is where the distinction between direct and indirect causation gets fuzzy, and thus, does not seem to serve our purposes. Among all of the causal factors contributing to a proximal stimulus, how do we identify one item in the world as the item impinging on the subject's sensory apparatus?

In chapter III, we hoped to get along without the distinction between direct and indirect causes, for the distinction is difficult to make precise. We should do the same here, and attempt to solve the tiger/hyena problem employing CR1 rather than CR2. According to CR1, there are other causes of S's tokening of 'tiger' in TC which would seem to be equally good candidates for replacement by our counterfactual hyena.[x] Consider again the example of the heart from chapter III. CR1 implies that S's heart was a legitimate cause of the tokening of 'tiger' in TC. The heart had causal effects on S without which S would not have tokened 'tiger'. Why shouldn't we substitute the hyena for S's heart when we construct a relevant counterfactual situation from TC?

We might rule out the heart on the grounds that the heart is internal to S, and that if you put a hyena inside S, in the place of S's heart, S wouldn't token any LOT term except possibly 'aarrgh'. We could try replacing S's heart with a hyena, but it would only lower the success rate of hyenas relative to S's term 'tiger'.[xi]

But what about the case of an external entity whose presence in TC was such that, without its causal effects on S, S would not have tokened 'tiger'? What might such a thing be? Consider a past case, TF, where a comrade prompted S to look at a tiger, where S's tokening of 'tiger' ensued, and where the tiger would have gone unnoticed by S without the comrade's prompting. Were it not for the presence of S's comrade, S would not have tokened 'tiger'. This makes S's friend a legitimate cause of 'tiger' according to CR1. Thus, our substitution strategy seems to tell us that we should replace S's comrade in TF with a hyena and see whether or not S still tokens 'tiger'.

Substituting a hyena for S's comrade substantially alters TF, and it's difficult to know what the result would be. It makes good sense to say that if the friend in TF had walked on without saying a word, S would never have noticed the tiger. However, when we replace the friend with a hyena, whether or not S notices the tiger depends largely on how the hyena acts. (Does it, for example, run toward the tiger?) The difficulty of

our epistemological situation matters not, however. There is a reasonable chance that in TF, when the friend is replaced by a hyena, S would not token 'tiger'. This is enough, together with the outcome in various other relevant counterfactuals, to bring down the success rate of hyenas relative to S's LOT term 'tiger'.[xii]

Version BTT-C seems to solve the tiger/hyena problem (as well as some other problems discussed below). However, BTT-C does so at some cost. The counterfactual-based version of BTT is nearly byzantine in its structure. This can be seen particularly well when one considers the enormous range of facts about counterfactual situations that go into determining the extension of one LOT term. Dissatisfied with this aspect of BTT-C, I continue to offer alternative, non-counterfactually-based replies to the objections raised below, even in cases where an appeal to BTTC might seem to do the trick.

B. Objection Number Two

1. Multiple winners

The Best Test theory does not tell us what to do when the stable success rates of two or more natural kinds relative to a given term are equal (but substantially higher than any other natural kind's success rate relative to t).

2. Discussion and reply

Objection #2 as stated does not specify exactly what circumstances might give rise to equal or nearly equal success rates. Specification of such circumstances is essential to our dealing with Objection #2, for different types of cases are to be dealt with differently. One type of case would be the case where, intuitively speaking, a subject cannot differentiate between two natural kinds, and thus categorizes the members of both kinds under the same heading. Frequently discussed cases of this type are the elm tree/beech tree case (Putnam 1975, pp. 226-230) and the case of jadeite and nephrite (Putnam 1975, pp. 241). In both of these cases, a subject uses one LOT term (e.g., 'elm') to refer to two natural kinds (e.g., elms and beeches) because the subject cannot tell the difference between the two kinds. In terms of success rates, both kinds in question (elms and beeches) have high, roughly equal (probably fairly stable) success rates relative to the term in question ('elm').

In a case like the elm tree/beech tree case, we may be inclined to say that the LOT term in question refers to both of the natural kinds in question. However, the verdict depends on the details of the particular

case. Which content-determining principle is operative is one of the relevant factors in determining extension in the cases under consideration. Assume that BT1 applies to a given term t. In this case, when two or more natural kinds have equally high success rates relative to t, and no other natural kind has a higher success rate relative to t, then t has truly disjunctive reference.[xiii] I take this situation to be unproblematic. That it is unproblematic is especially clear if one assumes BTT-C. On BTT-C if two kinds have equally high success rates relative to t for S, this means, in effect, that S can't tell these things apart (at least not in the relevant counterfactual circumstances). And since S has no intentions to differentially apply t vis a vis the two natural kinds in question, there seems to be no reason to think that t refers to one of the kinds and not the other.

Now assume that BT2 is operative in grounding the extension of an LOT natural kind term, say, 'elm'. Whether or not 'elm' has disjunctive reference for subject S depends on the samples toward which S has borne kind-minded intentions and on the nature of S's specific intentions. Imagine a case where S has interacted with numerous samples of elms and beeches. Imagine further that S has borne toward these samples kind-minded intentions which depend for their reference-fixing efficacy on a representation of observable properties of the samples. Since, by hypothesis, S cannot tell elms from beeches, we can safely assume that there is no systematic difference between the observable properties referred to in S's intentions toward samples of beeches and elms. S's concept associated with 'elm' (in particular the portion of it present in S's descriptive, extension-fixing intentions) is an equally good test for beeches as it is for elms. And since S has borne such intentions toward samples of both kinds, BT2 implies that S's LOT term 'elm' refers to both natural kinds, elms and beeches.

Although there may be cases where extension is truly disjunctive, there may be other cases where a subject S cannot tell the difference between two natural kinds, but where we want to allow that S is only thinking about one of the two natural kinds involved, rather than both. For example, S may be reading a book about trees, and may want to think about species of trees grouped in the same way that the expert author of the book groups them. The best way to deal with cases of this type may be to modify BT2 to allow S, via reference-fixing intentions, to defer to an expert's concept to ground the extension of a term. When facing a sample elm, for example, S could intend that her LOT term 'elm' refer to the natural kind for which the expert's concept, the concept an expert would have in mind when facing the same sample, provides the best test. S uses her intentions to fix on the sample and identify the relevant LOT term. However, in our

modified version of BT2, we are directed to consult the expert's concept to find out which kind is the natural kind with the highest success rate relative to S's LOT term 'elm'.[xiv]

The approach outlined above determines a non-disjunctive extension only where S has interacted with samples of one, but not both, of the two natural kinds which are, by hypothesis, indistinguishable to S. If S has interacted with both elms and beeches, for example, and has tokened 'elm' in the reference-fixing intentions directed at both types of samples, then even if S has deferred to experts to ground 'elm', the extension of 'elm' for S will be disjunctive. When S faced elms, the expert concept which would have been operative in those same situations, and to which S deferred, would provide the best test for elms. At the same time, in those cases where S interacted with a beech, S's deference to experts would have succeeded in getting beeches into the extension of 'elm'; for in those cases, the expert's concept, which she would have had in mind in when facing the same sample, would have been the best test for beeches. Thus deference to experts seems to remove the disjunctive nature of the extension of 'elm' just in cases where S has borne kind-minded intention toward samples of one of the two natural kinds which are indistinguishable to S. (This shortcoming of the deference-to-experts approach could be removed were S has to have an expert on hand to walk S through the grounding of 'elm' so that S doesn't accidentally have reference-fixing intentions toward any beeches. This is hardly likely to happen, however. And if S were to have such an expert on hand, S might as well just ask the expert to teach her the difference between elms and beeches.)

There may yet be a way for S to defer to experts to rid 'elm' of its disjunctive extension, even when S has borne reference-fixing intentions toward both elms and beeches. However, this may require that S defer to the use of experts' natural language terms. Consider the following principle, BT4:

> BT4- Assume that for each natural language term t, a subject has a unique LOT term which is most closely associated with it (but not necessarily vice versa). Assume also that a subject S intends a natural kind term t of S's LOT to have the same extension as the LOT term which an expert most closely associates with the natural language term identified by S in her content-determining intention. In such a case, t has the same extension as the expert's LOT term which is most closely associated with the natural language term specified in S's extension-fixing intentions.

Principle BT4 applies in the presence of thoughts of roughly the following type. "I want my LOT term 'elm' to refer to whatever an expert is thinking about when she uses the English term 'elm'." The assumption that S shares certain natural language terms with some experts limits the range of application of BT4. This limitation is not as restrictive as it might seem, however. It is common for a subject to not think about a kind until which time the subject makes acquaintance with the associated natural language term.

Subject S can render the extension of an LOT term more determinate in other ways as well. By building a uniqueness clause into her intentions, S can fix the reference of a natural kind term to one natural kind Ki even when other natural kinds Kj-Kn exist which are, to S, indistinguishable from Ki. Subject S can achieve such an effect by intending that the LOT term in question refer to the unique natural kind membership in which is responsible for the presence in the sample of certain specified properties (typically, observable properties).

The explicit uniqueness intention may justify our feeling that in certain cases, we can discount some of S's reference-fixing intentions. Imagine that there is a local kind, say gold, which is quite commonly found in S's environment. Imagine also that there is a type of mineral which looks a lot like gold (call it 'fool's gold') and is very rarely found in S's environment. (Contrary to the situation in the real world, fool's gold is so rare that the average member of S's community will have only one or two encounters with it in her life.) Assume now that S has had only one encounter with fool's gold and on that one occasion bore the intention toward fool's gold that S's LOT term 'gold' refer to it. Assuming that S has had numerous reference-grounding interactions with real gold, and that S has always intended that 'gold' refer to the unique natural kind responsible for, say, gold's beautiful luster, we may be justified in saying that for S, 'gold' refers to gold, not gold and fool's gold. Our grounds for saying so may be simply that the number of S's interactions with real gold far outweighs S's single interaction with fool's gold. Given that S intends 'gold' to refer to the unique natural kind responsible for the observable properties of her samples, and given that gold is the natural kind with which S has regular interaction, it seems reasonable to claim that gold is the sole natural kind in the extension of S's LOT term 'gold'.

In preceding sections, I emphasized the importance of the stability of extensions in the process of giving accurate psychological predictions and explanations. The solution just proposed to the gold/fool's gold case seems to rely on the same kind of considerations. However, two comments are in order. Firstly, there is a point of disanalogy between the discussion of the gold/fool's gold example and our earlier discussion of the stability of extensions. In the tiger/hyena case, for example, we assumed that S would distinguish between hyenas and tigers upon repeated exposure to hyenas. This was assumed to be true whether or not S had an explicit uniqueness intention. The current case is slightly different in that we are assuming that S is not able to tell the difference between gold and fool's gold, and may have no expert to defer to in order to give gold its unique status as the extension of 'gold'. While an explicit uniqueness intention might have done the trick for S in the tiger/hyena example, it was important in that case to see what

could be said on S's behalf even if S did not have an explicit uniqueness intention. Thus the discussion earlier of the stability and instability of extensions.

My second comment on the connection between stability and the uniqueness intention is related to the first. In the tiger/hyena case, considerations of stability helped us solve problems of prediction and explanation. Such problems would not seem to arise in the gold/fool's gold case. For example, since S cannot tell gold from fool's gold, there is no reason to predict that S will react any differently when, and if, S has a second encounter with fool's gold than S did upon first encountering the mineral. With respect to its effect on S's behavior, gold and fool's gold seem to be on a par. In fact, you may wonder why we should not simply treat this case as a case of disjunctive reference. One reason we might not be inclined to treat this as a case of disjunctive reference is because of the extreme rarity of fool's gold (taken together with S's uniqueness intention). Were S to be made aware of the fact that she once had ahold of a substance essentially different from her everyday gold, she may well deny that the oddball stuff is the kind of stuff she thinks about when she thinks 'gold'.

Some may place little weight on such predictions about what S would be likely to say. However, the reader who is inclined to think that gold really is the sole natural kind in the extension of S's LOT term 'gold' can hold on to this intuition and still give a reasonable psychological explanation of why S reacted to fool's gold in the same way she normally reacts to real gold (and of why S, remaining unenlightened would react in the same way were S to ever see fool's gold again). It's true that in S's first encounter with fool's gold, S acted just like S does when dealing with real gold. But this is not to be explained by adverting to the extension of 'gold'. One type of substance is in the extension of 'gold' and according to the current argument, the other kind, i.e., fool's gold, is not (or if it is, it is not a stable aspect of the extension). Thus, the fact that S behaves the same way in the presence of either substance is explained not by reference to the extension of S's LOT term 'gold', but by reference to the fixed extensions of the descriptive LOT terms S applies to both gold and fool's gold. Both gold and fool's gold are in the extension of S's LOT term 'having that gold luster'. Facts like these explain why S responds in the same way to samples of both types, even though one type is in the extension of S's LOT term 'gold' and the other type is not.[xv]

C. Objection Number Three

1. A close second

What about nearly equal success rates? How does BTT deal with the case where one natural kind has a slightly higher success rate than another relative to a given natural kind term in a given subject's LOT?

2. Discussion and reply

In the discussion of Objection #2, I talked only about problem cases where two natural kinds have equal success rates relative to a given term. In reality, success rates of two natural kinds relative to one term in a subject S's LOT may rarely be perfectly equal. This seems to be especially true if we stick by the actual history version of BTT (although similar difficulties may also arise if we calculate success rates using BTT-C). A reasonable approach to take here is to simply discount small variations in success rates when they do not seem statistically significant.[xvi] Assume that the success rate of elms relative to S's LOT term 'elm' is 98% and that the success rate of beeches relative to S's LOT term 'elm' is 97.5%. Assume further that the next highest success rate of any natural kind relative to S's LOT term 'elm' is 5%. (Practically speaking, this means that there is another kind of tree which, one out of 20 of the times S encounters its members, S tokens 'elm'.) The difference between the success rates of elms and beeches relative to 'elm' is one one-hundred and eighty-fifth the size of the difference between the success rate of beech and the success rate of the next highest natural kind (i.e., 0.5 is 1/185 of 92.5). Thus it seems reasonable to group the elms and beeches together in the extension of S's LOT term 'elm', and to leave all other natural kinds out. The general response to Objection #3, then, is that the case just described is typical of cases where two (or more) natural kinds have roughly equal success rates (and no other natural kind has a success rate substantially higher than the highest of the roughly equal rates).

D. Objection Number Four

1. Superordinate terms and the qua problem again

What about superordinate terms, for example, 'mammal'? Won't the success rate of tiger relative to 'mammal' be just as high as the success rate of mammals relative to 'mammal'? How do we choose which of the two natural kinds, mammal and tiger, is the extension of 'mammal'?

2. Discussion and reply

This appears to be a very serious problem, in many ways similar to the qua problem addressed in chapter III. However, this problem is easily solved. First off, notice that 'tiger' does not refer to mammals, for mammals have a fairly low success rate relative to 'tiger'. Mammals cause the tokening of too many other terms besides 'tiger'. But what happens when we consider the term 'mammal'? If S believes that all tigers are mammals, then there seems to be a perfectly legitimate sense in which tigers cause the tokening of 'mammal' in S every time tigers cause the tokening of 'tiger'. This means that the success rate of tigers relative to 'mammal' is very high, at least as high as the success rate of mammals relative to 'mammal'. How does BTT identify mammals, rather than tigers, or tigers and mammals, as the true extension of 'mammal'?

The superordinate problem can be solved quite easily. The simplest approach here is to concede all, and note that doing so does no harm. Assume that tigers (and perhaps any other single mammal group) have as high a success rate relative to 'mammal' as mammals do. The discussion of Objections #2 and #3 implies that BTT assigns the disjunction of the natural kinds mammal and tiger as the extensions of 'mammal'. Notice that we get exactly the extension we want, after all. The union of the collection of tigers and the collection of mammals is identical to the collection of mammals.

We've solved the superordinate problem for extensions. The extension of a term is simply the group of objects to which the term refers. Assigning the same object to a group twice is merely redundant. It doesn't change the identity of the group. Some might insist that a difficulty remains, however, in the realm of referential content. Recall that in chapter I, we distinguished between extension (and reference) and extensional content (and referential content). Referential content is determined by the type of thing to which a term refers, and there is a clear sense in which the abstract type mammal is different from the abstract type tiger or mammal. These two types seem to correspond to two different properties. Does BTT imply that the referential content of 'mammal' is the abstract type tiger or mammal? If so, isn't this an embarrassment to BTT?

According to BTT, the determination of extension is logically prior to determination of referential content. Referential content may correspond to something like an intension for natural kind terms in LOT (as was suggested in chapter I). However, referential contents are, first and foremost, determined by extensions. The referential contents of mental states or LOT terms are typed according to the kinds to which they refer.

Thus, before referential content can even exist for t, t must have an extension.[xvii] Having solved the superordinate problem for extensions, we have automatically solved the superordinate for referential contents. The extension of 'mammal' is the group of actual mammals. Referential content is characterized by finding the natural kind membership in which is shared by all and only those things in the relevant extension). Therefore, BTT assigns mammal as the referential content of 'mammal'.

The principle BT2 was included in BTT partly in order to solve the qua problem, discussed in chapter III. The relationship of a superordinate category to other categories contained entirely within the superordinate category is one of the most common causes of the qua problem. By solving the superordinate problem above, without using any of the machinery of BT2, we have shown how BT1 alone solves the qua problem, at least for a wide range of cases. Given this success, one may wonder whether there is any need for BT2 to solve the qua problem. Fodor has argued that, in effect, any of the intentions that are operative in fixing the content of a term must be contained in a non-intentional, physical description of the subject (and, presumably, her environment [Fodor 1987, pp. 121-122, Loewer and Rey 1991, pp. 313-314, Fodor 1990b, pp. 124-127]). It may be that content is built up piece by piece with the old pieces helping to shape the new, as I have claimed. But Fodor's view seems to imply that any effect that established content can have in fixing new content can be captured, in principle, anyway, just by specifying the non-intentional, physical facts of the situation. Doesn't all of this make BT2 obsolete (or, at least, redundant)?

We shouldn't dismiss BT2 too hastily. First off, we may reasonably doubt the accuracy of Fodor's views just described. Secondly, even if Fodor is right that a complete non-intentional description of the world fixes the content of subjects' LOT terms and the relations between them, these physically specifiable facts won't change the brute success rates. As noted in chapter III, BT2 functions partly to explain how reference can be fixed in a way that is substantially independent of success rates. Principle BT2 allows intention to take control of the extension-fixing process, the result being that brute success rates can be overridden. However, this also opens the door to the qua problem, simply because success rates alone no longer fix extension. At the level of the intention-fixed extensions, then, BT2 seems needed to solve a re-emergent qua problem. Thirdly, even if Fodor is right about the in principle availability of the relevant non-intentional descriptions, BT2 still has special explanatory value. If we want to capture what is common to the way numerous different thinkers use the same intentions to ground a new term, focusing on the non-intentional physical mechanisms at work gets us nowhere. Different subjects are differently constructed,

at least to some degree. Thus, even if a group of subjects all use the intentions with the same contents to fix the extension of a new term, these intentions will be instantiated differently in the different subjects. The non-intentional specification of the mechanisms that they use to fix the extension of the new term will be no more than a disjunction of non-intentional descriptions of the various subjects in the group. Only BT2 tells us what is common to the way the various subjects fix the extension of the new term in question. What is common to all of the subjects is the content of their extension-fixing intentions.

E. Objection Number Five

1. The Best Test Theory leads to phenomenalism

Assume that there are natural kinds of proximal stimuli. It seems possible that a certain natural kind of proximal stimulus would have a higher success rate relative to a given natural kind term than the natural kind of distal stimulus which, for the purpose of psychological explanation, we would normally think of as the reference of the term in question. Take, for example, the pattern of retinal cell firings which normally results from seeing a horse in profile from a certain distance (call this pattern 'P1'). It's easy to imagine that the success rate of such a pattern of retinal cell firings would be very high relative to a given subject S's LOT term 'horse'.[xviii] At the same time, one can easily imagine that horses sometimes cause the tokening of terms other than 'horse' in S, in which case the success rate of horses relative to S's term 'horse' will be lower than the success rate for P1. The Best Test Theory implies that, under the circumstances just described, the extension of S's LOT term 'horse' would be a natural kind of proximal stimulus, i.e., P1, rather than horses.

2. Two replies

First response: We might simply dismiss natural kinds of proximal stimuli as candidates for the reference of natural kind terms of LOT. The Best Test Theory is motivated by extension-based psychological explanations, and such extensions are normally external to the subject (i.e., beyond the subject's skin). Thus, the nature and the purpose of our project may motivate the outright dismissal of types of proximal stimuli as candidates for the extensions of natural kind terms in LOT. There are two problems with this approach, however. Firstly, P1 is not a member of a natural kind whose instantiations are underneath the skin of only one person. In so far as P1 is a characterization of a sensory stimulus taken from

the natural sciences, P1 is a natural kind which can, in principle, be instantiated in many different people. In other words, P1 is not some sort of subjective kind, which can be excluded from the category of legitimate natural kinds on the grounds of its subjective nature. Secondly, we should recognize that the sciences often talk about types of things that reside beneath the skin (or which exist at the level of the skin) of the subject. Internal organs provide examples of such types.

Second response: Given that the success rate of horses relative to 'horse' will be close to that of P1 relative to 'horse', the extension of 'horse', according to BTT, is the set of horses together with all of the proximal stimuli with equal or roughly equal success rates as horses have relative to 'horse' (see the discussion of objections #2 and #3 above).[xix] Is this a problem? What difference does it make to psychology whether 'horse' refers to horses or the disjunction of all such proximal stimuli with horses added in? It seems that the extension of 'horse' should only consist of the class of horses. However, if embracing the disjunctive option does not reduce psychology's empirical power, then perhaps we should bite the bullet, and accept this strange sounding extension for 'horse'. If, on the other hand, the assignment of proximal stimuli as parts of extensions for terms such 'horse' reduces psychology's empirical power, then we have trouble.

Trouble is averted, however. Assume that BTT assigns, as the extension of 'horse', the set of horses together with all such proximal stimuli as meet the general description given above of P1. The predictive and explanatory power of psychology then remains intact. When making predictions about S's behavior upon seeing the next horse, the fact that horses are in the extension of 'horse' for S to will facilitate reasonably accurate prediction. If we knew ahead of time which proximal stimulus the next horse was going to cause in S, we could check the extension of 'horse' to see whether it's there and base our prediction on such knowledge, but it's not important that we go through this process. In most cases, the exact stimulus which the next horse will cause in S is not in the extension of 'horse' at the time of the prediction of S's behavior. This is because the proximal stimuli which cause S to token 'horse' vary quite a bit, and are rarely repeated, and if a natural kind has no members which have ever caused to tokening of t, then the success rate of that kind relative to t equals 0. However, given that the next horse already is in the extension of 'horse' for S, we can rely on this fact in making our prediction.

What about explanation? How, for example, do we explain what's common to the behavior of different subjects in response to horses? To do this, we seem to need to abstract away from differences in

proximal stimuli and appeal to the fact that the subjects all have an LOT term which refers to horses and whose tokening (partly) causes the behavior in question. Variations in proximal stimuli in the extensions of the various subjects' 'horse' terms only seem to impede the explanation.

At this point, we should to return to the probability-based considerations detailed in the discussion of Objection #1. There it was claimed that, since the generalizations of an extension-based psychology are probabilistic in nature, the application of these generalizations gives the most reliable results when the extensions involved are highly confirmed. (In responding to Objection #1, we relied on the fact that the class of tigers was a highly confirmed extension for S's term 'tiger'. In contrast, hyena was a poorly confirmed extension of the same term because hyenas' high success rate relative to tiger was based on only one encounter between S and a hyena.) Such probability-based considerations are relevant here in the following way. Stimulus P1 is a type of proximal stimulus which can be physically specified, i.e., a specific pattern of retinal cell firings. Given the number of retinal cells of which an eye is composed, and the great variety of human experiences, P1 is a type of stimulus with which a given subject S is unlikely to have many encounters compared to S's total number of encounters with horses. Historically, phenomenalist reductions have faced the problem that there exists a wide variety of sensory stimuli associated with any given idea or mental term. The corresponding insight in the current case is that there are hundreds, or maybe thousands, of different patterns of stimuli which may have caused S to token 'horse'. For each of the proximal patterns that have caused S to token 'horse', S has causally interacted with it on only a very rare occasion. Even if BTT tells us that P1 is the extension of 'horse' for S, our probability-based considerations tell us to ignore that fact for the purposes of making accurate predictions and giving good explanations of S's relevant behavior. Stimulus P1 has a high success rate relative to S's LOT term 'horse', but it is based on very few, likely only one, encounters between S and P1. Horses also have a high success rate relative to S's term 'horse'. In contrast, though, horses' success rate is highly confirmed in that we can assume that S has had many encounters with horses.

Some might object to the assumption that S has had numerous encounters with actual horses. Everyone has to start learning a term at some point, and S's first encounter with a real horse will also be S's first encounter with some particular type of proximal stimulus (call it 'P2'). At this point in the learning process, S has interacted with members of P2 only once, and S has interacted with only one horse. After the initial encounter only, P2 and horses will have equal success rates (100%) relative to S's term 'horse'. How do we decide which part of the extension of 'horse' should guide prediction and explanation?

The sensible thing to do in this case would be to rely on past experience with humans. We face a case where two natural kinds have equal success rates, but we can be fairly confident that one of these kinds, horses, will very quickly emerge as the stable well-confirmed aspect of extension of 'horse'. S will have more encounters with horses in all likelihood, but S will not have many more, if any, encounters with P2. S is a standard human subject, and thus there is no reason to think that S will limit S's tokening of 'horse' to P2 and fail to treat newly perceived horses as horses.

Assume that S runs away in fright from the second horse S sees. How do we best explain S's behavior? Should we explain the behavior by saying that S was afraid of the horse or by saying that S was afraid of the proximal stimulus caused by the horse? Again, it seems only reasonable to base our explanations on what we know about humans in general. The average person who is afraid of horses has encountered many of them. Thus horses are the stable, well-confirmed element of the extension of the average horse-fearing person's LOT term 'horse'. We can see this stability in action when we note that if a person who is afraid of horses runs away from the horse, which stimulus caused the person to run away makes little explanatory difference. The stimulus which caused the person to run away was, in all likelihood, not even in the extension of that person's LOT term 'horse' prior to the event in question in virtue of the novelty of the stimulus. This justifies our saying that people are rarely afraid of types of proximal stimuli, while they are frequently afraid of horses. We can safely assume, then, that it is S's fear of horses, not the fear of a certain proximal stimulus, which caused S to run away on the occasion of S's second meeting with a horse. (There is no circularity involved here because the past experiences on which we are basing our assumptions about S are experiences of other subjects who do have well-confirmed extensions for the LOT term 'horse'.)

F. Objection Number Six

1. Nature won't cooperate

Many natural kinds, especially biological natural kinds, are not discrete. Thus the idea that various natural kinds, e.g., tiger and horse, have different success rates relative to a given LOT term is founded on an erroneous assumption, i.e., the assumption that there is a full complement of objectively definable natural kinds.

2. Discussion and reply

Chemical kinds and physical properties like mass pose some problems of indeterminacy. However, these problems have been dealt with fairly effectively elsewhere. (Cf. Maudlin 1986's discussion of the isotopes of gold [pp. 49-50] and Hartry Field's discussion of 'mass' [Field, 1973].) Recent attacks on the integrity of natural kinds have come primarily in the area of biological taxonomization, e.g., the separation of groups of animals into species. I focus on these attacks in what follows.

It has been claimed that treating species as natural kinds displays a failure to appreciate that the members of a given species lack any underlying shared essence. We should not think of species on the model of chemical elements, it's been said, because species are not definable in terms of their intrinsic, physical properties.[xx] The types classified under biological taxa are not static types and thus cannot figure into timeless laws of nature in the way that, for example, mineral kinds do. Dupre 1981 brings some of these concerns directly to bear on Putnam's theory of reference for natural kind terms, claiming that without underlying essential properties which are necessary and sufficient for inclusion in a species, Putnam's theory won't work. Dupre's criticism of Putnam's theory seems on the mark, at least with respect to the application of Putnam's theory to biological kind terms. In presenting his causal theory of reference for natural kind terms in a public language, Putnam assumes the importance of microstructural properties in defining biological kinds, e.g., the natural kind lemon (Putnam 1975, pp. 141-142). BUT IN PUTNAM'S DEFENSE HE MITIGATES THIS SUGGESTION BY GIVING THE STANDARD 'WHATEVER-SCIENCE-COMES-UP-WITH-AS-AN-ESSENCE-WILL-COUNT' LINE; SEE THE PARAGRAPH BEGINNING AT THE BOTTOM OF 140 Does Dupre's criticism extend to BTT?

It seems not. The Best Test Theory makes no mention of underlying defining microstructural properties. The theory is not committed to the shared microstructural properties view of natural kinds. Although some of the examples used in earlier chapters may seem to assume the shared microstructure view of natural kinds, all that is necessary for BTT to work is that there be natural kinds, however they might be defined by the relevant natural sciences.

It might be rejoined that species which lack microstructural essences cannot be defined in any other way. The most obvious alternative methods for defining natural kinds do not yield determinate, acceptable natural kinds. For example, species as defined by evolutionary biology are determined partly by ancestral relations. Species, so defined, are not the sorts of things to which scientific laws can make reference. Thus, the objection might run, such kinds are not truly natural kinds at all. As Maudlin correctly notes (Maudlin

1986, p. 123), this rejoinder depends on a view of scientific laws that is highly tendentious and not particularly secure. It was assumed in chapter I that the relevant natural sciences determine which natural kinds exist and how they are defined. No claims were made as to what the laws of those sciences must be like. Specifically, no claim was made that scientific laws must be syntactically characterizable (as Hull seems to assume [Maudlin 1986, pp. 120-121]), and nowhere was it claimed that scientific laws must involve only kinds which are timeless, unchanging, and specifiable without any reference to individuals or events extrinsic to those kinds (as is assumed by Lakoff's seemingly straw proponent of objective natural kinds [Lakoff 1987, chpt. 12, passim]).

Our line, then, is that whatever science gives us, BTT will use. So long as there are such things as species, and membership in species is a determinate matter, then the preceding observations about the nature of species are irrelevant to our evaluation of BTT. If the individual or sample in question has objective membership in a species, then the extension-fixing process can succeed. Essences are irrelevant.

Dupre's view of species still poses an apparent problem for BTT in that species, according to Dupre, do not have clearly demarcated boundaries. Members of a single species often vary with respect to genetic material as well as with respect to observed properties. There is some degree of reproductive isolation which characterizes species, but the wall against interbreeding is not inviolate. According to Dupre, all of these factors and more enter into the identification of a species.[xxi] Dupre claims that "the number of species can be a determinate matter, whereas the assignment of individuals to species may be only partially determinate." (Dupre 1981, p. 90) The basic problem is that while a species may be well-defined as a statistical clustering of properties, an individual may be somewhere in the no-man's land between two statistical clusterings. Dupre makes his point with respect to a group of closely related fruit fly species.

While recognizing the epistemological challenge of identifying the parents of a given fruit fly, one would think that a given fruit fly's lineage, and thus its species membership, is entirely determinate. If we find an individual fruit fly that, with respect to its phenotypic characteristics, lies between the statistical means for two species, it may be hard to tell of which species that fly is a member. However, the fruit fly still seems to have determinate membership in one natural kind, i.e., the kind to which its parents belonged. Dupre seems to infer metaphysical indeterminacy from epistemological disability.[xxii] At some point in the borderline fruit fly's history, some member of its ancestral lineage is certain to have not been precisely on the borderline between the two relevant species.[xxiii] Ancestry would then seem to provide all the reason we

need to view the fruit fly as belonging determinately to one species rather than the other (even if we don't know which one it is).

A final concern one might have about species is that even if there are well-defined biological natural kinds, they may not consist of the groups people normally take themselves to be referring to. For example, Gould 1983 challenges the unity of the pretheoretical natural kind zebra. According to at least one method of determining species, the mountain zebra is more closely related to horses than it is to the other two species of zebras (Burchell's zebra and Grevy's zebra) (Gould 1983, p. 361). Dupre also provides numerous examples of ways in which biological taxonomization may not match up with the way people separate the world, manifest in the way people apply natural language terms (Dupre 1981, pp. 74-79).

Principles BT2 and BT4 are sufficient to fix extensions in a satisfactory way in the kinds of cases presented by Gould and Dupre. Take the zebra case. Even if the hypothesis Gould suggests is correct (an hypothesis which originates from Debra K. Bennett), reference to zebras, and only to zebras, can be achieved via BT2. The subject specifies in her kind-minded intentions the properties of the sample that are relevant to grounding the LOT term 'zebra'. By intending that 'zebra' refer, for example, to the natural kind membership in which gives zebra their stripes, the subject determines the extension of 'zebra' to be the disjunctive set of the members of all three of the zebra species, i.e., all three of the natural kinds which meet the descriptive specification. In cases where subject wants to use a term in the way that an expert would, no matter how strange the result, BT4 is operative.

G. Objection Number Seven

1. Terms without extensions have referential content

If extension is prior to referential content, how come terms like 'unicorn' seem to have referential contents, but have empty extensions?

2. Discussion and reply

Terms that have, and always have had, empty extensions, are introduced by description. I will not argue here that it is impossible to have a semantically primitive term in LOT which has, and always has had, an empty extension. There just aren't any, so far as I can tell; and psychology isn't obliged to give explanations of what isn't. The most obvious candidates for being terms which don't have, and have never had, extensions are LOT terms like 'unicorn', 'witch', 'god', and 'soul'. However, we associate all of these

terms with rich descriptions. These descriptions themselves are constructed out of LOT terms which have, or have had, actual extensions. Thus, extension is prior to referential content for such terms as 'unicorn'. We couldn't assign referential content to 'unicorn' at all, if it weren't for our knowledge of the extensions of the constituents of the descriptions by which 'unicorn' is introduced and its use sustained.

In the preceding, I spoke of terms which have, and always have had, empty extensions. I did so with two other kinds of cases in mind. In both of these cases, to be described straightaway, we seem to have LOT terms which are presently void of extension, but whose referential contents are not calculated solely from the extensions of the terms which comprise the descriptions associated with these extensionless terms. Firstly, consider innate representations with no current extensions, but which once had extensions. If there are such things, we may want to say that their referential contents now are determined by BTT applied to our relevant ancestors (those who lived when the terms actually had extension) to yield actual extensions. Referential contents would then be characterized by reference to the extensions as they were back when the innate representations were selected for, thus becoming innate.

The second case I have in mind is the case of extinct species. In the case of, say, the LOT term 'dodo bird', we characterize the referential content of the term in the way that we do because of certain relations that hold through time. The description of dodos that we associate with our LOT term 'dodo' is composed of terms with the same referential content as those terms which constituted the concept which our relevant ancestors associated with an LOT term that had dodos as its extension (as assigned by BTT).

As presented, BTT is a narrow theory. It is, first and foremost, a theory of extension for the natural kind terms of LOT. In the preceding, I hope to have made plausible the claim that there is a well-defined class of natural kind terms in LOT, and that BTT provides satisfactory results as to what the extensions of those terms are. Along the way, I have shown how BTT can be applied to solve the disjunction problem for natural kind terms in a given subject's LOT. Additionally, it is hoped that the reader has gained some impression of how we might apply BTT's central ideas to other types of LOT terms; either by arguing that all reference can be ultimately traced back to reference to natural kinds, or by constructing new extension-fixing principles analogous to those of BTT, but which make no essential mention of natural kinds.

Notes to Chapter VI

---

[i] Strictly speaking, BTT does not lead to any specific psychological predictions. It must be supplemented, at least, by a theory which explains the role of extension in psychology. I offer no such theory here. In chapter I, I introduced examples of cases where extension seems to be invoked in psychological explanation. It is in the same spirit that I discuss here the use of extensions in the prediction and explanation of behavior.

[ii] One way for BTT to avoid the tiger/hyena problem would be to defer to intensional psychology, and say that extensions are not relevant to the prediction of S's future behavior. However, the situation as described seems to be a situation in which extensionally-based explanations and predictions are likely to be offered by folk psychology. Why did S exhibit fear-of-tigers behavior around tigers in the past? Because, the folk-psychological explanation may well run, S is afraid of tigers. Why did S run away when S saw hyena? Because, the folk would say, S thought that he saw a tiger. To say that S thought that he saw a tiger is to give an extensional explanation. S's mental state is characterized referentially, i.e., in terms of what the state (or one of its components) normally refers to. Deferring to intensional psychology to solve the tiger/hyena problem thus seems unjustified. Besides, doing so would set an awful precedent for responding to objections.

[iii] Notice, however, that without such a concept in place, we may lose the conviction that t is misapplied in the case in question.

[iv] We needn't worry much about this number increasing. Given that S has a standard tiger concept, even a nominal number of future interactions between S and hyenas is almost certain to reinstate tigers as the extension of S's LOT term 'tiger'. There exists the alternative possibility that S would continue classifying hyenas under the heading 'tiger' indefinitely because, for example, S can't tell the two species apart. However, such a case would provide an example of an entirely different sort than the one under discussion. Examples where a subject cannot tell two species apart are discussed below under the heading 'Objection #2'.

[v] Notice that if S does not have a concept of unfamiliar species whose content has been determined by past encounters with members of unfamiliar species, then there would be little grounds for the claim that our psychological theory should predict that S will behave in a cautious but curious manner when S meets another hyena.

[vi] This maneuver may sound suspiciously like the DR theory of primary extension, an approach dismissed in chapter V. Notice, though, that the tiger/hyena case is different than the examples discussed in chapter V. A DR theory of primary reference must apply to every natural kind term that the subject has ever mistakenly tokened. For a DR theory, extension is disjunctive for virtually every natural kind term in LOT. In order to make consistently accurate predictions of S's behavior, one must know the primary reference of virtually all of the natural kind terms in S's LOT. In contrast, the statistics-based maneuver described in the text only applies to cases where we have truly disjunctive reference, i.e., where there is a grouping of success rates at the top of the list of success rates relative to a given term (see the discussion of Objection #2 below). Even if the DR strategy of calculating primary reference does not seem different in principle than the approach to prediction advocated in the text, recall that DR approaches were rejected for a number of reasons, not solely because they required a theory of primary reference.

[vii] Note a special aspect of the tiger/hyena case that distinguishes it from typical cases. In the typical case, the extension of a subject's natural kind term of LOT will consist of the members of a natural kind which has a success rate that is much higher than the success rate of any other natural kind. But in the tiger/hyena case, the relevant success rates are 1 and 0.95. This is quite a small gap between success rates. I claim below that a natural kind term t has truly disjunctive reference in cases where two natural kinds have, equal, or roughly equal, success rates relative to t (and where both of these success rates are far higher than the success rate of any other natural kind relative to t). Thus, tigers never actually leave the extension of S's LOT term 'tiger', not even in the period immediately following S's first encounter with a hyena.

On the disjunctive view, tigers never leave the extension of 'tiger', both tigers and hyenas are in the extension of 'tiger' after S's first meeting with a hyena. Considerations of sample size guide our decision regarding which element of this disjunctive extension of 'tiger' to pay attention to when predicting S's behavior. Tigers have caused numerous of S's 'tiger' tokens, and hyenas have only caused one such tokening. For the purpose of predicting S's behavior, then, tigers would have a privileged place in the disjunctive extension of S's term 'tiger'.

[viii] This is especially important given that the tiger/hyena objection is based on intuitions about what 'tiger' should refer to. These intuitions may not be very strong in a case where the LOT term in question is subject to BT1 and thus has no compound concept associated with it. In the tiger/hyena objection, what seems to be doing most of the work is the assumption that S has a standard tiger concept, which means that BT2 applies to 'tiger', not BT1.

[ix] I talk throughout of the counterfactual-based version of BTT. In reality, I am offering a counterfactual-based version of the success rate function. BTT's principles themselves need not be altered at all. But in so far as these principles all make reference to success rates, once one changes the way success rates are calculated, the entire theory becomes counterfactually-based in an obvious way.

[x] Admittedly, talking of the counterfactual hyenas sounds strange. How do we know what a counterfactual hyena looks like, sounds like, etc.? To be precise here, every member of the species hyena is tested in each of the relevant counterfactual situations to determine the success rate of hyena relative to S's LOT term 'tiger'. Each counterfactual situation yields a percentage equal to the percentage of times that a counterfactual hyena caused the tokening of 'tiger' in that counterfactual situation (of all the counterfactual hyenas run through the single counterfactual testing situation, we find the percentage of them that caused S to token 'tiger'). All of the relevant percentages, one yielded by each of the relevant counterfactuals, are then averaged together to give us the counterfactual-based success rate of hyenas relative to S's LOT term 'tiger'.

[xi] Such results will hold no matter what you replace S's heart with in TC. If, for example, you replace S's heart with a tiger in TC, S will die (without tokening 'tiger', most likely, but it's hard to say).

[xii] Assume that S does token 'tiger' when S's friend is replaced by a hyena, because, for example, the hyena attracts the tigers attention, and the tiger comes running toward S. Such results are not likely to substantially increase the success rate of hyenas relative to 'tiger', if for no other reason than that when the hyena is substituted for the friend, the hyena will likely cause S to token other LOT terms. And when the hyena causes S to token these other LOT terms, the success rate of hyenas relative to S's LOT term 'tiger' is lowered as a consequence.

[xiii] Two important qualifications included in the discussion of BT1, and in the statement of the Objection #2, are assumed to apply throughout the remainder of chapter VI. Whenever I talk of two kinds having equal success rates, the points made also apply to cases where more than two kinds have equally high success

rates. Furthermore, such cases are only interesting when the kinds that have equal or roughly equal success rates are at the top of the heap. Therefore, I will not continue to point out that, in the cases of interest, the equal or roughly equal success rates relative to t are substantially higher than any other natural kind's success rate relative to t.

[xiv] Suggestions along these lines are originally due to Putnam and his hypothesis of the division of linguistic labor, although Putnam is concerned with natural language, not LOT (Putnam 1975, pp. 227-229). For an illuminating discussion of the nature of deference to experts, see Fodor 1994, pp. 33-39. According to Fodor, when we defer to an expert, we are, in effect, intending to use the expert as an instrument to get us into the right causal relation with the natural kind to which we want our term to refer.

[xv] Similar remarks seem to apply to Quine's much discussed Ortcutt example (Quine 1956), which is often taken to illustrate the poverty of extensional psychology. The Ortcutt case bears a deep resemblance to the gold/fool's gold case. In the Ortcutt case, we have two different terms being tokened in two different situations. These two terms ('man at the beach' and 'man who looks like a spy') have the same extension but lead to different types of behavior. Why? The reason for S's differential behavior should be explained not by citing the sameness of extension for 'spy guy' and 'beach guy'. This only leads to puzzlement. The difference in S's behavior is to be explained by the extensions of other terms that S tokens (if it is to be explained by citing extensional contents at all). The reason S treats Ortcutt differently on different occasions is because 'shady', 'dressed in black', and 'suspicious' have extensions which trigger in S different behavior than do the extensions of 'upstanding', 'happy', and 'plays with his family'.

[xvi] A more rigorous approach would be to apply a specific mathematical test for statistical significance. However, I think my point can be made without discussion of ANOVAs or the like.

[xvii] There are complications here. See the discussion of 'unicorn' below. Such complications, however, do not affect what is said here in response to Objection #5.

[xviii] The name of Objection #2, 'BTT leads to phenomenalism', may seem to misrepresent Objection #2 as it is developed above. In contrast to our characterization of proximal stimuli as physical types, a common version of phenomenalism holds that the reference of a term (in LOT or in a public language) is a class of sense-data. On this view, sense-data are ideas presented to the mind, rather than physical, natural kinds of proximal stimuli. With respect to such insubstantial phenomenal kinds, we can dismiss them as being inappropriate candidates for reference when applying BTT. If these insubstantial sense-data make up natural kinds of use to psychology, they are kinds which play a role in an intensional psychology. As noted in chapter I, BTT may give us some idea what the intension of a natural kind term is, but BTT does so by looking at the causal interactions between subjects and individual members of kinds. Insofar as sense-data have a mysterious or mentalistic nature, we can dismiss them because they do not enter into the same types of causal interactions the individual physical members of natural kinds do. On the other hand, if sense-data correspond to types of physical, proximal stimuli, then we face Objection #2 as it is framed in the text, i.e., as a problem regarding the success rates of physically specifiable natural kinds of proximal stimuli.

[xix] In some cases BTT does not include P1s as part of the extension of 'horse', and any problem caused by Objection #5 is solved. For example, it's possible that a subject other than S, call her 'D', would experience P1 (i.e., a proximal stimulus of the same natural kind as that experienced by S in the example given in the text), and that through D's behavior (e.g., her saying, "Lo, a cow"), D's P1 could have some effect on S. If P1, through D's behavior, causes S to token an LOT term other than 'horse', e.g., 'cow', this lowers the success rate of the natural kind P1 relative to S's LOT term 'horse'. If this happens often enough, P1's success rate will become so low relative to S's LOT term 'horse' that P1 will no longer be in the extension of S's LOT term 'horse'. Unfortunately, this type of series of events seems as if it would be rather rare. Thus, I proceed in the

text to give a response to Objection #5 which assumes that P1s are in the extension of the LOT term 'horse' for the average subject.

[xx] David Hull is one of the foremost architects of this type of criticism. For an approving review of Hull's arguments, see Lakoff 1987, chapter 12. For a not-so-approving review, see Maudlin 1986, pp. 119-126.

[xxi] In biology itself, there exist three opposing schools of thought regarding the nature of species. Cladistics, phenetics, and evolutionary systematics all lay some claim to defining species, but all on differing grounds. (Cf. Gould 1983, Maudlin 1986, and Lakoff 1987, for descriptions of the ways in which the three methods differ.)

[xxii] The charge of confusing epistemology with metaphysics is a serious one. However, there is some evidence in Dupre's article that his worries about the difficulty in divining ancestral relations have metaphysical implications. At one point in his search for objective defining characteristics of species, Dupre considers focusing solely on evolutionary history as the basis for species differentiation. He rejects this option largely on the grounds that we could never construct an entire evolutionary history of the earth's species (Dupre 1981, p. 87). However, Dupre does not make clear the relevance of our limited human abilities to the metaphysical question of species individuation. If ancestral relations exist, and are sufficient to demarcate species, then our inability to discern these ancestral relations has no bearing on the metaphysical determination of the species.

One important aspect of Putnam's theory is the idea of the linguistic division of labor, whereby one can defer to experts when determining the reference of a term. If one interprets Putnam to be saying that reference to a natural kind always requires an expert somewhere who is able to define the natural kind in question, then Dupre's epistemological point about our inability to identify all species may be relevant to an evaluation of Putnam's theory of reference. However, such an interpretation of Putnam seems totally unwarranted. Putnam gives no indication that people couldn't determinately refer to water until some experts found out that water was H2O. In fact, Putnam claims precisely the opposite (Putnam 1975, p. 224). Furthermore, none of the principles BT1, BT2, or BT3 assumes that there must be experts who can recognize and define a given kind in order for a natural kind term of LOT to refer to that kind.

[xxiii] If all of a given fruit fly's ancestors had been precisely on the border between two species, that would seem to be adequate grounds for identifying the fly and its ancestors as a new species, resolving our problem altogether. Even if this hypothetical situation would not provide reason to introduce a new species, the incredible unlikelihood of such a situation ever obtaining seems reason enough for the natural sciences to dismiss concerns about its actuality.

References

Antony, L. and Levine, J. (1991). 'The Nomic and the Robust'. In Loewer, B. and Rey, G. 1991.
Armstrong, S. L., Gleitman, L. R., and Gleitman, H. (1983). 'What Some Concepts Might Not Be'. Cognition, 13, pp. 263-308.
Baker, L. R. (1991). 'Has Content Been Naturalized?' In Loewer, B. and Rey, G. 1991, pp. 17-32.
Berlin, B. and Kaye, P. (1969). Basic Color Terms. Berkeley: University of California Press.
Biederman, I. (1990). 'Higher-Level Vision', In Osherson, Kosslyn, and Hollerbach, 1990
Blackburn, S. (1984). Spreading the Word. Oxford, UK: Clarendon Press.
Block, N. (ed.). (1981). Imagery. Cambridge, MA: MIT Press.
Block, N. (1986). 'Advertisement for a Semantics for Psychology'. In French, P., Euhling, T., and Wettstein, H. (eds.), Studies in the Philosophy of Mind. Vol. 10, Midwest Studies in Philosophy. Minneapolis, MN: University of Minnesota Press (1986).
Boghossian, P. A. (1991). 'Naturalizing Content'. In Loewer and Rey, 1991.
Bower, T. G. R. (1989). The Rational Infant: Learning in Infancy. New York, NY: W. H. Freeman and Company.
Brown, R., and Herrnstein, R. J. (1981). 'Icons and Images'. In Block, 1981.
Burge, T. (1986). 'Individualism and Psychology'. Philosophical Review, XCV, pp. 3-45.
Campbell, K. (1982). 'The Implications of Land's Theory of Color Vision'. In Logic, Methodology and Philosophy of Science VI. Proceedings of the Sixth International Congress of Logic, Methodology and Philosophy of Science, Hanover 1979. North Holland Publishing Company and PWN-Polish Scientific Publishers 1982. (Reprinted in Lycan 1990).
Carey, S. (1985). Conceptual Change in Childhood. Cambridge, MA: MIT Press.
Carey, S. (1990). 'Cognitive Development'. In Osherson and Smith, 1990.
Carey, S., and Gelman, R. (eds.). (1991). The Epigenesis of Mind. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
Cartwright, N. (1983). How the Laws of Physics Lie. Oxford, UK: Oxford University Press.
Cherniak, C. (1986). Minimal Rationality. Cambridge, MA: MIT Press
Churchland, P. M. (1981). 'Eliminative Materialism and Propositional Attitudes'. Journal of Philosophy, 78, pp. 67-90. (Reprinted in Lycan, 1990).
Churchland, P. M. (1986). 'Some Reductive Strategies in Cognitive Neurobiology'. In Mind, vol. 95, no. 379. (Reprinted in Silvers, 1989).
Churchland, P. M., and Churchland, P. S. (1983). 'Stalking the Wild Epistemic Engine', In Nous, 17, pp. 5-18. (Reprinted in Lycan, 1990).
Clark, A. (1989). Microcognition. Cambridge, MA: MIT Press.
Clark, A. (1991). 'In Defense of Explicit Rules'. In Ramsay, Stich, and Rumelhart 1991.
Cohen, L. J. (1986). 'How is Conceptual Innovation Possible?'. Erkenntnis, 25, pp. 221-238.
Cram, H. (1992). 'Fodor's Causal Theory of Representation'. The Philosophical Quarterly, Vol. 42, No. 166, pp. 56-70.
Cummins, R. (1986). 'Inexplicit Information', in Brand, M. and Harnish, M. (eds.), Problems in the Representation of Knowledge and Belief, University of Arizona Press, pp. 116-126.
Cummins, R. (1989a). 'Representation and Covariation'. In Silvers 1989.
Cummins, R. (1989b). Meaning and Mental Representation. Cambridge, MA: MIT Press.
Cummins, R. (1991). 'The Role of Mental Meaning in Psychological Explanation'. In McLaughlin, B. 1991.
Davidson, D. (1986). 'Knowing One's Own Mind'. In the Proceedings of the American Philosophical Association, Vol. 60, pp. 441-458.
Davies, M. (1991). 'Concepts, Connectionism, and the Language of Thought'. In Ramsey, W., Stich, S., and Rumelhart, D. 1991.
Davies, M. (1995). 'Two Notions of Implicit Rules', in Tomberlin, J. (ed.), Philosophical Perspectives, 9: AI, Connectionism, and Philosophical Psychology, Atascadero, CA: Ridgeview, pp. 153-183.
Dennett, D. (1978). 'A Cure for the Common Code?'. In D. Dennett, Brainstorms. Montgomery, VT: Bradford Books.
Dennett, D. (1981). 'Two Approaches to Mental Images'. In Block, 1981.
Dennett, D. (1982). 'Beyond Belief'. In A. Woodfield (Eds.), Thought and Object, Oxford, U.K.: Clarendon Press, 1982. Reprinted in Dennett, 1987.
Dennett, D. (1987). The Intentional Stance. Cambridge, MA: MIT Press.
Dennett, D. (1991). Consciousness Explained. Boston, MA: Little, Brown, and Company.

Devitt, M. (1989). 'A Narrow Representational Theory of Mind'. In Silvers 1989.
Devitt, M., and Sterelny, K. (1987). Language and Reality. Cambridge, MA: MIT Press.
Diamond, A. (1991). 'Neurophysiological Insights into the Meaning of Object Concept Development'. In Carey and Gelman, 1991.
Dickson, T. R., (1983). Introduction to Chemistry, fourth edition. New York, NY: John Wiley and Sons.
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. Oxford: Oxford University Press. (Reprinted in Lycan, 1990.)
Dretske, F. (1988). Explaining Behavior. Cambridge, MA: MIT Press.
Dretske, F. (1990). 'Seeing, Believing, and Knowing'. In Osherson, Kosslyn, and Hollerbach, 1990.
Dupre, J. (1981). 'Natural Kinds and Biological Taxa'. Philosophical Review, XC, No. 1.
Edelberg, W. (1986). 'A New Puzzle about Intentional Identity'. Journal of Philosophical Logic, 15, pp. 1-25.
Evans, G. (1973). 'The Causal Theory of Names'. Aristotelian Society Supplementary Vol. 47. (Reprinted in Martinich, 1985).
Field, H. (1973). 'Theory Change and the Indeterminacy of Reference'. Journal of Philosophy, LXX (14): 462-481.
Finke, R. (1980). 'Levels of Equivalence in Imagery and Perception'. Psychological Review, Vol. 87, pp. 113-132.
Fodor, J. (1975). The Language of Thought. Cambridge, MA: Harvard University Press.
Fodor, J. (1980). 'Methodological Solopsism Considered as a Research Strategy in Cognitive Psychology'. Behavioral and Brain Sciences, Vol. 3, No. 1. (Reprinted in Fodor, 1981).
Fodor, J. (1981). Representations. Cambridge, MA: MIT Press.
Fodor, J. (1983a). 'Observation Reconsidered'. Philosophy of Science, 51, pp. 23-43.
Fodor, J. (1983b). The Modularity of Mind. Cambridge, MA: MIT Press.
Fodor, J. (1987). Psychosemantics. Cambridge, MA: MIT Press.
Fodor, J. (1990a). 'Psychosemantics or: Where do Truth Conditions Come From?'. In Lycan 1990.
Fodor, J. (1990b). A Theory of Content. Cambridge, MA: MIT Press.
Fodor, J. (1994). The Elm and the Expert. Cambridge, MA: MIT Press.
Fodor, J., and Pylyshyn, Z. (1988). 'Connectionism and Cognitive Architecture: A Critical Analysis'. Cognition, 28, pp. 3-71.
Forster, K. (1990). 'Lexical Processing'. In Osherson, D. and Lasnick, H.1990, pp. 95-131.
Frege, G. (1980). 'On Sense and Meaning'. In Geach, P, and Black, M. Translation from the Philosophical Writings of Gottlob Frege, third edition. Oxford: Basil Blackwell. (Reprinted in Martinich, 1985).
Geach, P. T. (1967). 'Intentional Identity'. In Journal of Philosophy, Vol. LXIV, No. 20.
Giere, R. N. (1988). Explaining Science: A Cognitive Approach. Chicago, IL: University of Chicago Press.
Goldman, A. (1976). 'Discrimination and Perceptual Knowledge'. Journal of Philosophy, 73, pp. 771-791.
Goldman, A. (1986). Epistemology and Cognition. Cambridge, MA: Harvard University Press.
Goodman, N. (1983). Fact, Fiction, and Forecast, fourth edition. Cambridge, MA: Harvard University Press. (First published, 1955)
Goschke, T. and Koppelberg, D. (1991). 'The Concept of Representation and the Representation of Concepts in Connectionist Models'. In Ramsey, Stich, and Rumelhart 1991, pp. 129-161.
Gould, S. J. (1983). 'What, if Anything, is a Zebra?', in S. J. Gould, Hen's Teeth and Horse's Toes, New York, NY: W. W. Norton and Company, 1983.
Hacking, I. (1983). Representing and Intervening. Cambridge, UK: Cambridge University Press.
Hardin, C. L. (1990). 'Color and Illusion'. In Lycan, 1990.
Hatfield, G. (1991). 'Representation in Perception and Cognition: Connectionist Affordances'. In Ramsey, Stich, and Rumelhart 1991, pp. 163-195.
Haugeland, J. (1981). 'Semantic Engines: An Introduction to Mind Design'. In J. Haugeland (ed.), Mind Design: Philosophy, Psychology, Artificial Intelligence. Montgomery, VT: Bradford Books.
Hinton, G.E., McClelland, J. L., and Rumelhart, D. E. (1986). 'Distributed Representations'. In Rumelhart, et. al., 1986.
Horgan, T. and Woodward, J. (1985). 'Folk Psychology is Here to Stay'. Philosophical Review, XCIV, No. 2. (Reprinted in Lycan, 1990).
Inhelder, B. and Piaget, J. (1964). The Early Growth of Logic in the Child. New York: Norton.
Jackendoff, R. (1989). 'What is a Concept, that a Person May Grasp It?'. Mind and Language, 4, pp. 68-102.
Johnson-Laird, P. N. (1983). Mental Models. Cambridge, MA: Harvard University Press.

Kahneman, D., Slovic, P., and Tversky, A. (eds.). (1982). Judgement Under Uncertainty: Heuristics and Biases. Cambridge, UK: Cambridge University Press.
Katz, J. J. (1995). 'What Mathematical Knowledge Could Be'. Mind, Vol. 104, No. 415, pp. 491-522.
Keil, F. (1989). Concepts, Kinds, and Cognitive Development. Cambridge, MA: MIT Press.
Kosslyn, S. (1981a). Image and Mind. Cambridge, MA: Harvard University Press.
Kosslyn, S. (1981b). 'The Medium and the Message in Mental Imagery'. In Block, 1981.
Kosslyn, S. (1983). Ghosts in the Mind's Machine. New York, NY: Norton.
Kosslyn, S. (1990). 'Mental Imagery'. In Osherson, Kosslyn, and Hollerbach, 1990.
Kripke, S. (1980). Naming and Necessity. Cambridge, MA: Harvard University Press. (First published, 1972).
Kuhn, T. S. (1962). The Structure of Scientific Revolutions. Chicago, IL: University of Chicago Press.
Lakoff, G. (1987). Women, Fire, and Dangerous Things: What Categories Reveal About the Mind. Chicago, IL: University of Chicago Press.
Landau, B., and Gleitman, L. (1985). Language and Experience. Cambridge, MA: Harvard University Press.
Loewer, B and Rey, G (eds.) (1991). Meaning in Mind: Fodor and his Critics. Oxford, UK: Blackwell.
Lycan, W. G. (ed.). (1990). Mind and Cognition: A Reader. Oxford, UK: Basil Blackwell.
Macnamara, J. (1982). Names for Things. Cambridge, MA: MIT Press.
Manfredi, P. A. and Summerfield, D. M. (1992). 'Robustness without Asymmetry: A Flaw in Fodor's Theory of Content'. Philosophical Studies, Vol. 66, pp. 261-283. Dordrecht, Netherlands: Kluwer.
Markman, E. M. (1989). Categorization and Naming in Children. Cambridge, MA: MIT Press.
Marr, D. (1982). Vision. New York, NY: W. H. Freeman and Company.
Martinich, A. P. (ed.) (1985). The Philosophy of Language. Oxford, UK: Oxford University Press.
Maudlin, T. (1986). Reasonable Essentialism & Natural Kinds. Doctoral Dissertation, University of Pittsburgh.
McKay, T., and Stern, C. (1979). 'Natural Kinds and Standards of Membership'. Linguistics and Philosophy, 3, pp. 27-34.
McClamrock, R. (1995). Existential Cognition, Chicago, IL: University of Chicago Press.
McClelland, J. L., Rumelhart, D. E., and Hinton, G. E. (1986). 'The Appeal of Parallel Distributed Processing'. In Rumelhart, et. al., 1986.
McLaughlin, B. P. (ed.) (1991). Dretske and His Critics. Oxford, UK: Basil Blackwell.
Millikan, R. G. (1984). Language, Thought, and Other Biological Categories. Cambridge, MA: MIT Press.
Millikan, R. G. (1991). White Queen Psychology and Other Essays for Alice. Cambridge, MA: MIT Press.
Montgomery, R. (1989). 'Discrimination, Reidentification and the Indeterminacy of Early Vision'. Nous 23, pp. 413-435.
Murphy, G. L. (1982). 'Cue Validity and Levels of Categorization'. Psychological Bulletin, 91, pp. 174-177.
Murphy, G. L. (1988). 'Comprehending Complex Concepts'. Cognitive Science, 12, pp. 529-562.
Murphy, G. L., and Medin, D. L. (1985). 'The Role of Theories in Conceptual Coherence'. Psychological Review, 92, pp. 289-316.
Nisbett, R., and Wilson, T. (1977). 'Telling More Than We Can Know: Verbal Reports on Mental Processes'. Psychological Review, 84, 231-259.
Osherson, D. N., Kosslyn, S. M., and Hollerbach, J. M. (eds.) (1990). Visual Cognition and Action: An Invitation to Cognitive Science, Vol. 2. Cambridge, MA: MIT Press.
Osherson, D. N., and Smith, E. E. (eds.) (1990). Thinking: An Invitation to Cognitive Science, Vol. 3. Cambridge, MA: MIT Press.
Peacocke, C. (1992). A Study of Concepts. Cambridge, MA: MIT Press.
Pinker, S., and Prince, A. (1988). 'On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language Acquisition'. Cognition 28, pp. 73-193.
Putnam, H. (1975). Mind, Language, and Reality. Cambridge, UK: Cambridge University Press.
Pylyshyn, Z. (1981). 'The Imagery Debate: Analog Media versus Tacit Knowledge'. In Block, 1981.
Pylyshyn, Z. (1984). Computation and Cognition. Cambridge, MA: MIT Press.
Quine, W. V. (1953). From a Logical Point of View. Cambridge, MA: Harvard University Press.
Quine, W. V. (1956). 'Quantifiers and Propositional Attitudes'. Journal of Philosophy, 53, pp. 177-187. (Reprinted in Martinich, 1985).
Quine, W. V. (1960). Word and Object. Cambridge, MA: MIT Press.
Ramsey, W., Stich, S. P., and Rumelhart, D. E. (eds.) (1991). Philosophy and Connectionist Theory. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Reed, S. K. (1972). 'Pattern Recognition and Categorization'. Cognitive Psychology, 3, pp. 382-407.
Rey, G. (1983). 'Concepts and Stereotypes'. Cognition, 15, pp. 237-262.
Rosch, E., and Mervis, C. (1975). 'Family Resemblances: Studies in the Internal Structure of Categories'. Cognitive Psychology, 7, pp. 573-605.
Rosch, E., Mervis, C., Gray, W., Johnson, D., and Boyes-Braem, P. (1976). 'Basic Objects in Natural Categories'. Cognitive Psychology, 8, pp. 382-439.
Rumelhart, D. E., Hinton, G. E., and McClelland, J. L. (1986). 'A General Framework for Parallel Distributed Processing'. In Rumelhart, et. al., 1986.
Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). 'Learning Internal Representations by Error Propagation'. In Rumelhart, et al., 1986.
Rumelhart, D. E., and McClelland, J. L. (1986). 'PDP Models and General Issues in Cognitive Science'. In Rumelhart, et. al., 1986.
Rumelhart, D. E., McClelland, J. L., and the PDP Research Group (1986). Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1, Foundations. Cambridge, MA: MIT Press.
Samet, J., and Flanagan, O. (1989). 'Innate Representations'. In Silvers, 1989.
Schwartz, S. (1978). 'Putnam on Artifacts'. Philosophical Review, LXXXVII, pp. 566-574.
Schwartz, S. (1979). 'Natural Kind Terms'. Cognition, 7, pp. 301-315.
Searle, J. (1980). 'Minds, Brains, and Programs'. Behavioral and Brain Sciences, vol. 3.
Shapiro, L. A. (1993). 'Content, Kinds, and Individualism in Marr's Theory of Vision'. Philosophical Review, Vol. 102, No. 4, pp. 489-513.
Shepard, R., and Cooper, L. (1982). Mental Images and Their Transformations. Cambridge, MA: MIT Press.
Silvers, S. (ed.). (1989). Rerepresentation: Readings in the Philosophy of Mental Representation. Dordrecht, The Netherlands: Kluwer Academic Publishers.
*Simon, H. (....). 'On the Forms of Mental Representation'. In Posner ? (see Smith 1989).
Slote, M. A. (1966). 'The Theory of Important Criteria'. Journal of Philosophy, LXIII, 8, pp. 211-224.
Smith, A. D. (1990). 'Of Primary and Secondary Qualities'. Philosophical Review, XCIX, pp.221-254.
Smith, E. E. (1989). 'Concepts and Induction'. In M. I. Posner (ed.), Foundations of Cognitive Science, Cambridge, MA: MIT Press.
Smith, E., and Medin, D. (1981). Categories and Concepts. Cambridge, MA: Harvard University Press.
Smith, E. E., Osherson, D. N., Rips, L. J., and Keane, M. (1988). 'Combining Prototypes: A Selective Modification Model'. Cognitive Science, 12, pp. 485-527.
Smolensky, P. (1988). 'On the Proper Treatment of Connectionism'. Behavioral and Brain Sciences, 11, pp. 1-74.
Spelke, E. S. (1990). 'Origins of Visual Knowledge'. In Osherson, Kosslyn, and Hollerbach 1990, pp. 99-127.
Spelke, E. S. (1991). 'Physical Knowledge in Infancy: Reflections on Piaget's Theory'. In Carey, S. and Gelman, R. 1991, pp. 133-169.
Sterelny, K. (1990). The Representational Theory of Mind. Oxford, UK: Basil Blackwell.
Stich, S. (1978). 'Beliefs and Sub-Doxastic States'. Philosophy of Science, 45, pp. 499-518.
Stich, S. (1981). 'Dennett on Intentional Systems'. Philosophical Topics, 12, pp. 38-62.
Stich, S. (1983). From Folk Psychology to Cognitive Science. Cambridge, MA: MIT Press.
Stich, S. (1990). 'Rationality'. In Osherson and Smith 1990.
Stillings, N. A., Feinstein, M. H., Garfield, J. L., Rissland, E. L., Rosenbaum, D. A., Weisler, S. E., and Baker-Ward, L. (1987). Cognitive Science: An Introduction. Cambridge, MA: MIT Press.
Teller, P. (1989). 'Relativity, Relational Holism, and the Bell Inequalities'. In James T. Cushing and Ernan McMullin (eds.), Philosophical Consequences of Quantum Theory. South Bend, IN: University of Notre Dame Press.
Van Gelder, T. (1990). 'Compositionality: A Connectionist Variation on a Classical Theme'. Cognitive Science, 14, pp. 355-384.
*Van Gelder, T. (1991). 'Connectionism and Dynamical Explanation'. Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society.
Wittgenstein, L. (1953). Philosophical Investigations, translated by G. E. M. Anscombe. New York, NY: Macmillan.

Yuille, A. L., and Ullman, S. (1990). 'Computational Theories of Low-Level Vision'. In Osherson, Kosslyn, and Hollerbach, 1990.