The Best Test Theory of Extension: First Principle(s)

Robert D. Rupert (appears in *Mind and Language*, Sept. 1999)

1. The Best Test Theory and BT1

1.1 Motivation for the Best Test Theory: Misrepresentation and the Disjunction Problem

Over the past two decades, naturalistically[1] minded philosophers have expended a fair amount of

---

This paper presents the leading idea of my doctoral dissertation and thus has been shaped by the reactions of all the members of my thesis committee: Charles Chastain, Walter Edelberg, W. Kent Wilson, Dorothy Grover, and Charles Marks. I am especially grateful for the help of Professors Chastain, Edelberg, and Wilson; each worked closely with me at one stage or another in the development of the ideas contained in the present work. Shorter versions of this paper were presented at the 47th Annual Northwest Conference on Philosophy (1995), the 1996 Mid-South Philosophy Conference, the 1997 meeting of the Central Division of the American Philosophical Association, and at the University of Washington, Seattle; thanks to all audiences for their insightful comments and questions and also to my conference commentators, Eric Gampel, Jonathan Cohen, and Bruce Glymour, respectively, each of whom offered a thoughtful critique. Lastly, I extend my gratitude to anonymous referees, including two from *Mind and Language*, whose remarks led to significant improvements in the paper.

Address for correspondence: 13416, 4th Ave. S., Seattle, WA 98168, USA

E-mail: *robrupert@aol.com*

---

[1] When using 'naturalistic' and its cognates, I have in mind the view that the methods employed in the natural sciences are more likely than any others to yield knowledge. On this view, philosophy's primary task is to help lay the conceptual foundations for natural science in an interactive way, i.e., in a way that is sensitive to the aims and results of the sciences.

energy trying to explain how extensions are fixed for human beings' natural kind concepts (Fodor 1987, 1990, Dretske 1981, 1988, Millikan 1984, Sterelny 1990, Cummins 1989, Maloney 1994).

In what follows, I outline a new entry into the field of naturalistic theories of extension for natural kind concepts (or, as I will call them, for reasons to be explained below, 'natural kind terms in the language of thought'). I have labeled this theory 'the best test theory of extension' (henceforth, 'BTT'), and as do some of the most prominent naturalistic theories of extension (Fodor 1987, 1990, Dretske 1981, 1988), BTT claims that causal relations fix the extension of natural kind terms in the language of thought. Relying on causal connections to fix term extension poses a difficult theoretical challenge, however. Of all of the causal relations into which a given language of thought term *t* enters, we must identify, in a principled way, those causal relations that fix the extension of *t*.

Consider a simple version of a causal theory of extension for natural kind terms in the language of thought. According to the simple theory, the extension of *t*, for subject S, includes whatever kinds of things cause S to token *t*, i.e., the extension of *t* for S consists of all of the members of any natural kind whose members ever cause S to token *t*. As Jerry Fodor notes (Fodor 1987, pp. 101-102), such a theory does not leave room for S to misrepresent her environment. No matter what item causes S to token *t*, that item is in the extension of *t* simply in virtue of its being a cause of S's *t*-tokening. Closely related to this problem is what Fodor dubs 'the disjunction problem'. It seems that on the simple causal theory, every extension is disjunctive in nature. For any natural kind term *t*, its extension consists of a disjunctive set of all of the members of all of the various natural kinds that ever cause *t*. On account of their being radically counterintuitive, disjunctive extensions are rejected by Fodor, as well as others. Thus, the felt need to solve the disjunction problem, accompanied by the hope that doing so will explain the nature of misrepresentation.

I hope to solve the disjunction problem by beginning with the idea that a natural kind term in the language of thought refers to the natural kind for which the concept associated with that

term provides the best test.[2] To develop this idea in sufficient detail, specific content-determining principles must be formulated to explain how reference is fixed in different kinds of cases. The most basic distinction here is twofold: Some terms, at some times, have their extensions fixed as the effect of the *content* of a subject's intentions directed toward those terms; of course, these extension-fixing intentions must themselves have their contents already fixed. In contrast, some terms, at some times, have their extensions fixed independently of the content of any of the subject's intentions. The latter case is the more fundamental of the two and is the subject of BTT's first content-determining principle (and the only one to be discussed in this paper), BT1:

> BT1: If a subject S bears no extension-fixing intentions toward *t*, and *t* is an atomic natural kind term in S's language of thought (i.e., not a compound of two or more other natural kind terms), then *t* has as its extension the members of natural kind K if and only if members of K are more efficient in their causing of *t* in S than are the members of any other natural kind.

BT1 clearly requires explication. The following sections explain in more detail how BT1 is to be interpreted and applied in order to solve the disjunction problem.

---

[2] I say 'concept associated with that term' for a distinct reason. Given the myriad associations readers are likely to make with the term 'concept', I wish to avoid the outright identification of concepts with language of thought terms. I do not here offer a theory of language of thought term individuation, but I treat at least some of these terms as atomic symbols. This contrasts with one common usage of 'concept' whereby concepts are meant to be complexes made up of various, related ideas, features, or properties. At some level of cognition, where we might talk of atomic concepts, a language of thought term and a concept might be the same thing; this is a possibility I want to leave open. However, given our inclination to think of concepts as complex, I will avoid talking in a way that identifies concepts with language of thought terms.

1.2 BT1: Application and Interpretation

1.2.1 BT1 and Success Rates

BTT assigns extensions to natural kind terms in the language of thought via the comparison of success rates. In order to compare the success rates of different natural kinds relative to *t* in S, we must first have in a hand a success rate relative to *t* in S for each natural kind. Success rates are determined by the success rate function, f< K, S, *t*, m>. This function takes four arguments, one each from the following four categories: a natural kind (K), a subject (S), a natural kind term in that subject's language of thought (*t*), and a time (m). When we plug an ordered quadruple from its domain into the success rate function, it yields an output ranging from 0 to 1 (although I frequently express the output's value in the form of a percentage, where, for example, 0.5 = 50%). The output of the success rate function is determined by the ratio of the number of times members of K have caused a tokening of *t* in S to the number of times members of K have caused a tokening of any language of thought term in S.[3] If, for example, members of K have caused the tokening of some term or other in S's language of thought 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 0.45. 45% of the times that members of K caused S to token any language of thought term at all, they caused the tokening of *t*.[4] It is in this sense that the success rate of K relative to *t* in S is a measure of

---

[3] 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 language of thought = 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, the success rate equals 0 by stipulation.

[4] What I say here gives the impression that an individual item causes the tokening of only one term per causal interaction with the subject. This seems unrealistic; frequently, if not always, the subject's reaction to an object she

how efficiently members of K cause the tokening of *t* in S.

Of the various kinds that can cause the tokening of *t*, some kinds may be very efficient in causing tokens of *t*, and many other kinds may not be. It should be clear that, relative to S's language of thought 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.

Along these same lines, it is crucial to note that 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* that have been caused by Ks. For example, of all of my tokenings of 'horse',[5] it may be that only 5% of them are actually caused by horses. Most of my tokenings of 'horse' probably result from other causes, such as thoughts in a chain of reasoning or other people's mentioning horses. However, the success rate

---

encounters includes a multitude of conscious and subconscious associations. Thus, I propose to complicate the calculation of success rates in the following ways: (1) any event which includes a member of K's causing the subject to token *t* (among other terms, possibly) is counted in the numerator, (2) any event which includes a member of K's causing the subject to token *t* (even if the subject tokens other terms as well) is counted only once in the denominator, and (3) in the event that a member of K causes the tokening of one or more language of thought terms none of which is *t*, this event is counted only once in the denominator. Among other things, what this means is that the tokening of a language of thought term other than *t*, in response to k, does not lower the success rate of K relative to *t*, so long as *t* was also tokened on that occasion. An anonymous referee brought to my attention the need to make my view on these matters explicit.

[5] On pain of circularity, we can not individuate language of thought terms according to their extensions before those extensions have been fixed. Thus, a more neutral description of 'horse' would be something bland such as 't'. By calling the language of thought term of interest 'horse' in this case, I only mean to let the reader know that the term of interest here is the language of thought term which we would, pretheoretically, identify as referring to horses.

of horses relative to my term 'horse' is *not* 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* language of thought term whatever and ask on what percentage of these occasions the subject tokened 'horse' as opposed to some other language of thought 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. 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 a language of thought term(s) in response to horses, I token 'horse' almost every time.[6] No other natural kind has a success rate relative to 'horse' in me which even approaches the success rate of the natural kind horse. This is why horses, and not the members of any other natural kind, constitute the extension of my language of thought term 'horse'.

In general, then, 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 (relative to *t*), the members of K constitute the extension of *t*.

Herein lies BT1's solution to the disjunction problem. Relative to the term 'horse', for example, other natural kinds whose members may occasionally 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 language of thought terms. Of the total number of occasions on which cows cause a subject to token any term of the language of thought at all, very few of these will be tokens of

---

[6] I do not here offer a theory of what exactly is required in order for a language of thought term to have been tokened by S. The paradigm case of a tokening is when S consciously thinks, e.g., 'horse'. However, it is very likely that many language of thought term tokenings are subconscious. If, for example, a language of thought term has even a subtle psychological effect, such as a priming effect (Stillings et al. 1987, pp. 28-29, Forster 1990, pp. 102-103), there seems to be no obvious reason for saying that the language of thought term has not been tokened.

'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.

1.2.2 BT1's Range of Application

While the 'horse' example effectively illustrates how BT1 assigns extensions to natural kind terms in the language of thought, BT1 does not necessarily apply to the typical subject's language of thought term 'horse'. BT1 only applies to atomic natural kind terms in the language of thought whose extensions are determined without the aid of any extension-fixing intentions. Some subjects' 'horse' tokens, at some times, may fit this description, but many subjects' 'horse' tokens, 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 bears an intention toward *t* whose extensional content is relevant to the determination of the extension of *t*. 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. For the purposes of this paper, questions about the precise nature and role of the relevant intentions will be set aside. It should suffice to say that if these intentions are present, and if their content plays an explanatory role in the fixation of content for *t*, then *t* is not subject to BT1. In remainder of the present work, we must keep clearly in mind the narrowness of the range of language of thought terms to which BT1 applies. This is especially important because many of the examples of natural kind terms in the language of thought that are likely to come to people's minds (such examples as 'horse') are not necessarily, and probably aren't, terms to which BT1 applies.

The reader may wonder at this point whether there are any natural kind terms in the language of thought to which BT1 applies. What is required is that we find natural kind terms in the language of thought whose extensions are not fixed via the subject's employment of

intentions whose contents play an essential role in the extension-fixing process. Such terms are most likely to be language of thought terms as they appear early in cognitive development, before the subject develops intentions to treat language of thought terms in specific ways. Some of these terms may be terms for which we lack lexically simple correlates in English. Whether we can confidently identify such terms depends on the resolution of a host of developmental issues which I cannot hope to effectively address in the present work. Still, in order to appreciate more clearly BT1's range of application, it may be worthwhile to inquire further after terms to which BT1 would apply. One likely candidate is the infant's language of thought term 'object', where 'object' is the infant's language of thought term that controls her successful responses in object recognition and manipulation tasks.[7] In the early development of the infant's object concept (as it's often called, though see note 2), the subject would seem to bear no conscious intentions toward the language of thought term 'object' that might play an essential

---

[7] There may be some disagreement as to whether terms such as 'object' are natural kind terms. 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 which we normally take to denote natural properties as natural kind terms. In the present case, we are justified in treating the property of being an object as a natural property given physicists interest in phase transitions (see, for example, Gleick 1987, pp. 126-128). However, note that as time passes in a subject's life, BT1 is not likely to remain the appropriate content-determining principle for the language of thought term 'object'. When, for example, 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'. The presence of these intentions would imply that for the subjects in question, BT1 is no longer the relevant extension-fixing principle relative to 'object'.

role in fixing the reference of 'object'.[8]

Here the concern may arise that the infant's use of 'object' is guided by implicit or subconscious intentions, the content of which is operative in fixing the extension of 'object'. If this were correct, BT1 would not apply to the infant's language of thought term 'object'. One source of the concern here might be David Marr's claim that the detection of the edge of a solid object proceeds via the detection of a zero-crossing segment in the firing of retinal cells (Marr 1982, pp. 64-66). Why shouldn't we say that the zero-crossing segment is part of an implicit intention directed toward 'object'? If so, wouldn't it be the extension of 'zero-crossing segment', rather than the extension of 'object' which is fixed by BT1? Then again, why should we think that zero-crossing segments have extensions at all?[9]

A reasonable, general answer to the question whether symbols of type X have extensions is to advert to the science of human psychology. Representations of type X have extensions whenever the assumption that they do facilitates the construction of good psychological explanations. If psychology gains empirical power from talking about Xs as have extensions, then this is reason enough to develop a theory of extension that applies to Xs. This leaves open the specific question whether a zero-crossing segment has an extension, and thus leaves open the question whether any extension-fixing principle of BTT should apply to zero-crossing

---

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

[9] Many of the mental structures in which cognitive scientists are interested, e.g., representations of rules in phonology, appear at the subconscious level. Note that some cognitive scientists are skeptical of the value of attributing extensions to such subconscious mental structures. See, for example, Jackendoff 1989, p. 76.

segments.[10]

1.2.3 Concepts

Concepts are usually thought of as concepts *of* something. But despite their own semantic natures, concepts are often thought to play a role in determining content. According to what we might call a 'feature-match' theory of extension (a theory distinct from BTT), a language of thought term *t* refers to a certain group of objects because those objects match, to a satisfactory degree, the features that make up the concept associated with *t*. For example, 'bird' would refer to those animals which possess a sufficient number of the right features, e.g., 'flies', 'sings', 'lives in trees'.[11] On such a view, the extension of a natural kind term in the language of thought is determined by testing the concept associated with that language of thought term against the various natural kinds to see which natural kind is best described by the concept's weighted feature list.[12]

The problem with a feature-match theory of extension is that it cannot be as general or as independent as we would like. A feature-match theory of extension cannot apply to the features

---

[10] 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.

[11] A feature-match theory of reference for language of thought terms is inspired by two sources: (1) examples given in psychological work on concepts (e.g., Smith and Medin 1981) and (2) what has come to be known (and criticized) as the description theory of reference (see Kripke 1980. p. 71 and passim).

[12] 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.

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 would 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 a language of thought term unless the content of some other language of thought terms had first been determined in accordance with a different theory.

In contrast to the semantic basis of a feature-match theory of extension, BT1 specifies a causal, non-semantic mechanism of extension fixation. Instead of looking for semantic fit between concepts and natural kinds, BT1 focuses on the patterns of causal connections among members of natural kinds and the natural kind terms in a subject's language of thought. Here it is of utmost importance to bear in mind that BT1 only applies in cases where the subject bears toward *t* no intentions the contents of which are operative in the fixation of *t*'s extension. If we wish to talk of concepts here, such concepts should be understood as non-semantically individuated collections of psychological and physiological mechanisms that mediate the tokenings of the relevant language of thought terms. According to BTT writ large, concepts play an intermediate role (sometimes an essential, intermediate role) in the extension-determining causal chain that runs from members of natural kinds to a subject's tokening of terms of the language of thought. However, at the fundamental level at which BT1 applies, extension is ultimately determined by the patterns of causal connections, specifically those represented by the success rate function; and success rates are only sensitive to causal connections between the members of the natural kinds and tokenings of terms in a language of thought.[13]

---

[13] To those familiar with 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 essentially the same viewpoint with respect to concepts in presenting his asymmetric dependence theory of content. As Fodor puts it, "It's the existence

1.2.4 Causal Connections

In order to calculate the success rate of natural kind K relative to a language of thought 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 thus far from the explanation of success rates is an explanation of what it is for a member of a kind to cause a tokening of a language of thought term. While it is beyond the scope of the present work to analyze the concept of a causal interaction, it may be useful to give some idea 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 the tokening *t* in S if and only if had k (or any member of group k1...kn) not had causal effects on S, S would not have tokened *t* on the occasion in question.[14]

---

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 a language of thought term), then we can't dismiss the presence of these states as mere mindless mechanisms at work. In such cases, we must go beyond BT1 and apply an extension fixing principle which explains the role of the content of these intentions in fixing the extension of the relevant language of thought term(s).

[14] Note that CR1 allows for the group causation of the tokening of a language of thought 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 one instance of a member of K causing the tokening of *t*. Ultimately, this way of counting events may be unsatisfactory, for it erases any extra contribution that recognition of a group of ks (as opposed to an individual k) might make to the fixation of *t*'s extension. Given the complexity which the necessary formal apparatus is likely to possess, I do not

CR1 seems too liberal: CR1 seems to allow distant causes to be relevant to content determination. Consider your heart. It seems that, according to CR1, your heart causes the tokening of each and every language of thought term you token. For any given tokening of a language of thought term, had your heart not had some causal effect on your brain, you would never have tokened that language of thought term.

The heart example is not as troubling as it may seem. Because your heart makes a causal contribution to *every* tokening of every language of thought term, the success rate of hearts will be very low with respect to *any particular* language of thought term. Choose a language of thought 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 hearts causally contributed to the tokening of 'object' by the number of times hearts causally contributed to the tokening of any language of thought term whatsoever. The former number will be but a minute percentage of the latter. Thus, the success rate of hearts relative to 'object' will be much lower than the success rate of objects relative to 'object'.

Problems may arise, however, when CR1 is used as the basis of the assignment of an extension to 'heart'. As was just noted, S's heart is a cause of the tokening of every language of thought term that S ever tokens. This greatly reduces the success rate of hearts relative to 'object', but it seems to equally reduce the success rate of heart relative to 'heart'. In this case, deference to some principle beyond BT1 might seem in order. As a matter of empirical fact, it would seem that most, if not all, of the subjects who have a language of thought term whose extension would be intuitively identified as the set of hearts are subjects who have intentions toward hearts that can help ground the extension of the language of thought term 'heart'. However, given that I've presented no clear criterion for determining when such intentions exist,

---

here attempt to take into account any extra extension determining contribution made by the group.

and when their content is relevant, it would be better if this problem could be solved without deference to extension-fixing principles beyond BT1.

I believe that a BT1-based solution can be found. Note that for the members of a natural kind to be the extension of a natural kind term in the language of thought, the success rate of the kind relative to the term does not have to be particularly high. It merely has to be higher than the success rate of any other kind (where the necessary degree of difference can be fleshed out with reference to standards of statistical significance). Thus, so long as the success rate of hearts relative to 'heart' is significantly higher than the success rate of any other natural kind relative to 'heart', hearts can yet be the extension of 'heart'. This can be the case even though hearts contribute causally to the production of every language of thought term and thus have a fairly low success rate relative to 'heart'. (Here I am not trying to convince the reader that the extension of 'heart' is ever, in fact, determined in the way just described. I merely want to show that BT1 leaves open the possibility that a subject could have a language of thought term that refers only to the members of natural kind K even when members of K contribute to the tokening of every language of thought term for that subject, and even when the subject does not bear any extension-fixing intentions toward members of K.)[15]

2. Theoretical Assumptions: Methodological Issues

2.1 Representation and Psychological Explanation

It is tempting to view a naturalistic theory of extension as no more than an attempt to fit the phenomenon of representation, pretheoretically conceived, into the natural order. To know

---

[15] To this point, I have only addressed the relevance of actual causal interactions between the subject and items in her environment; however, circumstances may arise where it seems better to consider counterfactual causal interactions when assigning extensions to natural kind terms in the subject's language of thought. This is an issue well worth discussing, but one that limitations of space prevent me from exploring in any detail in the present work.

whether a theory of extension for natural kind terms in the language of thought succeeds on these terms, we only need know whether the theory assigns extensions that accord with our relevant semantic intuitions. For example, an obvious reason for wanting to solve the disjunction problem is that people take their thoughts about horses to be thoughts *about horses* and only horses. They do not take these thoughts to be thoughts about the disjunctive set that includes whatever items sometimes cause (what they would describe as) thoughts of horses.

The desire to respect subjects' semantic intuitions provides a respectable motive for our aspiring to solve the disjunction problem. However, the project so, characterized, faces two legitimate concerns. First, meaning isn't always what we take it to be. The naturalist wants to find out what meaning is, not just what people think it is, and this point applies equally well to our interest in extensions. According to a naturalist, our theory of extension should fit into our overall scientific view of the world, i.e., it should cohere with the entire body of empirically successful science, not just subjects' intuition checks. Secondly, if the motivation behind a theory of extension is only to respect semantic intuitions, then we risk letting the evaluation of a proposed theory of extension degenerate into a tussle among conflicting semantic intuitions. Here the naturalistic emphasis on empirical success can provide a more precise, although not completely precise, criterion for use in evaluating naturalistic theories of extension.[16]

---

[16] Many of the points made in this paragraph echo points made by Michael Devitt (See Devitt 1994). In some ways, though, my view diverges from Devitt's. Devitt claims that an investigation of meaning should not begin with an investigation of reference or truth-conditions because such an investigation would not be appropriately basic and would bias the issue as to the nature of meanings (Devitt 1994, p. 555). Thus, Devitt might say that if I intend to construe representation as something like 'relation-to-a-referent', I am starting the investigation from a biased standpoint. My reaction to Devitt's claim consists of two parts. First, I am not claiming here to be giving a theory of meaning, rather to be saying something interesting about mental representations and their possible connections with the world beyond the subject. Secondly, it seems to me that representation as a mind-world relation is an appropriately basic object of naturalistic inquiry, insofar as thinking in terms of mind-world relations is a basic

Assume, then, that the measure of a naturalistic theory of extension is its success in providing to empirically successful theories that which they need to effect their best explanations. Fair enough, but to which theories do we philosophers look? Views on this matter vary. Claiming that there really is no pretheoretical notion of representation (at least it is "not a commonplace of ordinary discourse" [Cummins 1989, p. 12]), Robert Cummins counsels us to construct theories of intentional content for one scientific theory at a time, so to speak (Cummins 1989, pp. 12-13). In *Meaning and Mental Representation*, Cummins pursues well his own counsel and presents a theory of representation suited to explanation within the framework of one promising theory of cognition, the computational theory of cognition.

Cummins takes the right approach in giving empirical success the primary, motivational role. However, for the naturalist, matters may be more straightforward than Cummins makes them out to be. As Fodor is at pains not to let anyone forget, folk psychology is among our most successful, empirical, psychological theories; folk psychology works, day in and day out (Fodor 1986, pp. 420-421, Fodor 1987, chapter 1, 1994, pp. 3-4; cf. Churchland 1981, Horgan and Woodward 1985). And folk psychology clearly employs the idea of a person's thinking about something, or in the case of natural kinds, a person's thinking about a group of things. ("Why did John go to the dude ranch for his summer vacation?" we ask the folk. "Because he was thinking about horses, and thought he might like to try riding one," the folk may well reply.)

Thus, to fairly evaluate BTT would require first identifying specific empirical theory(ies) that BTT's component principles are intended to serve. Subsequently, we would need to examine various explanations offered by this empirical theory(ies), checking to see whether the

---

accomplishment that qualifies the child as an "expert at identifying meanings" (as Devitt might put it). See Astington 1993 (especially chapter 8) for an argument that at least a limited understanding of representation as a relation between a mental state and items in the world is acquired by the child in the very earliest stages of the child's understanding of the mind. And one might reasonably think that acquiring a theory of mind is necessary for the child's true understanding of the types of meaning normally thought to be relevant to semantic theory.

extensions assignments made by BTT are the same as those assumed by the relevant explanations. Though I do not go so far as all this in the present work, some general remarks about the role of BTT in the naturalistic enterprise are called for. BTT is intended to serve everyday intentional psychology (i.e., folk psychology) and those theories in cognitive psychology that are built largely out of the material offered by everyday intentional psychology (where these materials are representational states, such as beliefs and desires, though perhaps refined in some ways).

2.2 Evolution, Optimization, and Animal Representation

Before leaving the question of how precisely to evaluate BTT, I should like to discuss one other oft-pursued theoretical approach, the ethology-based approach. In the discussion of theories of reference for language of thought terms, non-human animals' sensitivity to features of their environments is often treated as a kind of proto-example after which a theory of extension for natural kind terms in the language of thought should be modeled. Furthermore, it is widely thought that, to be successful, a theory of content for terms in a human language of thought should make the 'correct' content assignments to elements of feature-detection devices in non-human animals. (Millikan 1984, 1990, Dretske 1986, 1988, Godfrey-Smith 1991, 1992, Neander 1995, and Brown 1996, among other works, all discuss examples where animals selectively respond to their environments as if these were relevant to the development and the evaluation of theories of intentional content for humans.) Some authors are careful to clearly demarcate the simple animal response to the environment from full-blooded representation (a distinction of this sort is a key component in both Millikan's and Dretske's theories of content). However, it seems generally accepted among naturalists that human representation is of a piece with animal representation and that, therefore, a theory of extension for natural kind terms in the language of thought had better be suited to results in ethology and in the biology of nonhuman animals.

Assume for the moment that the ethology-based approach is the correct one. BTT seems

faced with a difficult challenge, stemming, in particular, from the work Peter Godfrey-Smith (Godfrey-Smith 1991, 1992). Employing Hartry Field's distinction between head-to-world correlation and world-to-head correlation (Field 1990), Godfrey-Smith argues for two related theses: (1) theories of content which give elevated status to head-to-world correlation are misguided, and (2) the whole idea of placing emphasis on one of these factors or the other, or some set proportion of the two, ignores the fact that which kind of correlation is useful to the organism depends greatly on the particular environmental circumstances. In what follows, I argue that neither of Godfrey-Smith's arguments impugns the value of BTT. First, however, we should have a full understanding of Field's distinction and BT1's position in relation to that distinction.

A maximum head-to-world correlation exists between *t* and its extension when a subject's tokening of *t* guarantees that an item in the extension of *t* caused that tokening of *t* (if it's in your head, then it's in the world). On the other hand, a maximum world-to-head correlation exists if the presence of a member of the extension of *t* always causes the tokening of *t* (if it's in the world, then you notice it). The first case (head-to-world) leaves open the possibility that many members of the extension of *t* can go unnoticed, while the second case (world-to-head) leaves open the possibility that many other things besides those in the extension of *t* can cause the subject to token *t*. Insofar as the measure of success rates fits into either one of these two categories, success rates are a measure of world-to-head correlation.[17] As I tried to make clear in section 1.2.1, the cause of a subject's 'horse'-tokening is very often not a horse, even when the success rate of horses is very high relative to that subject's language of thought term 'horse'. All

---

[17] A qualification is required here. BT1's assignment of content does not depend on any measure of the rate at which local items 'get noticed', so to speak. BT1 is based on a qualified world-to-head connection. It says that *if* you token any term at all in response to the presence of an item, then that tokening counts in the calculation of the relevant success rates. If there is, for example, a horse around, but it doesn't cause you to token any language of thought term at all, consciously or subconsciously, the calculation of success rates is in no way affected.

that is required here for a high success rate is that when horses cause the subject to token any language of thought term at all, this should frequently be 'horse' (or, to be more precise, 'horse' should be among the terms horses cause to be tokened on these occasions; see note 4; also, recall that having a *high* success rate, on some absolute scale, is not required for K to be the extension of *t*, only that K's success rate be significantly higher than that of any other natural kind). Thus, on BT1, given the knowledge that a subject is tokening 'horse', there is no reason to infer that a horse caused that tokening, even if BT1 assigns horses as the extension of 'horse'. This is to say that BT1 in no way predicates reference on any kind of head-to-world correlation (of a high degree or otherwise). Given that the greater part of Godfrey-Smith's critical remarks are aimed at theories such as Dretske's, theories that assign extensions on the basis of high head-to-world correlation, BT1 is untouched by these critical remarks. Thus, we can leave behind Godfrey-Smith's first thesis listed above.

What about Godfrey-Smith's second thesis? Here Godfrey-Smith's point requires more explanation. Using decision-theoretic matrices, Godfrey-Smith argues that whether one should want high head-to-world correlation or high world-to-head correlation depends on the structure of payoffs that result from basing one's behavior on the tokening of *t* (Godfrey-Smith 1991, pp. 718-721, 1992, pp. 298-308). For example, if food is scarce and there is no danger to you in preying on animals that are not nourishing to you, then high world-to-head correlation is desirable (regardless of whether or not your symbol for prey has high head-to-world reliability). Since attacking by mistake causes no great loss, you will be able to tolerate many false positives, so long as this means that when your prey does come around, you'll be sure to notice. On the other hand, if, for example, tokening *t* when members of the extension of *t* are not present is absolutely deadly (because, for example, you wind up attacking animals that prey on you), then you would presumably want high head-to-world correlation. In this situation, if you think you see prey (i.e., you token *t*), you want something close to a guarantee that prey is near, rather than an animal who wants to eat you.

Godfrey-Smith makes two related points, both based on the idea that whether head-to-world

or world-to-head correlation is useful to the organism is a matter of context. "Except in the most epistemically benign and friendly environments, the two kinds of reliability cannot be bought as a package deal; rather one must be traded off against the other." (1991, p. 712) And secondly, an organism will want "an inner state with both head-to-world and world-to-head reliability weighted equally...[o]nly when the costs and the benefits are symmetrical." (1992, p. 303) Let's take the points in reverse order. Since BT1 does not assign extensions on the basis of an equal consideration of head-to-world and world-to-head correlations, Godfrey-Smith's second way of making his point does not constitute a criticism of BT1.

With respect to the first passage quoted above, Godfrey-Smith's point is that across a wide range of biologically realistic cases, head-to-world reliability and world-to-head reliability can be, and are, traded off against one another depending on the details of the case. Godfrey-Smith does not explicitly discuss a theory of content that places all of its eggs in the world-to-head basket, as BT1 does. Still, his point would presumably apply to world-to-head theories just as much as he takes it to apply to theories that endorse the need for a high rate of reliability of both types. A criticism of BT1 seems lurking here; for if world-to-head reliability is negotiable, then BT1's emphasis on the success rate measure (a world-to-head measure, of sorts) seems misguided. Consider Godfrey-Smith's example of a case where high head-to-world reliability is useful to the organism without any special regard for world-to-head reliability. In this case, we are to imagine that there exist females of a species for whom it is very costly to get pregnant by a member of a different species. These females are best served by a high head-to-world correlation between their mental term for 'male conspecific' and male conspecifics. If a female believes that there is an appropriate male present and prepared to mate, she wants to be as certain as possible that there really is such an appropriate male present, rather than a male of a different species who happens to resemble her male conspecifics. This need for high head-to-world reliability can exist even when the world-to-head correlation between male conspecifics and the relevant symbol is very low. Imagine that there are so many male conspecifics in the immediate environment of any individual female that the female might fail to notice (or even misidentify)

nine out of ten appropriate males that pass by without thereby giving up the opportunity to mate and produce viable offspring in a biologically timely fashion.[18]

This example does not constitute a substantive objection to BT1 for a number of reasons. First off, the example as given may not be appropriate because it seems to involve a compound language of thought term, 'male conspecific'. So long as the female members of the species in question have a language of thought term 'male' and a language of thought term 'my species', then it would seem that 'male conspecific' is not an atomic term and, thus, not a term to which BT1 does applies.

Setting aside concern about the compound nature of 'male conspecific', the example seems yet to fail as an objection to BT1. Simply because a given female fails to notice, or even misidentifies, nine out of ten passing male conspecifics (i.e., just because the world-to-head correlation is low) does not show that BT1 assigns the wrong extension to the female's language of thought term 'male conspecific'. The assignment of extensions according to BT1 is based on the *comparison* of success rates for the various natural kinds. As the example stands, it would seem that male conspecifics *do* have the highest success rate relative to the female's language of thought term 'male conspecific'. Thus, even if the success rate of male conspecifics is only, say,

---

[18] This is an elaboration of an example Godfrey-Smith describes only briefly at 1992, p. 303. Godfrey-Smith works this example out in mathematical detail at 1991, pp. 720-721. I exclude discussion of Godfrey-Smith's detailed version of the example, for as he works the example out in detail, world-to-head correlation remains highly useful even though reliability of the head-to-world correlation is more useful. Godfrey-Smith's way of filling in the details does not quite serve present purposes for it seems to give some special status to world-to-head correlation, given that world-to-head correlation remains high even where head-to-world reliability is supposed to be in the spotlight. I am here trying to elaborate a possible criticism of BT1 based on the idea that world-to-head correlation is negotiable, so to speak. However, if world-to-head correlation remains high even when head-to-world correlations are especially useful, world-to-head reliability does not seem negotiable after all, leaving the criticism of BT1 on shaky ground.

0.1, relative to a given female's language of thought term 'male conspecific', male conspecifics can still be the hands down winners in the competition for the highest success rate.

Perhaps, in the interest of fairness, we can alter the example so that male conspecifics don't have the highest success rate of all natural kinds. Let's throw into the mix a third species whose males have a higher success rate relative to 'male conspecific' than the females' male conspecifics do, but with whom mating has negligibly negative consequences for our hypothetical females. Imagine that for some reason, the males of this third species, with whom the females cannot fruitfully interbreed, strike a breeding interest in the females in question. Though the addition of this third species may make the example quite unrealistic, a point of some interest to naturalists, it seems to yield an example where BT1's extension assignment is at odds with the assignment preferred by Godfrey-Smith. While the altered example may deserve more scrutiny, I propose to leave it here. This is partly because it's not clear what the correct extension assignment should be in a case where the females' mating behavior still yields the biologically important payoff, offspring, even though the females frequently engage in innocuous copulation with males with whom they cannot conceive. If Godfrey-Smith were to claim that the extension of 'male conspecific' should be the group of male conspecifics in this case, we would like some argument as to why this is the correct extension assignment. However, larger issues loom, and it is to those I shall turn.

While this is not the place to provide a full critique of Godfrey-Smith's work, there are fundamental, unanswered questions regarding his approach and its relation to BTT. Firstly, the value of Godfrey-Smith's optimization assumption is unclear in the context of the current project. Godfrey-Smith is at pains to show that if we take into consideration the full structure of costs and payoffs, it's not always optimal for the organism to employ terms having only one of the two types of reliability, head-to-world or world-to-head. Which of the two is optimal for the organism depends on the specific circumstances, including the relevant payoffs and costs. When it comes to assigning extensions, then, Godfrey-Smith suggests that when *t* is tokened in an optimal way in response to members of a given kind (that is, when it is tokened with just the

right degree of mind-to-world and world-to-mind reliability given the historical circumstances in which the tokening of that term has arisen in the organism), then *t* refers to the members of that kind (1991, p. 721). However, it's doubtful that considerations of optimality should drive our theory of extension for human language of thought terms when this theory of extension is intended to serve good psychological explanations of human thought and behavior. Typically, human behavior is not sensitive to the full structure of costs and payoffs. Humans usually are not aware of the full structure of payoffs and costs, and thus they act in ways which hardly look rational from the standpoint of the omniscient observer attempting to assign content by identifying optimal behavior in response to the environment. Furthermore, when humans do know the full structure of payoffs, they don't always act rationally (and more probably than not, I understate my case here).[19]

A related fact about the human condition makes it yet less likely that Godfrey-Smith's decision-theoretic analysis, couched in terms of costs and payoffs, bears directly on the construction and evaluation of a naturalistic semantics for natural kind terms in a human's language of thought. Godfrey-Smith bases his analysis on the assumption that the tokening of a mental/sensory structure automatically causes a given behavior (such as attacking what is perceived to be prey) (1991, p. 710). Godfrey-Smith finds this useful because it allows him to assess the worth of various degrees of head-to-world and world-to-head correlation by comparing calculable payoffs to their costs. Given the assumption that tokenings of *t* automatically lead to a particular behavior type M, Godfrey-Smith can assign a cost or benefit to the tokening of *t* itself. However, such assignments of costs and benefits to term tokenings themselves *cannot* be neatly accomplished if one assumes a fair degree of mediation between term tokening and behavioral response. This is a problem in the current context because across a wide range of cases, there is no automatic response that a human makes to the tokening of a

---

[19] There is some question as to whether it's possible for humans to act in a way that approaches optimal responsiveness to their environment. See, for example, Cherniak 1986.

given language of thought term. For humans, there is typically an enormous amount of cognitive mediation between perception and behavioral response, mediation that does not exist in most other (probably any other) earthly species. Thus, Godfrey-Smith's analysis seems tangential to the evaluation of a theory of extension for human language of thought terms.

Godfrey-Smith considers the history of the use of a language of thought term (either learning history or evolutionary history) and on that basis assigns an extension. This is a naturalistic approach to language of thought semantics; its aim is to place reference in the natural world in way that is inspired by evolutionary theory and biology. The approach I'm suggesting is equally naturalistic, though different in important respects. On the approach I've suggested, we look first to the most successful explanations of human capacities and behaviors as they now exist and ask what theory of reference is required to validate the talk of representation on which these explanations are premised. My preferred approach does not imply the uselessness of ethological results or evolutionary speculations. It may be that once we know exactly what we're looking for in a theory of reference for human language of thought terms, we will find it in a theory of the nature of Godfrey-Smith's. However, assuming that a theory of reference should fall out of the natural history of the wide range of earthly species introduces potential pitfalls. By associating human mental representation too closely with what we take to be its simple, and perhaps ancestral, roots in the animal kingdom, we run the risk of making assumptions such as Godfrey-Smith's assumption that a given representation type is tied in a knee-jerk fashion to a given behavior type. In contrast, much of what makes us take the idea of mental representation seriously in the first place is lacking from most species.[20] Humans reflect on events in progress. Humans deliberate and decide what to do, rather than immediately reacting with instinctive or

---

[20] For arguments in support of explicit representation in humans, see Fodor 1987, appendix, Fodor and Pylyshyn 1988, Clark 1991, and Clark and Toribio 1994. The arguments found in these works appeal to facts about human mental capacities, such as the capacities for abstraction, and for systematic and productive thought, that are found in very limited quantities in non-human species.

reflexive behaviors. Humans use an intentional psychology to make predictions about and give explanations of other humans' thoughts and behaviors. These human characteristics place us at such a distance from Godfrey-Smith's creatures who always and automatically react to their tokenings of *t* with a set behavior--and these characteristics place us at such a distance from frogs, hoverflies, and marine bacteria that possess magnetosomes (to list a few commonly used examples)--that one has to wonder at the value of such examples in our attempt to understand how reference can help to explain human behavior. While it often simplifies matters to begin inquiry with an artificial case, such as an investigation of what looks to be representation in animals,[21] we should not lose sight of the distance of such simplified cases from our own.

In reaction to my criticism of the ethology-based approach to theories of intentional content, one might worry that if my criticism of the ethology-based approach is cogent, BT1 is of only dubious value. For if we can find any language of thought terms to which BT1 applies, they will be precognitive, in the sense that they would not participate in the kinds of relations that motivate talk of human mental representation (e.g., the relations that terms bear to each other in systematic thought). Terms such as the young child's language of thought term 'object', i.e., terms to which BT1 is supposed to apply, may actually be located closer the sensory 'representations' Godfrey-Smith discusses, in which case BT1 would be subject to some of Godfrey-Smith's criticisms after all.

---

[21] I do not intend to be promoting a narrow-minded speciesism here. My point is that given all of the impressive results on animal cognition, humans remain the best examples we have of systems that we can confidently cite as full-blooded representational systems. Good naturalistic methodology seems to say that if we want to find out about representation as a natural phenomenon, and if humans provide our best examples of natural representational systems, then the investigation of representation in humans takes theoretical priority. Note also that the approach I advocate leaves open the use of ethological examples as means of illustrating theoretical concepts or ideas, which seems to be more the point of, for example, Dretske's discussion of magnetosome-possessing bacteria (Dretske 1986).

This defense of the relevance of Godfrey-Smith's work to the evaluation of BT1 misses the mark, for two reasons. First, on one plausible view of human sensory systems, which Kathleen Akins nicely characterizes as the 'pre-ontological' view, the most immediate deliverances of individual human sensory systems do not amount to detections of natural kinds or properties in the environment (Akins 1996). On this pre-ontological view of the senses, the sensory structures most likely to govern automatic, reflexive responses in humans, to the degree that we possess such structures, are not natural kind terms in the language of thought (and may not have extensions at all). On this view, tokenings of a language of thought term such as the young child's term 'object', which would seem to have ontological import, are at least one step removed from the immediate deliverances of the senses. This view of human sensory operations defuses Godfrey-Smith's criticisms as criticisms of BT1 by again placing distance between the basic natural kind terms in the language of thought, to which BT1 applies, and the proximal stimuli to which humans might react in an automatic, reflexive way, i.e., in the way relevant to Godfrey-Smith's discussion.[22]

Secondly, even if there is something correct about Godfrey-Smith's analysis at the level of the terms to which BT1 applies, world-to-head correlation can still have a privileged place in the explanation of human behavior. Recall that BT1 is intended to explain only the rudiments of reference. The larger goal, to be discussed in slightly more detail in the closing section, is to explain how the subject can use content determined by BT1 to fix and refine the extensions of

---

[22] I will not here attempt to resolve general questions regarding the nature or the functioning of human sensory systems. I assume that for humans the representation of natural kinds, even at the most basic level where BT1 applies, is mediated to a large extent by the processing of sensory stimuli. It would be quite a project to spell out how language of thought terms achieve the status of natural kind terms and to spell out to what degree, if at all, the achievement of this status depends on the existence of some other, contentful, but non-natural kind terms (which may, e.g., be representations caused by direct sensory stimuli). In a separate paper, I have begun this project.

other language of thought terms. These other language of thought terms are sure to be invoked in the explanation of human behavior that does not accord with the optimality assumptions on which Godfrey-Smith's arguments are premised. Thus, if the overall project of BTT is carried out with any degree of success, then BT1 is vindicated in the face of any of Godfrey-Smith's criticisms that might apply at the level of BT1, simply by virtue of what BT1 contributes to the wider success of BTT as a theoretical underpinning of successful extension-based psychological explanation of human behavior.

3. Theoretical Assumptions: Foundational Issues

3.1 Natural Kinds

Insofar as intentional psychology, as done by professionals or simple folk, assumes that human beings sometimes think about natural kinds, we should want to explain the relationship of aboutness that human thoughts bear to natural kinds. However, this project proceeds from an assumption that many find questionable, the assumption that there are such things as natural kinds (Lakoff 1987, chapter 12, Dupre 1981). Given requirements of space, I will mount only a brief defense of the existence of natural kinds, but I think that this will be adequate for present purposes.

First consider scientific success, in chemistry, for example. In their theoretical and applied pursuits, chemists have succeeded far beyond Priestley's or Lavoisier's wildest dreams (I suppose). In general, inference to the best explanation of scientific success suggests that the kinds of things that chemists, physicists, and biologists talk about really exist. It would seem to be just short of a miracle that, for example, antibiotics could be effectively manufactured and used successfully in treatment without there being a real kind *bacteria* in the universe.

The preceding defense of natural kinds suggests a perhaps familiar definition of natural kinds, that they are the kinds which appear in statements of scientific laws. I will adopt a more liberal definition, however, assuming that natural kinds are any of the kinds that successful non-

intentional science finds theoretically interesting and useful. This view of natural kinds inspires the following, somewhat glib response to the objection that there are no natural kinds: The exact nature of natural kinds matters little in the explication of BTT; natural kinds are whatever good science says they are.

Further concerns arise, however, when we turn to the question of biological kinds. This is of special importance *vis-à-vis* BTT, for biological kinds seem to be among the natural kinds that the average person is most likely to think about. In biology itself, there have existed in recent times 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 (See Gould 1983, Maudlin 1986, and Lakoff 1987, for descriptions of the ways in which the three methods differ). In addition to problems raised by discord among biologists themselves, the nature of the relevant scientific laws poses problems. Per the strict view stated above (that natural kinds are the kinds mentioned in statements of laws of nature), we can glean from the laws of biology a partial list of existing natural kinds, including species, so long as the relevant laws of biology are genuine laws of nature. However, it has been widely claimed that generalizations in biology that mention species are not true laws, simply because of the nature of species.[23] At least if one is a cladist or an evolutionary systematist, evolutionary biology defines species in such a way that being a member of species X depends on your possessing a relational property, e.g., the property *being a descendant of population X* or *being able to interbreed with members of population X* (Hull 1978; also see the works cited in Lange 1995). Such properties are thought to rob species of the abstract nature that is required of the kinds of thing that could

---

[23] Notice the gap between talk of biological kinds in general and species in particular. Within evolutionary biology, we will likely find theoretically important kinds that are not to be thought of as species, e.g., *predator* (Sober 1984, 335, 1980, 184). Bear in mind that the criticisms of biological kinds discussed in the text are directed specifically at the idea of species as natural kinds, though the critic may be able to modify the criticisms so as to apply to other, non-species kinds such as *predator*.

appear in statements of universal laws of nature; instead, it is claimed, species are defined in such a way that they seem more like individuals, and we don't expect there to be laws of nature that mention individuals.

Taking these two objections in order, it seems, first, that disagreement among biologists as to the proper way to define a species is little proof that there are no true species. Great disagreement once existed as to the nature of electricity, but this did not show that there was no such thing as electricity. If the debate over the nature of electricity was satisfactorily resolved, why not think that the debate over the true nature of species will be similarly resolved? We may have partial resolution already, to the degree that phenetics is no longer treated by many as a viable theory of how to taxonomize organisms. Furthermore, if there remain good empirical reasons to use the three different methods of dividing organisms into species, we might just as well claim that the species as defined by each of the three approaches all constitute natural kinds (A promiscuity of this sort is encouraged to some degree in Mishler and Donoghue 1982 and Kitcher 1984). This results from a consistent application of the dictum 'good science tells us what the natural kinds are.' Of course the existence of more natural kinds requires the consideration of more success rates in the application of BT1 to any given term, but this does not count directly against BT1.

With respect to the second objection to the treatment of species as natural kinds, that species are individuals and not kinds, I can only point elsewhere. A recent defense of the idea that some species-specific generalizations should be regarded as genuine natural laws can be found in Lange 1995.[24] Also see arguments in Kitcher 1984 and Kitts and Kitts 1979 to the effect that

---

[24] Also see Horgan and Tienson's defense of 'soft' laws (Horgan and Tienson 1996). Though it is not their main point to argue for the lawlike nature of biological generalizations, some of their examples of soft laws make reference to biological species (e.g., 'Cheetahs are fast' [Horgan and Tienson 1996, 116]). Thus if soft laws are scientifically legitimate, so would be the types of biological generalizations discussed in Lange 1995. Ultimately Horgan and Tienson's arguments may be of more use to the defender of BTT, for Lange's analysis seems to imply

species may well be natural kinds, whether or not they figure in laws of nature; this is the position that I am more inclined to endorse.

Having pointed elsewhere for a defense of the claim that biological kinds are legitimate natural kinds, I should note that with respect to BT1, the entire debate about species may be moot. It's not clear that BT1 applies to species terms. The animate/inanimate distinction seems to emerge fairly early (much earlier than Piaget thought, for example; see Carey 1985 and Massey and Gelman 1988), but this distinction does not involve naming any particular species. Perhaps BT1 would apply to 'human', given the evidence of the infant's preferential treatment of faces (Karmiloff-Smith 1992), but even here we may not have reached the foundation of the reference-fixing process. Very young infants seem to be sensitive to amodal, structural properties of stimuli coming through the various sensory organs (Streri and Spelke 1988, Bower 1989), and perhaps it is here that BT1 finds its purview. BT1 likely applies to some very general natural kind (or property) terms, acquired early in development, that are then used by the child to fix reference to other natural kinds. On this picture, it is the content of the terms to which BT1 applies that fixes the content of natural kind terms as we might normally think of them (such terms as 'tiger' and 'gold'). Thus, some principle beyond BT1 is likely to tell the correct story about reference fixation for natural kind terms that are typically cited as such.

Before moving on, I wish to address a related issue regarding just what counts as a natural kind. Consider the kind *female chimp with oestrus swelling*.[25] Often, the kinds that are of importance to sentient creatures are kinds such as this, kind that may not seem to be natural kinds in their own right. Let me start by saying that we should be wary of this example for a couple of reasons. First, it would seem to involve the tokening of a complex language of thought

---

an interest-relativity of biological kinds, where BTT would seem to require an objective status for natural kinds, at least with respect to the kinds to which BT1 applies.

[25] This example was suggested by an anonymous referee.

term, thus one to which BT1 would not apply. We can best explain how reference is fixed in this case by allowing BT1 to fix the extensions of the components of the complex expression (all of which, except for 'with', have reasonable claim to being terms which, taken individually, denote natural kinds). Then we can go on to tell a story, which I will not tell here, about the combinatorial semantics for language of thought terms. Second, I have already made clear my misgivings about examples based on animal cognition (be there such a thing, in the relevant respect). Such misgivings persist. All this being said, is there a way to solve the problem without what might seem to be a dodging of the issue?

Assume that a male chimp has an atomic term *o* that we think should refer to female chimps with oestrus swelling. BT1 can explain how *o*'s extension is fixed so long as we allow the intersection, so to speak, of two or more natural kinds to count as a natural kind, at least in cases where the treatment of such kinds as natural kinds is motivated by natural (i.e., non-intentional) sciences. Reproduction is a biological fact the characterization of which requires no intentional or semantic terminology. For this reason, *being female* is a biological kind, as is the kind *oestrus swelling*. Of the natural kinds given to us by the non-intentional sciences, then, *female chimp with oestrus swelling* may well have the highest success rate relative to the male chimp's atomic representation we pick out with the our compound expression 'female chimp with oestrus swelling'. I am not entirely comfortable with this approach, for the solution I've proposed seems to lead to an unacceptable promiscuity with respect to natural kinds. Such promiscuity would seem to open the door to a wide range of counterexamples to BT1 that might be constructed by finding the intersection of just the right legitimate natural kinds to yield a new natural kind with a very high success rate relative to some *t*, even though the resulting extension assignment to *t* is counterintuitive. Given what I take to be reasonable discomfort with example itself, and given the fact that these examples are difficult to come up with when limiting ourselves only to terms to which BT1 is likely to apply (e.g., 'object' in the mind of a child), I leave matters here.

3.2 Natural Kind Terms

Beyond the assumption that natural kinds exist, my presentation of BT1 may seem to assume too much regarding the nature and the organization of human mental representations of natural kinds. In particular, one might wonder whether there is a class of language of thought terms, natural kind terms, deserving of a special theory of reference all its own. Furthermore, even if this class exists, one might wonder why our theory of extension for natural kind terms should treat these terms as semantically basic.

A fair amount of work in developmental psychology suggests that from an early age (ca. four to five years old), humans treat terms for natural kinds differently than they treat terms for other kinds (for example, artifactual kinds) (see Keil 1989, Markman 1989, and the work reviewed in Kornblith 1993). Thus, later in development, there would seem to be a class of language of thought terms that are marked as natural kind terms. However, this does not address cases of terms to which BT1 applies, if I am correct in identifying these as terms that emerge very early in cognitive development (probably the first two to three years of life). Thinking pragmatically, we might note that at this point in the history of the development of naturalistic semantics, it would seem a worthwhile achievement to provide a thoroughly naturalistic account of extension-fixing for *any* class of terms in the language of thought, on the idealizing assumption that these terms are distinct from any others that might exist. And to simplify the development of a semantics for our chosen class of terms, we can assume that this class of terms (natural kind terms, in the present case) is semantically independent. The diehard naturalist may prefer a reductionist characterization of the project, however, one that assumes a naturalistic semantics to be the only legitimate semantics we can ever have.[26] Here I do not defend this view,

---

[26] Fodor 1990 and 1994 provide reasons for qualification here. Fodor (1990, pp. 110-111) argues that the logical vocabulary in the language of thought (e.g., 'and') can be satisfactorily defined in purely functional terms, and that

though I consider some of its implications in the next section.

3.3 BTT and the Natural-Kinds-Only Assumption

BT1 solves the disjunction problem by making the natural-kinds-only assumption. By calculating success rates for homogeneous natural kinds only, BT1 assigns, in the typical case, extensions that are not disjunctive.[27] If, on the other hand, BT1 were to allow the consideration of the success rate for the disjunctive kind *horses or cows which are encountered under the specific circumstances that make them look like horses*, then BT1 could not solve the disjunction problem. Of course, the natural-kinds-only assumption alone does not solve the disjunction problem. The natural-kinds-only assumption only has the effect of limiting our options when considering candidates for the extension of a given natural kind term in the language of thought. A theory of extension must go beyond the natural-kinds-only assumption and tell us, as BT1 does, which natural kind, among those whose members cause the tokening of *t*, is the kind whose members constitute the extension of *t*. The question remains as to why are we allowed to simply ignore the success rates of strange, disjunctive sets in the application of BT1 in the first place. In

---

thus, a causal theory of extension would not be needed for such terms. Furthermore, Fodor (1994, pp. 70-71) argues that the meaning of the logical vocabulary in the language of thought must be fixed in order to solve Quine's problem of the inscrutability of reference within the framework of a causal theory of reference. Thus, if Quine's problem is to solved in the manner Fodor suggests, a theory of extension for natural kind terms must be complemented by a theory of content for the logical vocabulary.

[27] Throughout this section, when I talk about disjunctive extensions, I have in mind disjunctive extensions which are problematic in the way codified by the disjunction problem. In contrast, some extensions may be truly disjunctive, and thus should be assigned as such by BT1. Below I discuss truly disjunctive extensions in slightly more detail.

.

what follows, I discuss two justifications of the natural-kinds-only assumption, the first based on the reductive nature of the present enterprise, the second on empirical considerations.

On one reasonable understanding of naturalistic semantics, the project is motivated by the thought that we can fit intentional relations into the framework provided by those sciences that do not give intentional explanations. According to this view, we are not attempting a theoretical reduction (of, say, psychological laws to biological laws), but we are attempting to explain, in a reductive way, the intentional, relational predicate 'x refers to natural kind X' in the vocabulary of the natural sciences. Seen in this light, the naturalist's challenge begins with the presentation of two lists, each of which can be thought of as a list of types. On one list, we put language of thought term types, which, for the project to be naturalistically respectable, are to be characterized non-intentionally.[28] On the other list, we enumerate the full complement of natural kinds, where what goes on this list is determined entirely by the non-intentional sciences. The question is, "How do we define a relation between the types on the first list and those on the second so that the relationship is sufficiently like what we think reference is (or what good psychological explanations tell us it should be)?" Given that we wish to frame the answer solely in terms of causal relations holding between the kinds offered by non-intentional sciences, it's not at all clear why we should introduce kinds such as *horse or cows on dark nights* into the mix. The non-intentional sciences don't offer such kinds to us, so these kinds are simply not in the running, at least not from the reductive standpoint I'm advocating. One may be tempted to count the kind *everything that ever causes 'horse' tokenings* as a kind offered by the non-intentional sciences on the sole grounds that we can characterize this kind in non-intentional terms.

---

[28] Dynamical systems theory provides a promising set of tools for use in characterizing types of human language of thought terms non-intentionally. See the essays in Port and Van Gelder 1995, especially those by Esther Thelen, Jeffrey Elman, and Jean Petitot, for ideas as to how dynamical systems theory can be employed to individuate mental representations in the human cognitive system.

However, we should not assume that just because we can characterize a kind in terms of causal relations between physical events, we have therefore captured a natural kind. The kind ‘every solid object that has ever caused an automobile windshield to shatter’ may be naturalistically respectable in the sense that it is definable, via a brute list, in non-intentional terms. Yet this hardly makes the kind a natural kind, i.e., the subject of generalizations in the non-intentional sciences; it’s very unlikely that it will appear any time soon in the laws of physics, chemistry, biology, or even neurophysiology. This is enough reason to ignore such kinds when naturalizing the semantics of natural kind terms in the language of thought.

It seems, then, that to a great extent, the naturalist has been struggling to demystify reference on someone else’s terms.[29] For this reason, it is misleading to frame the naturalist’s problem of intentionality in the way I did at the outset of the present work. We are not simply

---

[29] Fodor seems to take the approach I advocate in the text when he makes claims that only nomic connections are to be considered for the purposes of applying his theory of content (Fodor 1990, pp. 100-103, 121, Loewer and Rey, p. 257). However, Fodor allows an odd lot 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 1990, p. 121) It’s not clear what the ‘qua’ is supposed to amount to here. Perhaps what Fodor has in mind 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. But if ‘nomic’ is taken to mean lawlike, it would seem to be stretching things a bit to characterize as nomic the relationship between being a cow on a dark night and ‘horse’. If this is a lawlike relation, why do cows on dark nights only sometimes (in fact, very rarely) cause the tokening of 'horse' in humans? The change in perspective that I’m suggesting would not substantially alter the presentation of Fodor’s asymmetric dependence theory. Instead of saying that cow-on-a-dark-night caused ‘horse’ tokens are asymmetrically dependent on horse caused ‘horse’ tokens, Fodor can say that cow caused ‘horse’ tokens are asymmetrically dependent on horse caused ‘horse’ tokens.

looking at any and all causal connections, trying to say which ones determine reference. Instead the appropriate question is, "Given the natural kinds described by the non-intentional sciences, what relation between these kinds could be sufficient for an intentional relation to hold?" Implicit in this question is a way of conceiving of the naturalistic project that seems most deserving of the title. The natural-kinds-only assumption should thus be seen as the placement of naturalistic semantics for natural kind terms onto the firm ground of theoretical consistency.

A second justification for the natural-kinds-only assumption follows from the claim that a naturalistic theory of extension for natural kind terms in the language of thought should serve up what psychology needs. Cognitive psychologists make the general assumption that natural kind terms refer to homogenous natural kinds, and this assumption results in more empirically powerful psychology than we would have without this assumption. Given these theoretical ground rules set by psychologists, we have all the more reason to only put natural kinds on our list of kinds with which language of thought terms are to be paired. According to the type of psychological theories that are the most successful empirically, strange disjunctive sets are just not the right kind of things for natural kind terms to refer to.[30] (Certain exceptions should be noted; for example, when two or more natural kinds tie for first or are very nearly tied for first in the success rate derby, and when no other contenders are close, as would be the case in Putnam's jadeite/nephrite example [Putnam 1975, p. 241], the extension may be recognized as truly disjunctive; more on truly disjunctive extensions below).

A criticism of the preceding argument runs as follows. In the interest of creating an empirically powerful psychology, it has been suggested that natural kind terms of the language of thought refer only to causally homogeneous natural kinds. However, it seems that we could

---

[30] I do not here argue for the assumption that a natural-kinds-only-based psychological theory is empirically superior to psychologies which accept, for natural kind terms in the language of thought, reference classes of the sort that cause the disjunction problem. However, notice that if the former theory is not the more powerful of the two, the naturalist has substantially less reason to think that it's important to solve the disjunction problem.

use the same style of argument to assign specific extensions to individual natural kind terms of the language of thought, putting BTT out of business If we can get rid of disjunctive extensions simply because doing so gives us better psychology, then why bother at all with BTT? Why not just assign horses as the extension of 'horse', cows as the extension of 'cows', etc.? Doing so yields better psychology, doesn't it?

Considerations of empirical fruitfulness alone should not be used to assign extensions to individual terms. BTT, or some such theory, is needed for this task. The stipulation that 'horse' refers to horses, 'cow' refers to cows, etc., gives our psychology the empirical power we desire. However, the addition of a theory such as BTT is illuminating because it tells us what all of these empirically superior extension assignments have in common. Given the preference in the sciences for laws and generalizations, this commonality is something we would want included in our best psychological theory (at the very least, as part of its philosophical underpinnings). In contrast, in the choice of the natural-kinds-only framework in the first place, there is no parallel explanatory work to be done. When we choose a framework for our theorizing, the empirical success of working within this framework is the only justification of our choice. We do not make several concurrent framework commitments, the commonalities between which must be explained by a general theory of reference. Unlike the assignment of extensions to individual terms, the choice embodied by the natural-kinds-only assumption is a singular choice of a theoretical framework within which to operate.

Are there other reasons for adopting the natural-kinds-only assumption, beyond the assumed added empirical power of natural-kinds-only-based theories? As noted above humans are disposed from an early age to treat natural kinds as such (Markman 1989, Keil 1989, Kornblith 1993). Furthermore, it would seem to be advantageous to humans to treat natural kind terms as if they referred to natural kinds, even if this means that they have to treat some tokenings of some terms as mistakes (See Clark 1993 where he spells out in detail the information-theoretic advantages of treating some of our language of thought term-tokenings as mistakes). Add to this the principle that if it's advantageous to do x and doing x comes naturally (seems innate) to the

members of a species, then x was probably evolutionarily chosen for because of the advantages x offers to members of that species. From all of this, we can infer that natural kind terms, as a type, may have the evolutionary function of referring to causally homogeneous kinds.

The suspicion may arise at this point that I am making an optimizing assumption similar to the one I criticized above when discussing Godfrey-Smith's work. However, there are two important differences between that case and the present one. Firstly, I am not suggesting, as Godfrey-Smith seems to be, that individual human mental representations have evolutionarily determined content. I am only suggesting that as a type of terms, natural kind terms in a human language of thought have an identifiable evolutionary function, i.e., to have a certain kind of content.[31] More importantly, the empirical facts about humans seem to favor the current use of evolutionary assignments of function over the use to which Godfrey-Smith puts the same conceptual tool. Recall how Godfrey-Smith's view assigns intentional content on the basis of optimal reactions to the immediate environment. However, as argued above, this way of assigning content rests on the faulty view that we can make the best sense of human behavior by assuming that humans respond (or have responded) optimally to their environment. In contrast, we find empirical facts in our favor when we assume that the function of natural kind terms in the language of thought is to refer to causally homogeneous classes. Humans engage in just those behaviors that are necessary for a system to take advantage of the benefits offered by treating their natural kind terms as referring to homogeneous, rather than disjunctive, classes. Such behaviors include multiple scannings of the environment, together with resistance to the idea that an object of one type suddenly changes into an object of a different type during rescanning (Clark 1993, pp. 304-310) (An inclination to construct a picture of the world out of mutually exclusive basic level categories would also seem to be operative here; see Markman 1989 for empirical evidence for what she calls the 'mutual exclusivity constraint' on categorization.)

---

[31] General suggestions along these lines have been made by Sterelny (1990, p.138) and Neander (1995, p. 136).

I am hesitant to place too much emphasis on the evolutionary argument. The reduction-motivated argument given at the outset seems to me to be much stronger. However, if one is inclined to think that, generally speaking, human thoughts' being about natural kinds matters to the science of human psychology, then it is not out of the question that human thoughts' being about natural kinds was chosen for by evolutionary processes, as would seem to have been the case for many of our cognitive capacities (e.g., the ability to use language).

4. The Qua Problem

In their critical discussion of causal theories of reference for natural kind terms in a natural language, Devitt and Sterelny raise what they call the 'qua problem' (Devitt and Sterelny 1987, pp. 63-65, 72-79), and it would seem that an analogous problem can arise for a causal theory of reference for natural kind terms in the language of thought. Causal theories of reference for natural language face the qua problem as a result of 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 the natural kind *kangaroo* is also a sample of the natural kind *mammal*. Thus, even if a speaker consciously intends that *t* refer to all members of the same natural kind as a sample at hand (as Putnam suggests, see Putnam 1975, chapter 12), the speaker will, in many cases, fail to pick out a single natural kind whose members make up the extension of *t*. Devitt and Sterelny solve the qua problem by claiming that the speaker applies a description to the sample she faces (Devitt and Sterelny 1987, p. 75) so as to focus, for example, on the kangaroo-ish nature of the sample rather than its mammalian nature. This attempt to solve the qua problem seems wanting. The speaker employs descriptions, either spoken out in natural language or in the form of descriptive intentions, to isolate the intended reference class. In this way, Devitt and Sterelny's solution assumes that the speaker has available to her descriptive terms, either in a natural language or in the language of thought, that already have their content fixed. This situation should be of great

concern to causal theorists. Regress threatens when we endorse a theory of reference which requires that for the extension of any natural kind term to be fixed, other extensions must already be fixed.[32] Regardless of how things turn out for causal theories of reference for natural kind terms in natural languages, a similar problem of dependence, and looming regress, arises when we focus solely on the language of thought.

There is a hierarchy of kinds in nature, and this seems to be the root cause of the qua problem. Because of the inclusion relations that constitute this hierarchy, we cannot simply point to a sample member of a natural kind and say that we want *t* to refer to members of 'that natural kind'. The specific question to be addressed here is whether BT1 can be of any help in fixing determinate reference for natural kind terms in the language of thought in a way that both avoids the qua problem and does not require that the subject be capable of formulating contentful intentions before the extension of a new term can be fixed, as does Devitt and Sterelny's solution.

BT1 solves the qua problem handily. Assume that BT1 applies both to 'tiger' and 'mammal'. BT1 should assign only tigers to the extension of 'tiger' and only mammals to the extension of 'mammal'. First off, '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', and this drastically reduces their success rate relative to 'tiger' by making the denominator in the relevant calculation much bigger than the numerator. Tigers themselves are much more efficient at causing tokenings of 'tiger' than mammals are, and thus, BT1 assigns tigers, not mammals, to the extension of 'tiger'.

---

[32] Sterelny 1990 (pp. 134-140) suggests a solution to the qua problem that puts an end to the looming regress. Sterelny's approach is different from mine in that he claims that in the case of basic reference, reference at the level where descriptive intentions play no role in fixing extension, it is teleology, rather than a principle such as BT1, which fixes reference.

Why, though, doesn't this same reasoning apply to that subset of tigers (call them the 'stereotypical tigers') that is more efficient at causing 'tiger' in me than are the other tigers? The stereotypical tigers are bound to have a higher success rate relative to 'tiger' than all tigers have.[33]

There are three reasons to doubt that this problem is fatal to BT1's solution of the qua problem. First, the set of stereotypical tigers would not seem to be a natural kind, and thus we are not to consider it as a candidate for the extension of 'tiger'.

To see the second reason, I must now confess that the proper application of BT1 is more complicated than I've made it out to be. Concerns about unhappy disjunctive extensions motivated the introduction of BT1. However, we want to allow that there may be some truly disjunctive extensions. Therefore, *t* has a truly disjunctive extension when the following conditions, (a), (b), and (c), are met:

> (a) two or more natural kinds have equal or roughly equal success rates relative to *t*; (b) no other natural kind has a success rate substantially higher than those kinds whose success rates are equal or roughly equal relative to *t*; and (c) the gap between the group of success rates at the top and those farther down is substantial.

Given this refinement in our interpretation of BT1, even were we to count the set of stereotypical tigers as a natural kind (though I see no reason to), so long as the success rate of tigers as a whole and the success rate for stereotypical tigers are clustered together at the top of the success rate heap, they will all be in the extension of 'tiger'.

The third solution to the stereotypical tiger problem is a dissolution of sorts. If neither of the first two solutions applies, it seems that the problem dissolves itself. Assuming 'tiger' is a term to which BT1 applies, it's not clear which good psychological explanations would motivate

---

[33] This example was proposed in a slightly different form by an anonymous referee.

us to require that all tigers, rather than just the stereotypical ones, be in the extension of ‘tiger’.

Let us return to the language of thought term 'mammal'. If S believes that all tigers are mammals, 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, possibly as high as the success rate of mammals relative to 'mammal', depending on the details of the individual case. How does BT1 identify mammals, rather than tigers, or tigers *together with* mammals, as the true extension of 'mammal'?

The simplest approach to take here in defense of BT1 is to concede all, and note that doing so does no harm. If tigers have as high a success rate relative to 'mammal' as mammals do, then BT1, as modified above, assigns the mammals *and* tigers as the extension of 'mammal'. However, extensionally speaking (which is, after all, how we’re speaking), this is precisely the assignment we want. The union of the collection of tigers and the collection of mammals is identical to the collection of mammals. 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 redundant; it doesn't change the identity of the group.

Difficulties may persist when we attempt to assign extensional contents, as opposed to extensions, to language of thought terms. The distinction is often made between extension, or reference, on the one hand and extensional content, or referential content, on the other. The reference of a term is that thing or group of things to which the term actually refers. Referential content is determined by the *type* of thing to which a term refers. Talk about referential content is theoretically useful in that it allows us to state psychological regularities in terms of the types of things a given language of thought term refers to. Even if two subjects have different language of thought terms with numerically different extensions (as will often be the case when we compare the extensions of natural kind terms for non-contemporaneous subjects), we can fruitfully generalize on their behavior if their two different terms refer to the same type of thing. (For discussions of the theoretical role of referential content, as opposed to reference itself, see 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). A potential problems arises because the abstract type *mammal* is different from the abstract type *tiger or mammal*. These two types correspond to two different properties. Does BT1 imply that the referential content of 'mammal' is the abstract type *tiger or mammal*? If so, isn't this an embarrassment to BT1?

Referential contents are, first and foremost, determined by extensions. The referential contents of language of thought terms are typed according to the kinds to which they refer. But in order for a term to have a content assigned to it on the basis of the kind of thing it refers to, the term must first determinately refer to something.[34] To make referential content theoretically prior to actual extensions would be to make referential content independent of reference in the way that intensions are often thought of as being. Doing so would seem to erase any value of talking about such content as *referential* and would turn referential contents into something more like Fregean senses. Thus, given that the present project consists in trying to give a theory of extension without adverting to traditional intensions (other than the success rate function, which could be seen as an intension), our only option seems to be to make reference prior to referential

---

[34] There are complications here in assigning referential content to terms such as 'unicorn' which have no extension. This is not to say that assigning referential content to such terms is an impossible feat. Some theories of content that, generally speaking, taxonomize contents by extension do so even for terms such as 'unicorn' (see, for example, Fodor 1990, pp. 100-101). Extensions are still basic in such theories in that in order to assign the referential 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 1990, p. 116].) One way to fill out this view is to say that all terms such as 'unicorn' are complex terms whose parts have actual extensions. (For present purposes, this amounts to saying that terms such as 'unicorn' are not terms to which BT1 applies.) We can then assign referential content to terms such as 'unicorn' by referring to the extensions of their simple parts (which parts have actual extensions). This approach has implications which some find unacceptable (Baker 1991, p. 21, Boghossian 1991, p. 77). However, a full treatment of such terms is beyond the scope of this paper.

content. Having solved the superordinate problem for extensions, then, we have automatically solved the superordinate problem for referential contents. The referential content of 'mammal' is the natural kind membership in which is shared by all and only those things in the relevant extension. The extension of 'mammal' is the group of actual mammals. Therefore, BTT assigns *mammal* as the referential content of 'mammal'.

5. Prospects for BTT's Expansion

At the outset, I noted the difference between natural kind terms in the language of thought that have their extensions determined independently of the content of any explicit intentions on the part of the subject and those which do not fall into this category. Throughout the paper, I have been concerned almost exclusively with the former case. However, we must recognize that many, perhaps even most, natural kind terms in the language of thought are of the latter type. Were further principles BT2...BTn to be added to BTT, they would jointly have to address the full variety of methods by which reference is fixed. By including such further principles, we could, for example, account for familiar facts regarding the division of linguistic labor (Putnam 1975, chapter 12). Further principles should also account for the various ways in which humans qualify the reference-fixing process by limiting the candidate kinds to local kinds or kinds with which the subject has interacted. Imagine, for example, that we're faced with a standard subject whom we would take to be thinking about dogs when she tokens 'dog'. But imagine also that relative to her language of thought term 'dog', some species of dogs have high success rates while other species have very low success rates.[35] If BT1 were to apply here, it would assign only the dogs whose species have high success rates to the extension of this subject's language of thought term 'dog'. But via some principle beyond BT1, BTT can yet assign dogs as the extension of 'dog'; this by taking into consideration the fact that the subject intends 'dog' to

---

[35] This is a slightly modified version of an example suggested by an anonymous referee.

refer to a natural kind to which all of the samples belong (samples of species with high success rates as well as samples taken from species with low success rates relative to ‘dog’). Of the natural kinds to which all of the sample dogs belong, the kind dogs has the highest success rate relative to the subject’s language of thought terms ‘dog’. Furthermore, allowing subjects’ intentions to come into play may explain how humans can refer to species the presence of whose members is difficult to detect. While considerations of space prevent my going into detail regarding principles beyond BT1, I hope that I have given the reader some idea of how the basic framework of comparing success rates could be qualified and extended to allow a subject to use language of thought terms whose contents have already been fixed by BT1 to fix the extensions of further natural kind terms in the language of thought.

Works Cited

Akins, K. 1996: Of Sensory Systems and the "Aboutness" of Mental States. *Journal of Philosophy*, 93, 337-72.

Astington, J. W. 1993: *The Child's Discovery of the Mind*. Cambridge, MA: Harvard University Press.

Baker, L. R. 1991: Has Content Been Naturalized? In B. Loewer and G. Rey 1991,17-32.

Boghossian, P. A. 1991: Naturalizing Content. In B. Loewer and G. Rey 1991, 65-86.

Bower, T. G. R. 1989: *The Rational Infant: Learning in Infancy*. New York: W. H. Freeman and Company.

Brown, D. 1996: A Furry Tile about Mental Representation. *Philosophical Quarterly*, 46, 448-66.

Burge, T. 1986: Individualism and Psychology. *Philosophical Review*, 95, 3-45.

Cherniak, C. 1986: *Minimal Rationality*. Cambridge, MA: MIT Press.

Churchland, P. 1981: Eliminative Materialism and the Propositional Attitudes. *Journal of Philosophy*, 78, 67-90.

Clark, A. (Andy) 1991. In Defense of Explicit Rules. In W. Ramsay, S. P. Stich, and D. E. Rumelhart (eds.), *Philosophy and Connectionist Theory*. Hillsdale, NJ: Lawrence Earlbaum Associates, 115-28.

Clark, A. and Toribio, J. 1994: Doing without Representing?. *Synthese*, 101, 401-31.

Clark, A. (Austen) 1993: Mice, Shrews, and Misrepresentation. *Journal of Philosophy*, 90, 290-310.

Cummins, R. 1989: *Meaning and Mental Representation*. Cambridge, MA.: MIT Press.

Devitt, M. 1994: The Methodology of Naturalistic Semantics. *Journal of Philosophy*, 91, 545-72.

Devitt, M., and Sterelny, K. 1987: *Language and Reality*. Cambridge, MA.: MIT Press

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

Dretske, F. 1986: Misrepresentation. In R. J. Bogdan (ed.), *Belief: Form, Content and Function*. Oxford: Oxford University Press.

Dretske, F. 1988: *Explaining Behavior*. Cambridge, MA.: MIT Press.

Dupre, J. 1981: Natural Kinds and Biological Taxa. *Philosophical Review*, 90, 66-90.

Field, H. 1990: "Narrow" Aspects of Intentionality and the Information-Theoretic Approach to Content. In E. Villanueva (ed.), *Information, Semantics & Epistemology*. Oxford: Blackwell.

Fodor, J. 1986: Banish DisContent. In J. Butterfield (ed.), *Language, Mind and Logic*. Cambridge, UK: Cambridge University Press. Reprinted in W. Lycan (ed.), *Mind and Cognition: A Reader*. Oxford: Blackwell, 420-38. Citations are to Lycan.

Fodor, J. A. 1987: *Psychosemantics*. Cambridge, MA.: MIT Press.

Fodor, J. A. 1990: *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, 3-71.

Forster, K. 1990: Lexical Processing. In D. Osherson and H. Lasnick (eds.), *An Invitation to Cognitive Science, Vol. 1*. Cambridge, MA.: MIT Press, 95-131.

Godfrey-Smith, P. 1991: Signal, Decision, Action. *Journal of Philosophy*, 88, 709-22.

Godfrey-Smith, P. 1992: Indication and Adaptation. *Synthese*, 92, 283-312.

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, 355-65.

Gleick, J. 1987: *Chaos: Making a New Science*. New York: Penguin.

Horgan, T. and Tienson, J. 1996: *Connectionism and the Philosophy of Psychology*. Cambridge, MA.: MIT Press.

Horgan, T. and Woodward, J. 1985: Folk Psychology is Here to Stay. *Philosophical Review*, 94.

Hull, D. 1978: A Matter of Individuality. *Philosophy of Science*, 45, 335-60.

Jackendoff, R. 1989: What is a Concept, that a Person May Grasp It?. *Mind and Language*, 4,

68-102.

Karmiloff-Smith, A. 1992: *Beyond Modularity: A Developmental Perspective on Cognitive Science*. Cambridge, MA: MIT Press.

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

Kitcher, P. 1984: Species. *Philosophy of Science*, 51, 308-33.

Kitts, D. B. and Kitts, D. J. 1979: Biological Species as Natural Kinds. *Philosophy of Science*, 46, 613-22.

Kornblith, H. 1993: *Inductive Inference and Its Natural Ground*. Cambridge, MA, MIT Press.

Kripke, S. 1980: *Naming and Necessity*. Cambridge, MA.: Harvard University Press.

Lakoff, G. 1987: *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. Chicago, IL: University of Chicago Press.

Lange, M. 1995: Are There Natural Laws Concerning Particular Biological Species? *Journal of Philosophy*, 92, 430-51.

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

Maloney, J. C. 1994: Content: Covariation, Control, and Contingency, *Synthese*, 100, 241-90.

Markman, E. M. 1989: *Categorization and Naming in Children*. Cambridge, MA.: MIT Press.

Marr, D. 1982: *Vision*. New York: W. H. Freeman and Company.

Massey, C. M., and Gelman, R. 1988: Preschooler's Ability to Decide Whether a Photographed Unfamiliar Object Can Move Itself. *Developmental Psychology*, 24, 307-17.

Maudlin, T. 1986: *Reasonable Essentialism & Natural Kinds*. Doctoral Dissertation, University of Pittsburgh.

Millikan, R. G. 1984: *Language, Thought, and Other Biological Categories*. Cambridge, MA.: MIT Press.

Millikan, R. G. 1990: Truth Rules, Hoverflies, and the Kripke-Wittgenstein Paradox. *Philosophical Review*, 99, 323-53.

Mishler, B. D., and Donoghue, M. J. 1982: Species Concepts: A Case for Pluralism. *Systematic*

Zoology, 31, 491-503.

Montgomery, R. 1989: Discrimination, Reidentification and the Indeterminacy of Early Vision. *Nous*, 23, 413-35.

Neander, K. 1995: Misrepresenting and Malfunctioning. *Philosophical Studies*, 79, 109-41.

Port, R. F. and Van Gelder, T. (eds.) 1995: *Mind as Motion*. Cambridge, MA: MIT Press.

Putnam, H. 1975: *Mind, Language, and Reality*. Cambridge: Cambridge University Press.

Shapiro, L. A. 1993: Content, Kinds, and Individualism in Marr's Theory of Vision. *Philosophical Review*, 102, 489-513.

Smith, E., and Medin, D. 1981: *Categories and Concepts*. Cambridge, MA: Harvard University Press.

Sober, E. 1980: Evolution, Population Thinking, and Essentialism. *Philosophy of Science*, 47, 350-83. Reprinted in E. Sober (ed.), *Conceptual Issues in Evolutionary Biology* (second edition), Cambridge, MA: MIT Press (1994).

Sober, E. 1984: Sets, Species, and Evolution: Comments on Philip Kitcher's "Species". *Philosophy of Science*, 51, 334-41.

Spelke, E. S. 1990: Origins of Visual Knowledge. In D. Osherson, S. Kosslyn and J. Hollerbach (eds.), *Visual Cognition and Action: An Invitation to Cognitive Science, Vol. 2*. Cambridge, MA.: MIT Press., 99-127.

Spelke, E. S. 1991: Physical Knowledge in Infancy: Reflections on Piaget's Theory. In S. Carey and R. Gelman, *The Epigenesis of Mind*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 133-69.

Sterelny, K. 1990: *The Representational Theory of Mind*. Oxford: Basil Blackwell.

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.

Streri, A., and Spelke, E. S. 1988: Haptic Perception of Objects in Infancy. *Cognitive Psychology*, 20, 1-23.