How (not) to give a theory of concepts

Draft 1

This paper presents the lineaments of a new account of concepts. The foundations of the account are four ideas taken from recent cognitive science, though most of them have important philosophical precursors. The first is the idea that human conceptuality shares important continuities with psychological faculties of other animals, and indeed that there is a well-distinguished hierarchy of such faculties that extend up and down the phylogenetic scale. While it would very likely be a mistake to look at some conglomeration of these simpler abilities and assert that we have produced a reductive account of human conceptuality, an examination of these will lend insights into essential features of human conceptuality in a non-reductive, non-exhaustive manner. The second idea is that an important function of both human concepts and of their protoconceptual ancestors in the animal kingdom is to make distinctions or discriminations. We shall thus look at the human conceptual apparatus as being, in large part, a discrimination engine. How are these discriminations realized in humans and other beings? Presumably, some discriminative mechanisms are innate, while others are acquired through learning. But how is learning accomplished? The third idea from cognitive science is that adaptive discrimination is realized through neural networks, and that the properties of this realizing system explains familiar features of human thought that seem puzzling when viewed through other lenses, such as the logical analysis of language. The fourth and

final key idea is that the mind is not a single central processing unit, but is highly *modular*-that is, it has faculties within it that are optimized for particular tasks.

This view of concepts has important philosophical payoffs. These come in the form of making sense of a range of philosophical issues that have generally been regarded as problematic in analytic philosophy. Among these are vagueness and indeterminacy of concepts and predicates, pragmatic constraints on the assessment of truth conditions, puzzles concerning how to understand animal cognition, difficulties in the incommensurability and irreducibility of local discourses within a single language, and the disunity of science. The plan of this paper is as follows: first, I shall locate the new approach to concepts against a backdrop of a more familiar, logically-motivated analysis of concepts modeled upon the logical analysis of language, and give some reasons why an alternative analysis is called for. I shall then proceed to outline the alternative account. Finally, I shall very briefly describe some of its possible fruits by showing that several issues that seemed philosophically problematic before are explainable as artifacts of the way conceptuality and cognition are realized in human beings.

Extensional Semantics -- Why Do We Need a New Theory of Concepts?

One might naturally ask why a new theory of concepts is needed. In particular, one might ask why existing semantic theories do not already provide a theory for concepts, just as they do for linguistic predicates. While Frege, viewed by many as the father of analytic philosophy, thought that a semantic analysis of concepts needed to focus on different things than a semantic analysis of words [], most recent philosophers in

the analytic tradition have taken the view that concepts and predicates are much the same for the purposes of semantics, in that their semantic properties are defined in terms of their extensions—the set of objects that satisfy the concept or the predicate. They differ, of course, in that predicates are parts of a public language while concepts have troublingly mentalistic overtones, but this need not matter to their semantic analysis.

This view of the semantics of concepts is part of what I shall call *the logical approach to language*. The approach is "logical" in the sense of being motivated by the concerns of logicians, and not in the laudatory sense. It is no secret that philosophy of language has been driven in the past century in large measure by concerns arising out of logic, and that many of its most influential innovators were and are, first and foremost, logicians. Logic is, by its nature, focused upon questions of *truth* and *derivability*, the first a feature of sentences, statements or thoughts and the latter a kind of relationship between these within a derivational system. Since truth is concerned with the relationship between linguistic tokens and non-linguistic states of affairs, it is natural that one would adopt an extensional analysis of the truth-values of statements: e.g., that if *P* is a primitive predicate, and *x* denotes an object, then *Px* is true just in case *x* is in the extension of *P*. The formalizers showed the way to another class of truths that are true on purely syntactic grounds: given the axioms of a system and formalized rules for derivation, a set of theorems is demonstrable as true within that system. To this Tarski added a semantic account of truth for things other than predicates, framed in terms of recursive functions.

The Logical Empiricists who set the agenda for 20th century philosophy of language were well aware that natural language did not look very much like the kind of

purified language they desired on logical grounds. In order to have a language that lacked ambiguity and indexicality and which conformed to the principle of bivalence it was necessary either to reform natural language (Russell[]) or ignore it (Tarski[]). Concerns for completeness and finite recursive definability were arguably only appropriate to the analysis of formal deductive systems in the first place. Attempts to deal with the social and public character of language eventually led to such bizarre theories as Quine's radical indeterminacy thesis [] and the view of Davidson [] and Lewis [] that we each speak our own language. Because logic is concerned with the truth values of symbols, philosophy of language spent much of the century ignoring the role of the speaker and of the utterance until the work of Grice [] and Austen [] in the 1950's, and likewise ignored the semantics of the vast panoply of speech acts (promises, apologies, performatives) that have no truth value because they are not used to assert anything.

Today, it is perhaps the predominant view in analytic philosophy that to "give a semantic theory" for a language is to provide a function from referring expressions onto their referents and from sentences onto the states of affairs that would make them true. In this sense, a "language" is a kind of ideal or abstract object [Lewis…], which can then be adopted or used by individuals or linguistic groups. This view of semantics is also embraced by many people in philosophy of mind, who view the project of "giving a theory of content" for mental states in terms of supplying a mapping from "tokens in the language of thought" to their referents and from "sentences" in this "language" to the states of affairs that would make them true. []

This logical approach to semantics is not without its detractors. Writers like Putnam [], Field [] and Blackburn [], for example, have pointed out that logical accounts

of semantic properties like meaning, reference, and truth at most provide a set of mappings from semantic primitives to objects, sets and classes, and a method for generating the semantic values of complex semantic objects as a function of the values of the primitives plus the syntax. But a mapping is not an *explanation* of a semantic property. A semantic "theory" that digs no deeper than a function from predicates to extensions or from sentences to truth values is no account of the nature of meaning, reference or truth. Moreover, the programme of linguistic reform endorsed by some of the Logical Empiricists has long since fallen out of favor. Semantic theories like that of Tarski [] which aim at such logically-motivated properties as consistency, bivalence, and closure achieve these only by restricting their scope to carefully circumscribed "languages" that do not have properties like indexicality, indeterminacy and vagueness, which are possessed by all natural languages. Since some of the theoretical desiderata of logical analysis of language are at odds with features of actual languages, we might do well to ask what alternative theoretical perspective might help us complete the picture.

Philosophers have also increasingly payed attention to the features of natural language that are unlike those of formal deductive systems. For example,

- As Austen pointed out, a truth- and inference-oriented approach to language leads us to forget that there are kinds of speech acts (perhaps the majority of actual speech act tokens) whose pragmatic aims are something other than truth. Yet these speech acts also have semantic properties beyond the sense and reference of the lexical units that comprise them. And even in the case of statements, it is not sentences but speech acts that are the bearers of semantic properties.
- Extensional logic has notorious problems in dealing with fictitious and imaginary referents.
- There are kinds of ways predicates and statements can be good and bad that fall outside of the web of the logical approach: for example, whether we are prepared to call a scientific claim about how the world is (e.g., a law of classical mechanics) *true* or *false* depends in large measure upon context. It is not simply that we are willing to *pretend* classical mechanics is true in some contexts, even though we "know better"; rather, our assessment of truth values depends not only upon predicates and the

world, but also upon assumptions about things like permissible margins of error in a given context. Thus the person who says "I stayed right where I was while the train moved by me" is not saying something false just because there is no such thing as absolute motion. (Though she would be saying something false if her statement was meant as an assertion that there was absolute motion.) A statement of a scientific law with a truncated representation of a constant with an infinite decimal series is not false just because it is inexact, etc. [Edidin]

- As Nancy Cartwright [] has shown, many scientific laws must either be regarded as false or as not being statements about how objects actually behave (i.e., not properly interpreted in terms of an extension class of objects and events). Because a law—say, of gravitation—idealizes away from other forces (e.g., mechanical force, electromagnetism) that are always at work *in vivo*, does not not state exact truths about actual events. It is not clear how a purely extensional notion of truth can accommodate this.

It is possible to build logical models of languages with such features as indeterminacy, vagueness and multivalence—e.g., in the form of many-valued logics. However, we do so more in an attempt at psychological modeling than at providing an alternative normative model of deduction and proof. This is as much a step out of the logical approach to semantics as a move within it. For a logic whose focus is truth-preserving deduction, such features can never be anything but aberrations to be shuffled off by way of idealization.

To these I should like to add some further concerns that apply specifically to the assimilation of concepts to sentences in a language. First, it is problematic whether it makes sense to extend a linguistic semantics to a semantics of thoughts and concepts, as at least a good portion of semantic conceptuality is continuous with the capacities of animals and pre-linguistic children. Some recent discussions of content in animal thought [] have asked such questions as whether the part of frog's eye that detects flying insects sends signals that "mean 'fly'" or "mean 'small moving dot'". This way of putting the question attempts to account for the frog's protoconceptual structures by translating them into English, or alternatively by giving, in English, a specification of their

extensions. However, it is not clear that this is the natural way to proceed if what we are asking is the question of what role this neural device plays in the inner economy of the frog's life.

More basically, the appeal of extensional semantics in logic is motivated by two prior goals of the logician: (1) an interest specifically in truth viewed as a correspondence between statement and world, and (2) an interest in truth-preserving inference. While it is the extensions of predicates that matter for the question of truth-preserving inference, it is the intensions of thoughts, as Frege recognized [], that matter for how we in fact reason and represent the world. Even if extensional semantics is appropriate for logic, it need not follow that it is appropriate for psychological semantics. In particular, if we view the minds of animals as things shaped by ecological and evolutionary forces, it is not clear that we should see animal conceptuality, or protoconceptuality, as being aimed at truth or even at reference in the case of the lower animals. Selective pressures favor cognitive systems that track the world well enough for practical purposes, and this may fall short of [] or even conflict with [] the production of *true* (i.e., veracious) cognitions, if indeed we may literally speak of "truth" with respect to animal cognition at all. We, of course, *do* have a language, and indeed public language shapes the child's conceptual space. And at least some of us are concerned with truth and with truth-preserving inference. However, even though this is so, our cognitive apparatus is still built upon an animal substrate. It matters greatly whether our minds are designed from the bottom up with things like logical inference and bivalence in mind or whether resources originally designed for very different tasks have been co-opted into these specifically human endeavors. Is the connection between concepts and reasoning a deep and elegant design feature of the

human organism, or is it a rather precarious kludge that appeared a few thousands or tens of thousands of years ago within genus *homo?* To view conceptuality solely from the standpoint of the interests of the logician is to assume the first scenario, and that means that it is to assume an answer to a substantive and empirical question. The tensions between logical theories of language and real human languages might well be taken as evidence that this assumption is mistaken.

More modestly, but no less importantly, we might simply observe that conceptuality is a rich real-world phenomenon. A theoretical stance we take towards it, like any other theoretical stance, will accentuate some features while idealizing away from others that may be equally important *in vivo*, even if they are not equally important to a given theoretical enterprise. The logician's theoretical interests rightly lead him to steer clear of psychologism, as psychological facts cannot determine logical norms. The resulting picture of *concepts* and semantics may or may not be psychologically illuminating. If, on the other hand, we want to study conceptuality in its own right, our priorities are very different: we are interested precisely in the specifics of human psychology (and animal psychology insofar as it sheds light on our own), and are interested in how real humans *do* think rather than how ideal logical reasoners *would* think. Thus, we might view a *biological* perspective on concepts first and foremost as a separate approach which might well turn out to be a supplement to, rather than a competitor with, the logical stance, much as we were forced, in the end, to view wave and particle descriptions of light as different aspects of the same phenomenon rather than competing theories.

Human and Animal Cognition -- A Dilemma

There are good reasons for looking at human conceptuality in a way that stresses its continuities with animal cognition as well as its uniqueness among terrestrial species. While there is much that is radically unique about human cognition, there is much that we do in common with the animals: we perceive through various modalities, and at least in the case of other mammals, through analogous parts of the brain; we learn, like almost all animals, through classical conditioning and, like the social mammals, through imitation and supervised instruction; the commonalities of what creatures with similar physical architectures seem to be able to represent about their environments is at least as striking as the differences; and we seem to find closer approximations of our own minds the closer we come to our own species' location on the phylogenetic tree. Moreover, we have every reason to believe that a great deal about our minds is closely connected to the design of our brains. And here again we find at least as much continuity as diversity: nervous systems in general work on similar principles, and the structural similarities with our own brains increase dramatically as we move out to our own branch in phylogeny. The distinctive feature of human brains, of course, is the radical enlargement of the neocortex, which is responsible for much of what we think of as thought, particularly reasoning and language, and surely is responsible for many of the differences between human and animal cognition. However, the effects of lesions in other parts of the brain vividly demonstrate that much of what goes into human thought is dependent upon parts of the brain that we share more directly in common with other animals. [] (Were this not so, neuroscience could not learn about human cognition through experiments on other species.)

Add to this the persuasive evidence of common sense: it is probably unrealistic to think that my cat reasons things out. But it surely makes sense to say that it wants its breakfast, or is afraid of the dog, or sees the cardinal in the dogwood tree, or that it likes one neighbor cat and not another. Philosophically, we take such attributions with a grain of salt: we know that there are surely some unspecified differences between feline cognition and emotion and our own mental life. (Indeed, when we use proprietary words like 'mental' we are often very cautious about attributing them to other species.) In some cases, we are even sure that we are speaking only metaphorically—e.g., if I say that my cat is "virtuous" or "philosphical". However, when I speak of my cat seeing something, or being afraid, or wanting its breakfast, I do not take myself to be speaking metaphorically or analogously. While I acknowledge that what I call "seeing" or "wanting" in a cat is probably different in important ways from the paradigm uses of those words as applied to myself and other humans, I assume that what goes in in my cat is *continuous with* my own cognition rather than merely *analogous to* it.

And yet there is also reason for extreme caution here as well. In our own mental economy, concepts are wrapped up in a rich web of mental phenomena. It is unclear that we could pry any one of our concepts away from a variety of other mental phenomena—its inferential role in reasoning, its semantic and emotional associations, its connection to words in a public language, our ability to think *about* rather than simply *through* our concepts, or the phenomenology of conscious thoughts involving the concept—without ending up with something altogether different. Indeed, if someone were to attribute a concept to a thing that was commonly agreed to lack consciousness, introspection, semantic connections and inference, we might well wonder whether she

was using the word 'concept' in the standard way. But this creates problems for a biological examination of concepts that highlights the continuities between human conceptuality and animal cognition. It is doubtful that any other terrestrial species possesses either language or the kinds of inference that are dependent upon it. It is at best a vexed question whether other species possess a conscious phenomenology, or whether they can think about their own mental states. And in the absence of full-fledged inference and consciousness, it is unclear what the remaining item on our list—cognitive and emotional associations—might come to beyond the psychological concept of conditioning. If concepts are intrinsically tied to any one of these things that other species lack, then members of these other species lack concepts. And yet this seems absurd.

What we have here is in part an issue over the definition of the word 'concept'—and indeed, over a great portion of the mentalistic vocabulary. But there is a real substantive issue here as well, and some familiar philosophical dangers lurk in the wings. For we *identify* notions like 'concept' *through our own case*, where they are closely tied to the whole rich web of human mentality. Indeed, philosophers have persuasively argued [] that the phenomenon we identify as conceptuality is essentially bound up with other things like inferential role [Brandom] or consciousness [Searle] or public language [Wittgenstein]. If this is the case, then an analysis of concepts that excludes these connections to language, inferential role and consciousness risks what I have elsewhere called the *fallacy of idealization*. This fallacy is committed when we take a rich phenomenon A, idealize away from some of its features in order to give a precise account of some of its salient features in terms of B, and then forget the idealizing move

to proclaim that B provides a complete analysis of A, such as an analysis in terms of necessary and sufficient conditions. It is this fallacy, I think, that lies at the bottom of functionalist and computationalist theories of thought: theorists have subtly slipped over from the interesting and innocent claim that reasoning can be given a characterization in functional terms to the claim that mental states are *defined* (or at least *definable*) in functional terms, as though nothing else mattered *in vivo*.

We thus seem faced with a bit of a dilemma. If we restrict our use of the word 'concept' to things that have the inferential and phenomenological connections found in human conceptuality, then we deprive ourselves of the resources to talk about the continuities between ourselves and other animal species, and indeed seem to paint a wider discontinuity between our minds and theirs than is warranted. If, on the other hand, we use the word 'concept' to embrace some broader territory that we share with other species, then we arguably leave out something that is essential to the paradigm cases of concepts, and risk commiting the fallacy of idealization by saying "concepts are just ….", where the lacunae are filled in in a way that makes no mention of inference or consciousness.

The Role of Idealization in Scientific Theory—Why Theories are Not Analyses

We can do justice both to the connections between conceptuality and inference and consciousness and to the continuities between human and animal cognition if we properly understand the role that idealization plays in scientific theorizing and distinguish a scientific theory from a philosophical analysis in terms of necessary and sufficient conditions. Scientific theories generally require both abstraction and idealization. Real-

world events are complex, and theorizing requires that we cut through the complexity by factoring some things out in order to attend to others. In *abstraction*, we simply ignore some features of a situation, often because they truly do not matter for our explanatory purposes. It was thus an advance in dynamics when Galileo moved away from the Aristotelian framework of explaining motions of bodies in terms of the specific natures (as, say, cabbages or cats) in favor of a few simple variables like mass. For purposes of dynamics, the specific nature of what you drop off the Tower of Pisa does not affect its gravitational acceleration. (Unless, of course, it can fly.) But scientific theories also need to factor out variables that *do* matter *in vivo*. For example, a theory of gravitation ignores the effects of electromagnetism, strong and weak force, and mechanical force. The same holds true, *mutatis mutandis*, for theories of the other fundamental forces. As a result, what any one of these theories ends up saying is *not* an accurate representation of how objects behave in reality. Gravitational theory *would* tell us how objects behave *if there were no other forces at work*. But the other forces are always at work, so gravitational theory *never* tells us *exactly* how real-world objects behave. Nancy Cartwright [] has eloquently discussed this topic, saying that the laws of physics "lie"—i.e., state things that are false. I prefer to say that they state truths, but that the truths they state are not universally quantified statements about objects. Instead, they are *idealized truths:* truths that get at fundamental realities only at the cost of screening out other fundamental realities. We risk the fallacy of idealization only when we miss this distinction, and treat idealized truths as though they were analyses of how things actually behave *in vivo*.

Sometimes, of course, we can retrieve a way of predicting events *in vivo* from several idealized theories—e.g., that we could derive the actual dynamics of objects by doing a summation of the contributions of gravitation, electromagnetism, mechanical force, strong and weak force. (Cartwright is quite skeptical about this.) However, there are certainly other sorts of cases as well, in which it is *not* possible to factor the variables that are at work *in vivo*. In the case of physics, we are in the happy situation of having a relative few fundamental variables, and of having these be *independent* of one another. (Though even in the case of three gravitational bodies the exercise of determining the unfolding of the dynamics is computationally impossible!) Often, however, scientific theory is not so lucky. Sometimes important variables are mutually dependent, either through part-whole relations (as in the relationship between an organism and its ecosystem) or through feedback relations (as in the case of the connections between different areas of the brain). In these cases, the relationship between the insight provided by an idealized theory and its capacity for prediction of real-world results tends to be rather distant. [Cartwright]

(Because the brain is one of most complex feedback systems we know, there are thus principled problems in arriving at neurological or psychological laws that have the degree of accuracy and reliability we find in physics. [] However, this does not so much mark a deep difference between psychology and the "hard" sciences as it does the number of mutually-dependent variables involved in the brain compared with the small number of independent variables we control for in physics.)

Concepts and Protoconcepts

I therefore propose to finesse the dilemma discussed above in the following way. If it is uncontroversial that there are aspects of our conceptuality that are shared with other species—and indeed, that there are increasing commonalities as one climbs out towards *homo sapiens* on our phylogenetic tree limb—but problematic that the paradigm instances of concepts are bound up with things like inference, language and consciousness, we may use another term, such as *protoconcept* to track the continuous elements, and reserve the notion of *concept* for the richly-embedded things we find in our own mental economy. Concepts are, in part, protoconcepts. If we idealize away from inference, language and consciousness in our own case, we are left with a protoconceptual kernel that arguably we literally share with other portions of the animal kingdom.

This does *not* imply that we can simply *build* an analysis of full-fledged conceptuality out of protoconceptuality plus something more, like language or consciousness. For the variables here may not be factorable in this way. Idealization is not simply subtraction, as in a feedback system *removing* one of the modules would not idealize but impair the performance of the others. Nor is embedding a phenomenon in a larger psychological economy simply a matter of addition, as the interactions within the larger economy may change and even redefine the nature of the ingredients.

Our analysis of concepts will thus be undertaken as an investigation of *protoconcepts*. It is thus an investigation of *one aspect* of our concepts, rather than their entire nature. Some of the things we initially factored out—association, language, inference and self-awareness—will eventually be brought back into the picture as special enhancements to protoconceptuality that we find in the "higher" animals. (Some we find

only in humans.) Consciousness and phenomenology will remain outside the picture entirely. The resulting story says a great deal that is revealing about human thought, but it is not meant to say *everything*, and indeed has idealized away from things that are essential. Still, like a theory of gravitation that ignores electromagnetism, it may be of some theoretical use, as we shall see.

The Discrimination Engine

The basic idea I wish to develop is that protoconcepts (and hence concepts) are elements in a system used for discriminations—for dividing the world in useful ways. In some animals, the resources they possess for making discriminations may be entirely innate, and not responsive to experience. We humans, however, do not merely *possess* protoconcepts, but can also create new ones and refine existing ones, an ability arguably shared with a number of other species. I shall develop this idea at greater length in what follows, distinguishing protoconcepts from simple tropisms and schemas and then situating them in their richer context within human conceptuality.

Even very simple animals have some internal mechanisms that distinguish between environmental variables that are correlated with things that are helpful or harmful to the individual or the population. There are small sea organisms, for example, that contain magnetically polarized bits that allow them to stay in oxygen-rich surface water and stay out of oxygen-poor deep water. Depending on which hemisphere they are in, the direction of the north pole will correlate either with a downward or an upward trajectory in the water, and so a magnetic tropism in the animal allows it to navigate to more promising conditions. It is, of course, a very crude hack, as the correlation is only a rough one, and there are many other factors that might affect the quality of living for these little creatures. And if we say that they are *thinking*, "Ooh, let's go up here where the water has more oxygen," we are surely telling a fictional story. In a simple tropism

like this, there is simply a mechanism in the animal that responds to an environmental condition in a fashion that is adaptive, and likely has been selected on an evolutionary time scale for its adaptive value.[1]

Both tropisms and more complex behaviors may be accomplished through some system within an organism which is capable of being of different states, whose function is to track features of the organism, its environment, or the interface between the two. At a first approximation, we might (at some peril) think of these as inner *models*. I say, "at some peril", because the word 'model' usually signifies something that is consciously and explicitly used to represent a reality to which we have an independent means of access. These "inner models", by contrast, need not be conscious, and are what mediates our access to the world. I shall therefore use another word, and refer to the "inner models" plus their sensory inputs and associated capacities for motor control as **schemas**.

As an example of a simple set of schemas, consider Werner Reichardt and Tomaso Poggio's analysis of the visual system of the fly [Reichardt and Poggio, 1976, 1979; Poggio and Reichardt 1976, reported in Marr] as reported by David Marr [1982].

> Roughly speaking, the fly's visual apparatus controls its flight through a collection of about five independent, rigidly inflexible, very fast responding systems (the time from visual stimulus to change of torque is only 21 ms). For example, one of these systems is the landing system; if the visual field "explodes" fast enough (because a surface looms nearby), the fly automatically "lands" toward its center. If this center is above the fly, the fly automatically inverts to land upside down. When the feet touch, power to the wings is cut off. [Marr 1982, pages 32-33]

As Marr observes, "it is extremely unlikely that the fly has any explicit representations of the visual world around him—no true conception of a surface, for

[1] For readers who do not believe in the theory of evolution, I should point out that nothing here depends upon accepting that theory. We could tell a cognate story about tropisms and the things that follow in terms of God's design for an animal instead of a selective story.

example, but just a few triggers and some specifically fly-centered parameters." (p. 34) And this points to an important way of partitioning the class of schemas. Some schemas are, as we may say, **object-centered:** they contain elements that covary with objects and properties of objects. Others are **interface-centered:** their elements covary with relational properties relating the organism with its environment. I shall call the elements in a schematic system that covary with objects and their features "representations". The reader should note that this is being used in a proprietary sense here: I use the word 'representation' here to mean precisely the elements in a schematic system that covary with objects and their features. None of the ordinary associations of the word 'representation'—e.g., conventionality, or semantics, or syntax—are intended to be operative in the technical meaning. This being said, we may say that Marr is right that the fly seems to have no representations, but merely interface-centered schemas. It is also the case that perception, cognition and action do not seem to be distinguished in the fly: the motor control mechanisms are directly driven by perceptual stimuli, without any apparant intervening level at which cognition takes place. Either the fly has no semantics at all, or else there is no distinction between semantics and pragmatics for flies: the activation of the fly's "landing system" might be equally well (or badly) reported by us as saying "there is a surface coming up," and as saying "Brace for impact, laddie!" The fly's brain contains a distinction device, but what it distinguishes are fly-relevant ecological conditions that are not factored out into states of affairs involving objects and properties.

It is quite possible that we have inner schemas of this interface-oriented sort as well. I am thinking in particular of various functions of the autonomic nervous system, which modulate all manner of important biological processes within our bodies, all without our conscious awareness. These would seem to involve mechanisms that respond to inner variables like oxygen level in the bloodstream with control mechanisms that do things like cause us to breathe and release hormones. But there seems little

reason to think that there are object-oriented representations involved in these autonomic control operations.

It is clear that higher animals—all of the mammals, probably all of the birds, quite likely a great deal more—can also do more than this. First, they seem to be able to model their environment in ways that involve treating the world as involving things or objects. A bird's responses in flight to a sudden updraft may be like the fly's landing mechanism, but when a robin listens for a worm, or a sparrow retrieves a piece of straw for its nest, or a mother bird picks out her own chick from thousands on the rocky beach, or a bird receives its mate into the nest but chases off other birds that approach, it seems clear that something is going on that involves representing (in my proprietary sense) in a way that tracks objects. It also seems to imply more than this:

(1) The inner system would seem to be tracking *kinds* of things (e.g., worms, food, nesting material, conspecifics, predators, perhaps even specific kinds of predators that need to be treated differently). Thus vervet monkeys give different cries for leopards, eagles and snakes.

(2) In at least some species—more or less the social animals—the inner representations seem to distinguish between different *individuals*, at least within one's own group or kind.

(3) At least in some species, there seems to be an ability to model different *states* of objects. The cat responds differently to a mouse that is animated and one that is still (and hence the mouse can sometimes escape by "playing dead"). It responds differently to another cat that is engaging in aggressive behavior from one that is sleeping. It does not regard cold catfood as on a par with room-temperature catfood, and so on.

In our own mental lives, all of these features are clearly present, and their presence is reflected in our language: we have both kind-terms and property-terms (Aristotle's substances and accidents), and we have proper names—first and foremost for our own kind, but it is an ability we can also extend to other things. It is not clear that all of these features need be found together in any organism that has any one of them. Indeed, it seems that there are plenty of species of nonsocial animals (all of the reptiles, for

instance, perhaps many birds) that divide the world into a number of kinds, and can spot members of a kind, and track the states of an object or organism, but cannot tell one individual from another.

It is now time to summarize just a bit. In order for protoconceptuality to come upon the scene, an organism must have a system of representations—that is, a system of inner states that can covary in regular ways with objects and properties of objects in their environment, in ways that can be put into the service of adaptive behavior. I should emphasize here an important caution. What I have been presenting is *not* a reductive analysis of concepts. I am pointing out some features that are present in conceptual thought in us, and which are continuous with things found in other animals. This analysis is completely silent on whether animals which have protoconcepts might also (or must also) have emotions or true intentionality or a phenomenology or language or reasoning. It is true that when you start with concepts and abstract away from things like consciousness, language, reasoning and phenomenology you are left with protoconceptuality. But it does not follow that what is left in its own right has the properties of the richer system, nor that the richer system is merely a combination of these parts *plus something more*. Idealization is not surgery and it is not undone by construction work.

Richer Features

In order to approach human-level conceptuality, we must add some additional features to protoconcepts. These additional features will not yield concepts full-stop—I do not intend to offer a compositive definition of these at all!—but will yield richer forms of protoconcepts. The first addition is *learning*. Some animals come with all or most of their protoconcepts "prewired". All of their discriminative abilities are built into the nervous system, and cannot be added-to or fine tuned through conditioning. The higher animals, however, are capable of learning about new kinds and new properties. The

social animals, additionally, can become acquainted with new individuals. I propose to look at this in terms of a protoconceptual *organ* or *module*, as it were, that functions as a **discrimination engine**. That is, such a module is in the business of responding to experience by creating, adjusting and refining a system of representations that capture environmentally-salient variables. If intelligence is fundamentally the ability to respond adaptively to one's environment, the possession of this ability to coin new categories and refine existing ones is a momentous gain in intelligence. Animals that lack it may be well-adapted to an existing environment but too inflexible to adapt to a changed environment, while animals that possess this advantage are better-suited to handling changes in their situation.

(I should emphasize that this discrimination engine is characterized in purely functional terms. Whether it is realized in the brain through a single physiological module or many such modules is a separate and empirical question.)

The functionality of this kind of discrimination engine is, of course, precisely the sort of thing that many well-known neural network systems are famously good at doing. [] For example, in a network system whose function is to "learn" to, say, distinguish rocks from mines,[] one level of the network L1 is given feature representations of sonar readings from rocks and mines. These cause initially random responses in a second level, L2, but the connections between L1 and L2 (which are mediated by a middle level in the case of PDP models, and by additional modules in the case of ART models) are given feedback in such a fashion that, after training, L2 becomes a bistable system, with one state reliably activated by reflectance patterns characteristic of rocks, and the other reliably activated by reflectance patterns characteristic of mines. [] L2 activity was originally random, but through training it has come to be an apparatus for discriminating between two environmentally-salient conditions: those in which sonar is bounced off rocks and those in which it is bounced off of mines. The training process amounts essentially to *the partitioning of a state space*. There are many possible activation states,

but they are molded through feedback into a small number of equivalence classes (in this case two).

It is quite plausible that this is how protoconcepts are formed in higher animals as well, and at least a part of the story of how humans learn concepts. The story may seem simple in its outlines, and perhaps even obvious to those who have cut their teeth on neural network models, but I shall argue later on that it has some powerful consequences for how we look at human conceptuality. And while the human nervous system is differs in a number of ways from the simple networks employed in Nettalk [], it shares with them the feature of being a highly distributed, highly connected, dynamical system. The fact that protoconceptuality is realized through a system of this sort, rather than through some other sort of system, like a rigid system of rules in an expert system, will produce certain artifacts in the form our cognition takes.

Language 1: Communication

In many animal species, heredity, conditioning and perception seem to be the entire story about the animal's ability to model its world. In a few species of social animals, however—and perhaps hive animals as well—information is also communicated from one individual to another. (I am being careful here to not use the word 'language' yet, since one might well reserve it for specific forms of communication.) Prairie dogs and vervet monkeys, for example, live in colonies. Both are moderately sized animals that are subject to predation both from the ground and from the air. In both species, members of the pack will take turns standing sentry while the others feed, and will sound a cry if a predator approaches. In the case of the vervets, at least, there are different cries for different kinds of predators (leopards, hawks and snakes), and these elicit different kinds of behavior on the part of the other members of the tribe. It appears that the entire warning is encoded in a specific call, and not composed of more atomic units expressing separate concepts as is the case in human speech. For all we know, a vervet cannot

simply *refer* to an eagle without *warning* of one's presence. However, the cry does seem to produce the effect of manipulating the other monkeys' inner models of their environment: the eagle-cry makes them get down under the canopy where they are less vulnerable and scan the skies, whereas the leopard-cry makes them take to the treetops, where leopards are too heavy to follow. This seems like far more than conditioned response: It seems very probably that they are in some fashion modeling the state of affairs of an eagle's being present, even though they have not themselves perceived one.

Here is a remarkable new behavioral resource! Lower animals can, of course, smell other animals' fear, or respond to the fact that they are suddenly bolting for cover, and this might set in motion their own flight mechanisms. However, there is no reason to think that this involves any modeling of their environment as specific as modeling the presence of a particular kind of predator. (Compare: people suddenly come rushing down the hall in a panic, and you are caught up in the panic and the flight even though you have no idea —perhaps even no hypothesis—as to what its cause might be!) In the vervet call we seem to see a mechanism for producing the duplication of a very specific kind of representation of the environment in other members of the group. True, the mechanism is not very flexible by our standards—it does not allow for any speech acts but alarms, does not involve a compositional syntax, etc.—but it is surely an advance over alerting the other members of the group to danger simply by running away oneself.

Secondary Control of the Distinction Engine

A second advance in intelligence that takes place in some of the social animals (but not the hive animals) is that individual animals do not need to learn everything from some combination of (a) closed instincts passed on through their genes, and (b) their own experience by trial and error, but can be taught by other members of the species. A kitten, for example, will naturally pounce on small moving objects, but needs to learn to give the killing blow by watching another cat. As a result, domestic cats that go feral are

often unable to feed themselves. Dogs and wolves apparently teach their young, not only about hunting and dangers in their environment, but also about the "social rules" within the pack. They could not do this, of course, unless there were mechanisms innate in the species that made this possible (turtles and even cats cannot be socialized in the same ways), but this supplies only the raw materials which are then shaped by "social" interaction. In the primates, including our own species, this ability is developed in very high degree, as we primates are very quick to pick up on what others are doing and add it to our repetoire.

In terms of our model, what is going on here is the addition of a new pathway by which the protoconceptual space of an individual animal can be shaped. The first pathway was the very slow one of evolution. The second was the faster and more efficient one of conditioning. The third and still more efficient is one (or more) mechanism by which distinctions and schemas in one member of a group can be passed on to others without the others having to go through independent learning through trial and error. Instead, one individual (the "teacher") provides supervised feedback to another (the "pupil"), which shapes not only its behavior, but its protoconceptual space. In primates, at least, this supervised instruction is supplemented by a propensity for imitation of the behavior of conspecifics. This is an enormous gain in intelligence, in that hard-won knowledge that it might take generations for an individual to come along and acquire *de novo* (and would take aeons to acquire genetically) can be retained and propogated throughout the group at much lower cost. (Protoconcepts with this additional mechanism might very well correspond to Dawkins' notion of a *meme*.)

It is worthy of note here that this sort of intelligence is not located in any one individual—indeed it *cannot* be localized in one individual!—but is socially distributed. It is by definition a process that takes place within a community of individuals where a particular kind of feedback takes place which shapes the conceptual space of one individual to approximate that of others. [Bullock]

A special and very important version of this mechanism is found in human language. A public language is something that is distributed across a community. It reflects, encodes, embodies a rich space of concepts (not merely protoconcepts, since we are talking about people here) which act as a kind of modulator of the conceptual space of individual language-users. I do not mean to take any stand at all here on the Whorfian hypothesis that humans are capable of very different fundamental ways of carving up the world. Whether the general form of our conceptual space is quite malleable or highly constrained, our learning and refinement of concepts is very much mediated through the vehicle of language. It is really quite remarkable once you start thinking about how much of what you know was learned, not through direct encounter with the world, but through what you read or heard someone say. And, more basically still, we do not in general develop first-hand our own concepts for all the things we encounter, much less those science supplies to explain the things we actually encounter, but rather we are "shaped into" the concepts used by those around us. Of course, this is not a one-way process—we may reject or refine our culture's concepts, and exceptional individuals may make lasting changes to a language and a culture. However, the point is that, for beings with language, the feedback relationship between a protoconceptual system and the environment is supplemented by a second feedback loop between the conceptual system of the individual and the set of distinctions encoded in a public language. Sometimes these two feedback systems are in step with one another. At other times they are not, and there are tensions between what seem like conceptual truths or truths of ordinary language and the feedback we get from our environments. And this keeps philosophers and scientists in business.

Language 3: Grammar and Compositionality

The cries of the vervet and the prairie dog are special-purpose and inflexible. In large measure, this is because, unlike human language, they are not articulated into a

grammar. Instead, semantics and pragmatics are all bound up together in a single type of cry. This not only necessitates a new kind of cry for each thing that is to be expressed, but also fails to capture the degrees of complexity that are very likely already there in the animal's protoconceptual system. Unlike the fly, the vervet's behavior is not simply under the control of special-purpose cells in its eye. It surely has a rich model of its environment, with variables (representatons in my technical sense) that track things like its position (in the tops of the trees, beneath the canopy, on the ground), other animals (including predators and conspecifics) and their relative positions, and schemas for its own repetoire of actions. There is every reason to think that the representational system employed has to have features that track individuals, kinds and variable properties, and that this involves a kind of articulated model that can simply substitute one element for another: i.e., common elements are used to represent a member of one's own tribe on the next branch and a member of an unknown band or a python on the next branch. In short, there is every reason to believe that the vervet's inner models are articulated in ways that its "language"—or, more accurately, its repetoire of cries—is not. The inner model, in short, is compositional and productive.

In human language, this "productivity" of the model is conferred upon the language through the addition of syntactic structure. Different grammatical categories, such as *noun*, *verb,* and *adjective*, at least roughly correspond to the different kinds of concepts we employ. Names express individual-tracking concepts, kind-terms kind-tracking concepts, verbs and adjectives track variable properties. This, of course, confers upon public language a far greater expressive ability—famously, it allows for the generation of an infinite number of sentences from finite numbers of lexical primitives and grammatical rules. But this easily obscures the fact that much or all of the *representational* power conferred here is already present in our conceptual models of the world, and even in the protoconceptual models of species that lack a language. Indeed, the mere grammatical categories of language radically underdetermine the rich semantic

relations that language expresses, because these are not encoded in the syntax, but are part of the model that is expressed through the syntax. For example, verbs and prepositions are used to express all sorts of relations, but the kinds of relations are of a much greater variety than one could divine simply by combinatorial grammatical analysis. In short, language is parasitic upon pre-existing models.

However, an articulated language does introduce a number of new elements. First, it is only here that pragmatics and semantics truly become distinct, with the ability to re-use symbols that express the same semantic value in a variety of speech acts with different pragmatic force and vice-versa. We might say that *you need a syntax to separate semantics from pragmatics.* Second, grammatical articulation is arguably necessary for language to perform the function described in the last section, in which the public language becomes a repository for the conceptual accomplishments of the group and a kind of template that shapes the conceptual space of the young. Third, as we shall see in subsequent sections, articulated language becomes the basis for the critical refinement of our thoughts, including several forms of reasoning, definition, and conceptual refinement. Fourth, there is at least the possibility that some aspects of the articulation of our *thought* as we know it as modern adult humans is itself a product of interaction with a linguistic grammar.

Reasoning

I shall distinguish between **logical** and **associative reasoning**. In logical reasoning, one passes from a thought A to a thought B because of some semantic or syntactic connection between A and B. In merely associative reasoning, there need be no such link, but merely some mechanism in mind or brain that causes one to think of B when one thinks of A. It is dubious that associative reasoning really merits the title *reasoning* at all, but I shall let that quibble pass. In speaking of reasoning, we once again need to be cautious about the problem of idealization. When we think of reasoning in our

own case, we may well be thinking of a conscious process that has a semantics and a distinctive phenomenology. Perhaps it even involves some critical oversight of the reasoning process to assure its syllogistic validity. Some of this may not be present in other animals. However, any animal capable of object-centered schemas and conditioning is likely to have something more or less continuous with associative reasoning: one representation is linked to another by dint of conditioning so that some unspecified inner mechanism causes one to be produced when the other is activated. We might not want to call this associative *reasoning* since we are in doubt of the semantics and the phenomenology of animal minds, but it is a common element of their lives and ours. And, more to the point for us, it does not require language.

Logical Reasoning 1: Material Inference

There are at least two importantly different kinds of logical reasoning and both of them depend in their own ways upon language. The first of these is what is sometimes called *material inference* and the second *formal* or *syntactic deduction*. 'Material inference' is a name we give to inferences in which we draw out the implications of the semantics of our premises. For example, knowing that John is a bachelor, I can infer that he is male and unmarried; knowing that Mars is the fourth planet, I can infer that there are at least three other planets; knowing the concept 'triangle' I can infer that it is a planar figure, etc. What we are doing here, in effect, is to tease out the information encoded in the inner models in which the concepts are embedded. The form in which the information is encoded in the model may very well be nonlinguistic. In order for the concepts 'bachelor' and 'unmarried' to be connected in the right way in an inner model, it is not necessary that there be an inner sentence that says "'bachelor' =df "unmarried man"'. Indeed, the constitutive relations between concepts are not the things that predicative language even normally expresses, and models exist in animals that do not

possess language. We must not confuse the embedding of information in a model with the sentence used to express it.

Logical Reasoning 2: Syntax and Syllogism

A second form of reasoning is syllogistic reasoning, in which conclusions are derived from premises, not on the basis of their semantics, but on the basis of their form. For example, if it is true that A, and if it is true that *If A then B*, then it must also be true that B, regardless of what propositions "A" and "B" express. This kind of reasoning is directly parasitic upon the articulation of representations into a syntactically-arranged language, and hence is a resource available only to beings that possess such a language.

Simulations

Whereas the inner representations of the fly and perhaps even the prairie dog are driven largely by perception, we are able to think about things that we are not perceiving, and even to think *creatively* about things that have never been. This requires a refinement of the conceptual system in which inner models are freed from perceptual control to be exploitable in other ways—in short, to run inner *simulations*. We do this not only in explicit fictions, but also in imagining things we want or fear, in planning, in free fancy, in philosophical thought experiments, etc. There is some reason to believe that other primates have at least limited capacities of simulation [], and it may extend to other animals as well. Goldman [] and Gordon [] have argued that simulation of other human beings' mental states lies at the basis of our ability to understand and talk about other minds.

Conceptual Refinement

In insects, protoconcepts are shaped on an evolutionary time scale by forces of selection. In lower vertebrates, they are shaped on a shorter timescale by other selective

forces in conditioning. These select somewhat blindly for whatever factors provide effective reinforcement. The protoconcepts that result—in our case, the concepts—are often fuzzy at the edges and ill-defined. When we try to apply them to the more exacting disciplines of science and philosophy, they often generate puzzles and confusion. As a result, the history of philosophy and the sciences reads almost like a history of the exploration of techniques for refining our concepts. I shall briefly describe some familiar examples in ways that highlight their role in the enterprise of conceptual clarification.

Definitions and Analysis

One important tool is that of definition and explicit analysis. If the associative links between concepts are loose and fuzzy, language gives us a tool for "fixing"—i.e., rendering more stable—more precise connections. The most heavy-handed of these is **stipulative definition**, for example of the sort that I have given for how I shall be using 'representation' in this paper. A stipulative definition anchors the meaning that a term is to be given in relation to other terms. Of course, it may be impossible to batten meaning down against all vagueness and variability, but this tool allows us to go far beyond what is accomplished by neural networks optimized for practical goals like finding berries and mates. More refined concepts can then be tested against the world empirically to see which of several possible revisions of an original concept matches up better against the world.

The flip-side of stipulative definition is **linguistic analysis**, in which one explores (and, indirectly, refines) existing meanings. This was a technique that experienced a major heyday in the middle part of this century in ordinary language philosophy. In linguistic analysis, one asks questions like "Well, would we say that a divorced man was a bachelor?" Here we are exploring what is already implicitly determinate in our use of the concept 'bachelor' by explicitly testing it against test cases that may fall somewhat far afield from the paradigm instances. This is a good technique for sharpening conceptual

edges, because the mechanisms by which neural networks create partitions of a feature space tend to emphasize central paradigms or exemplars, and not yield clear results along the borders. In this kind of testing of imaginary cases, we are, as it were, seeding conceptual space with new exemplars that are intentionally located along the boundaries in hopes of clarifying the boundaries. Sometimes it will have precisely this result, and sometimes it will have the very different result of a radical redefinition of the partitions. An important kindred enterprise is that of considering many examples that may lie some distance from the original exemplars. This can help iron out effects that are merely accidental epiphenomena of the original choice of examples. This technique is, of course, parasitic upon the linguistic articulation of a conceptual scheme. It also presupposes something more: the ability to *thematize* concepts and words, or treat them in their own right as objects of study.

Philosophical thought-experiments and Husserl's [] method of imaginative variation are closely related to this method. However, whereas linguistic analysis places special emphasis on what is encoded in the public language, these techniques tend to bracket previous encodings of ordinary language in order to explore and refine conceptual space more directly. This ability is parasitic upon the ability to run simulations and make comparisons. These two techniques–linguistic analysis and thought-experiments—complement one another, in that one is in dialogue with the slower-changing collective entity of a language, which encodes the normalized experience of many individuals, while the other explores within individual conceptual space, which is both more idiosyncratic and adaptable on a faster time scale.

Taxonomy

Another very old and useful enterprise in both philosophy and natural science is that of taxonomy. In taxonomy, we attempt to organize our concepts by grouping them in terms of similarity. In so doing, we sometimes discover what groupings seem to lend

more insight or have more objective validity than others. Aristotle, for example, in examining the physiology of animals with an eye towards taxonomic classification, came upon the important discovery that whales and dolphins are physiologically more like land mammals than like fish, in spite of their outward shape and the fact that they live in the water. This was a substantive biological discovery based upon an enterprise of grouping by shared properties.

Philosophical Implications

What has been said in the last few pages has been said very quickly and in a very sketchy way. Taken in its own right, I supspect that most readers will either find it pretty straightforward or else think it highly speculative. However, when you start looking at fundamental issues about language, thought, truth, meaning, reference and the like through this lens, you begin to realize that you have made a fundamental change in perspective from how philosophers have generally looked at these issues. In particular, discussions of concepts and thought have in large measure been assimilated to issues in the semantics of language. Issues in linguistic semantics, in turn, have been driven (within philosophical circles, at least), not so much by the questions of the psychologist, or even the descriptive linguist, as by those of the logician, interested in truth-preserving deductions. Hence, many classic discussions of "semantics" quickly turn away from messy natural languages altogether and into artificial languages that have rarefied properties like completeness or finite axiomatizability. And all of this is allright as far as it goes. The problem is that, if we look at language and concepts and semantics wholly through the logican's glasses, we may lose sight of some things that are fundamental for other purpose, and may in the end even distort our subject matter entirely.

And so, for the remainder of this article, I shall tease out some of the implications of the account of concepts I have given, highlighting how these may cut in different

directions from those generally taken in philosophy of mind and language. I shall organize this by the assumptions of the model that generate the philosophical implications: i.e., the characterization of schemas and protoconcepts, the embedding of these in models, the realization of protoconcepts in neural networks, and the relation between mental models and their linguistic expression.

Implications of Schemas and Protoconcepts

What is a concept? The analysis here gives no comprehensive answer, but suggests in part that concepts are protoconcepts as these are embedded in our particularly rich sort of experience. Protoconcepts, you will recall, are elements in object-centered schemas in which there are aspects of the model that correspond to objects and properties, whether variable features or kinds, and perhaps elements that correspond to individuals.

Concepts arise within local, special-purpose models

Protoconcepts, and hence concepts, arise as elements of inner models. In some sense the fundamental activity underlying concepts, thought and language is that of modeling things in oneself and in the world. There seems to be every reason to think that human and animal cognition are built by the accumulation and interconnection of numerous local and special-purpose models of different aspects of the world. In primitive nervous systems like that of the fly, the modules seem to be implemented directly in the hardware. Our own brains also contain a great deal of hardware modularity, such as having particular bits of tissue dedicated to things like recognizing human faces. But I wish to suggest something more here: that our minds involve an additional layer of "virtual" or "software" modularity, in the sense that when our neural nets generate a model of some feature of our environment, it functions semi-autonomously, so that changes can be made to the local model without causing

widespread global changes to everything else. In using one model we can exploit information from other models—e.g., I adjust my cooking schemas at high altitudes in light of my knowledge that both the boiling of water and our tastes change at high altitude. Here I *borrow* knowledge from one schema to modulate another, rather than making my knowledge about pressure part of my culinary knolwedge in its own right. But we start out learning many specialized things as a hodgepodge, and may easily overlook ways in which they do not go well together.

Modeling is ecologically-based and pragmatically charged

The enterprise of modeling one's environment is driven by what factors are relevant to the particular kind of organism you happen to be. Vervets have reason to model the presence of eagles overhead; lions do not. The possession of an opposible thumb and the possession of mental equipment that can model rotations and other physical transformations of solid objects are not only accidentally joined together: there is only reason to select for the latter ability if an organism's interface with the world supports the manipulation of solid objects. (We do not have a good interface for interacting with fluids, and so things like burbling streams and waterfalls seem strange and wondrous to us, and we can only find our way to an understanding of them through building great mathematical highways.) Both the sorts of things that we model and the constraints for what counts as "good enough" are conditioned by pragmatic concerns. In a biological system per se, the question about a model and a protoconcept is not that of truth or correspondence, but whether it is good enough for practical purposes. It is only later that models are adapted for things like truth and reference, and since they were designed for more basic tasks and then appropriated for new uses, we should not be surprised if some problems will arise.

Features, networks, and object-tracking functions

When we begin to look for the basis of a theory of reference in protoconcepts, we find an interesting set of familiar features that have appeared in theories of reference, often in conflict with one another. A protoconcept has the function of tracking an individual, a kind or a property in the environment. It arises within a model in which it is situated with respect to other elements of the model, and in response to particular kinds of perceptual input from other modules in the organism's nervous system. Moreover, the protoconceptual space is shaped by a distinction engine that forms partitions of a state space in response to feedback. In order for this to take place, it must be possible for the system to continue to track the same object, kind or property while adjusting both the input mappings and the partitioning of its own protoconceptual state-space. In other words, it is a constraint of this picture that protoconcepts involve:

1) an element that locates them with respect to other protoconcepts within the model (a network of protoconcepts)

2) an element that connects them with things outside the model, such as perceptual inputs and motor schemas, and

3) an ostensive element that "points" to the things that are to be tracked and continues to do so while the other elements are readjusted.

Here one sees key ideas of (1) the network theory of meaning, (2) a constitutive theory of meaning such as that based on sense-data, and (3) something that has overtones both of the Kripkean causal theory of meaning and Millikan's teleofunctional account.

Implications of Realization Through Neural Networks

The assumption that the distinction engine is realized through neural networks also has some implications. First, there are some things that this assumption gives us for free. One of these is the ability to "learn" distinctions through a feedback process. This is the sort of thing that neural nets are naturally suited to. Another is the fact that

concepts have fuzzy boundaries and are centered around paradigm examples. This feature of our concepts seemed puzzling from the standpoint of logic and formal semantics. But if we treat concepts from a biological standpoint, as things that are, among other things, protoconcepts realized through neural networks, this feature is completely natural.

Second, the embedding of protoconcepts within a network system once again reinforces the pragmatic aspect of conceptuality. The formation of a partition of a network's state-space is pragmatically-driven by the feedback that is supplied when the system is "learning". Again, both the shape of the concept-space and the degree of exactitude that counts as "good enough" is driven by the reinforcement cycle. [standards of truth internal to the model--Edidin's train example]

Implications of the Relation Between Concepts and Language

Philosophers have generally tended either to assume that concepts arise only within a linguistic system or else to project a language-like system in the mind that has all the features of adult human thought and locate concepts within this. And this may be a necessary truth when talking about concepts in the full-blooded sense. However, concepts are in part protoconcepts, and these occur in other animals without a language and arguably without many other features of our thought as well. It thus behooves us to examine some features of the relationship between concepts and language modeled here.

Protoconceptuality and models are prior to language

As we have seen, protoconceptuality can exist without a language. A language provides additional ways of exploiting conceptuality and of transmitting conceptual structures and conceptual thoughts from one organism to another, but in so doing it adapts protoconcepts to new uses as well.

The aggregative unity of language is not a unified model

One of the important things that a language does is to provide a single framework for expressing ideas that arose in diverse and local models of features of the world. This very likely provides necessary resources for working to unify our thoughts. However, it is important to see that language *directly* "unifies" models only by aggregation. Take two local models of the world that arise independently, such as aesthetics and physics. Notions like "beautiful" and "the pinnacle of Rennaisance portrait painting" arise in ways far removed from ideas like "parabolic trajectory". But I can say something like, "If you fire the Mona Lisa from a catapault, the pinnacle of rennaissance portrait painting would travel in a parabolic trajectory." And my ability to say such a thing does not count as a "unification of aesthetics with physics"! English is not parochial. You can take the concepts crafted in local models and just string them together in English. In fact, natural languages can express the statements of mutually incommensurable systems, like alternative geometries! So whatever the prospects for finding unifying connections between models, it will require something more than the kind of aggregative unity found in language. The mere fact that I can express propositions of Euclid's geometry and Reimann's in English—even in the same sentence!—does not count as having found a geometry that is both orthogonal and spherical.

Indeed, I should be inclined to put the point even more strongly. Although there are both good reasons and strong drives to unify our local models of the world, this kind of unity should not be a constraint upon language, and is not even a virtue for a language per se. For it is a virtue of natural language that it is able to express what is encoded in a great variety of local models, be they commensurable with one another or no. This confers upon us a great cognitive advance, and may even be necessary to the project of connecting, reconciling and unifying more local models. So while unified models may be valuable, it is also valuable that a language be able to accommodate disunity.

[disunity of science]

Language inherits the encoding of knowledge, distribution of linguistic labor

Insofar as a language may be said to have a semantics, this is inherited from that of conceptual models. However, a language as a public and transpersonal entity then acts as a "teacher" and regulator of the concepts of those who use it. So if I discover a new kind of animal and call it a "plippenloafer", and report it in the press, I first have a concept for this kind of animal myself, and then pass it on to the language at large. Then, someone who is without personal access to plippenloafers can learn of them through the language. And indeed, even someone who lacks the perceptual and practical schemas needed to identify a plippenloafer should he stumble upon one can in a somewhat extended sense refer to them by using the new English word 'plippenloafer'. The linguistic type takes on the function of encoding the concept, which has the function of tracking a kind of animal in this case. The layman who then uses the term intends to activate this function. Hence the division of linguistic labor.

Uses of language for reasoning are not only uses

It should perhaps be obvious that reasoning is not the only use of concepts or language. For one thing, the kinds of reasoning that tend to interest philosophers are not even present in other animals that have protoconcepts. For another, concepts and protoconcepts play other roles as well, in things like perception and action, and even language is used in plenty of kinds of speech acts that do not have truth values and cannot be stuck into a syllogism. [Austin 1962]

It should therefore be little surprise if concepts and natural language are resistive to some of the things that count as virtues in logical reasoning such as axiomatizability, consistency, completeness, and bivalence. For purposes of a truth-functional logic, we should want a language that is not vague, one which employs concepts whose boundaries

are not fuzzy, and for which every proposition is either true or false. However, these are not the kinds of concepts we have. It is therefore ironic, I think, that the study of language among philosophers has been so heavily driven by the interests of logicians and by the vision of what a language would have to look like to be good for particular kinds of logic. Why, for example, should we think of predicates in terms of their extensions? Because this is how one does it to model semantics for purposes of logic. But what if concepts were not built with logic in mind? What if God or nature did not intend semantics to serve the logician's ends, but the organism's? A more biological account of protoconceptuality seems to lead us in different directions.

There thus seems to be a tension here between the fuzziness of concepts and the requirements of the kinds of reasoning in which they are employed. Does this imply that something is wrong with our understanding of concepts or of reasoning? Not necessarily. If you assume, as the 17th century rationalists did, that God created our minds to have clear and distinct ideas of things as they really are, then perhaps there is a problem. But if you think that God or evolution shaped protoconcepts with pragmatic goals in mind, so that the resulting minds could approximate clarity and accuracy according to their needs and their measure, and that reasoning is built upon this engine, with all the particular facts about its implementation, then it may not be so surprising. Why should one think that either God or nature would build us in the way the rationalists assumed?

Semantics

Let us now try to look at the semantics of concepts from this new standpoint. The very word 'semantics' is arguably *homonymous* as applied (1) to concepts and (2) to symbols in a language, as the "meaning" of symbol consists in the ways it can be used to express concepts. (For a more exhaustive analysis of the semantics of symbols, see

Chapter 4 of Horst 1996.) But this difference need not concern us here. It is quite enough of a task to try to come to an understanding of conceptual semantics.

The first thing that becomes apparant is that the natural way to approach conceptual semantics is *not* by way of looking at extensions, but by way of the role they play in the overall mental and ecological economy of the concept-user in its environment. This is not to say that there is no role for the kind of semantic analysis that has been favored by logicians, but rather to recognize that that kind of analysis is an analysis with very special purposes in mind, which are neither the only purposes nor the most fundamental purposes of conceptuality in its protoconceptual roots. As a result, we do better to look first at the biological analysis of concepts and then see what follows for an extensional analysis rather than the other way around.

In the very sketchy analysis I have given above, I have suggested that there are at least three different kinds of concepts: (1) those whose function is to track individuals, (2) those whose purpose is to track kinds, and (3) those whose function is to track variable properties. I might add that there are surely other sorts as well:

- special-purpose concepts whose role is to pick things out by way of special relations—the indexicals I, you, this, that, here, there, now, then, etc.
- concepts that play a role in connecting other concepts, such as logical connectives
- concepts (or cousins to them) which play a role in differentiating the logical, deontic and pragmatic modalities of thoughts.

However, for present purposes, we may confine ourselves to concepts whose role is to track individuals and kinds.

Kinds and Properties

The ability to track kinds is a straightforward consequence of the right kind of network, supplied with appropriate connections to the environment and conditioning

cycles. Many kinds of neural nets will partition an input state space, either with the help of supervised feedback [Sejnowski, Rumelhart and McClelland] or driven by internal properties of the net [Grossberg]. Neural nets are naturally well suited to learning feature discriminations through trial and error [], in a fashion that approximates the kind of prototype-driven learning found in humans [Rosch; …] As others have noted [], if you assume that human conceptuality is realized through neural networks, you get a microexplanation of basic types of learning for free.

I should remind the reader here that I am *not* offering a reductive or compositive analysis of concepts or concept learning. The fact that a system such as a neural network can adaptively track salient environmental features (the salience being provided by the feedback of "reward" and "punishment" that drives the learning) is not a sufficient condition for network states or network dispositions to count as concepts, or indeed as anything mental at all. Nor does it even provide sufficient conditions for protoconceptuality: adaptive feature-tracking only amounts to protoconceptuality only when it is in the service of the adaptation of an organism. It is not clear that it becomes true conceptuality unless it is embedded in a mind like our own, which also has language, reasoning, self-awareness and consciousness. However, our true conceptuality is realized, in part, through neural networks, and the features of these networks have implications for the "shape" of our conceptual space.

Logicians tend to treat kinds and properties in exactly the same way, representing both by means of predicate letters. This is a useful simplification for formal purposes, and seems to have no serious consequences either in the extensional interpretation of the predicate calculus or in the preservation of truth in inferences. It is not, however, a good model of how humans think. Intuitively, we have difficulty teaching students to make the transition to the predicate calculus, because both our language and our psychology treat kind terms and property terms quite differently. Linguistically, we seldom employ anything corresponding to a bound variable, but use a noun both to denote an object and

to designate its kind. Millikan [] and others [] have argued that there is good developmental evidence to show that we think in this way too, and from a very early age. In short, human psychology is highly biased to model the world in "Aristotelian" terms—kinds and accidental properties (properties a thing may have or lack without ceasing to be a member of that kind)—rather than predicates and variables or even predicates and purely denotative names. The latter are applied first and foremost to a very limited class of objects—members of one's own species—and arguably there is nothing very much like this kind of representational resource in any but a few social species of mammals and birds.

["more mama"]

Individual-Concepts

I suggested above that the social animals seem to possess, and to need to possess if they are to be social animals, resources for distinguishing and re-identifying specific members of their own kind. Be this true or no, we humans, at least, can identify both individual human beings, and individuals of most medium-sized concrete objects. (This ability may break down at the level of subatomic particles, but that is no matter for this analysis.) Our concept for an individual would seem to involve at least three strands:

1) a strand that somehow encodes the function of picking out a single individual through various changes that both it and our ideas about it might otherwise undergo (perhaps the "pure x" that Husserl said lies at the core of the noema [])
2) a set of perceptual and motor schemas to be employed in re-identification of the individual, and
3) a network of connections to other concepts and conceptually-articulated beliefs and memories.

I hasten to say that these "strands" need not be interpreted as independent *things* or *representations*, and may not be independent *in vivo*, much less separable in fact. A change in our beliefs about a person—say, that she has grown up from the little girl we last saw twenty years ago, or has covered her body with tatoos or taken to wearing a Muslim veil—will change the set of perceptual schemas we use to re-identify her. Likewise, perceiving a change in what is presented for our perception and interaction may occasion changes in belief as well: we hear the familiar voice coming from behind the veil and come to believe that she has become a devout Shiite. And indeed, sufficient changes in perceptual schema or beliefs may cause our whole assumption that there is a single person we have been dealing with to crumble: we decide that this is an imposter and not the real Anastasia, or that we had been confusing two sisters all along. (Manifestation of multiple personalities may cause even deeper dissonances in our cognitive structure.)

When one thinks of an individual, these different strands may play roles of differing strengths. Consider these examples.

Things that go bump in the night

I am not a happy camper, in the quite literal sense that I am far too prone to fretting when something rustles in the bushes in the middle of the night. When this happens, I form a thought that might best be expressed by some sentence like "There's something out there." Of course, the logical analysis of this sentence— x(x is out there)—does not capture my thought. Rather, I hear particular kinds of rustling, and constitute the situation as one in which there is some particular thing—call it X—which is moving around and causing the rustling noises. I do not *call* it X, of course. I do not call it anything at all. But it is as though representational resources are being set aside for an individual animate thing, and the closest thing we have in language to express this is a variable letter. So I form this model of my physical environment—myself in the dubious safety of the tent, trees and bushes all about, the flashlight ever ready to my left

hand—and some thing, X, located at some inexact location *over there*, making noise in the night. I then think thoughts like, "What is it?" By which I mean an indefinite but nonetheless circumscribed set of possibilities: a skunk? a bear? a serial killer? another camper who is, as they say, "making like a bear"? In short, I am trying to pin down a kind-concept (and perhaps something much more concrete if the kind is human) to attach to X.

I should point out in particular that X is *not* felicitously expressed by "the Y such that Y caused the bushes to rustle." For if I find out that it was just the wind in the bushes, I don't conclude that "it"—X—is identical with the wind. Rather, I conclude that I was mistaken, and that there *was* no X of the sort I had postulated. I thought I heard some *thing*, but I did not. I might *say* "It was just the wind", but 'it' is not an anaphoric reference to the hypothetical X. There might be occasions on which I would be inclined to constitute the wind as a thing, but it does not fall within the boundary conditions (perhaps something like "medium-sized animate things") that were tacitly supposed in thinking about the hypothetical X.

What are we to say about the semantics of such a thought? There are really two very different sorts of questions here. The first is a question about the ways in which the concept goes about the task of triangulating an object in the world. The second is a question about what individual, if any, it picks out in fact. The second is dependent upon the first. In this particular example, the triangulation works something like this:

- Percpetual schemas are set off that are compatible with the interpretation that there is a moving object causing rustling noises
- I coin a representational resource 'X' for referring to the hypothetical cause of that noise
- The interpretation activated by the perceptual schema places particular conceptual constraints upon X—e.g., that it be animate, and of a suitable size to cause such sounds. (The exactitude of these constraints will be determined by

how finely I guage my objects-to-sounds-they-make mapping!) This does not determine a kind-concept, but sets constraints upon compatible kinds. (Too big for a gnat, too small for Godzilla, too fast for a turtle, etc.)

These elements change dynamically as new information comes in—a sudden crash and the size imputed to X moves up in our estimation, we hear species-specific sounds or a familiar voice and our interpretation gets locked onto an existing kind- or individual-concept.

What is in fact picked out, if anything, will depend upon whether there is indeed something corresponding to these conditions. If there is no animate object that caused the sounds, 'X' fails to pick anything out. If a single skunk is making the sounds, it picks out the skunk. If it was a family of skunks, Mama and her babies, then we were erroneous in thinking it was a single something. But it is also easy to think of grey areas here, where it is not clear what we should say: some of the sounds were made by the skunk and some by the wind. If 99% were made by the skunk, we would say we heard the skunk. If there was a skunk there, but the sounds we heard were made by the wind, then we did not track the skunk. Now suppose it is 98%, 97%, and so on..... There is no definite point at which we would say that veridical concept-formation ends and mistakes begin. Which is as it should be if protoconceptuality was devised to help us negotiate our environments rather than produce a neat semantic theory.

Here is a way *not* to think about this situation: that 'X' means something like "the Y such that Y is (an animate object and) the cause of sounds s1, s2....sn." If *this* was what 'X' meant, then 'X' would only refer if there was a unique Y that caused all of the sounds in question. But this is not a psychologically realistic story. This coining of an individual-tracking concept *is* occasioned by activation of a percpetual schema, but (a) we do not individually represent and quantify the sounds in question, and (b) we do not define 'X' in terms of a rule, even if the overall concept's conditions of reference may be expressed by a rule. (Which remains to be seen.) What happens is a series of events:

1) Events in the woods (perhaps involving a skunk) produce noises

2) Patterns in the noise activate a perceptual schema

3) This evokes the formation of an individual-concept and certain constraints

The steps from 1 to 2 and from 2 to 3 are all *causal* processes, albeit very likely causal processes involving feedback loops. There is no reason to think of them as processes of explicit reasoning or as propositional. That is, we do not need to think of this process as proceeding as follows:

a) I hear sounds s1, s2, ...

b) These sounds are compatible with the interpretation of being caused by a moving animal

c) I hypothesize that there is an animal X that caused these sounds.

Granted that there are important parallels, even continuities, between formation of the thought that there is something out there and this kind of explicit reasoning, it is nonetheless the case that causal processes in cognition might *be reason-shaped* without being examples of reasoning proper. When we force a causal process into the mold of a reasoning process, we are left with certain unwanted *artifacts*—we have to treat seriously the notational baggage of our logical interpretations, such as representations of sounds.

Philosophical Payoffs -- Question and Answer

I suggested at the outset that there would be substantial philosophical payoffs for this new theory of concepts. In particular, I suggested that a number of familiar philosophical puzzles might cease to be puzzling if we work from these assumptions about human cognition. There is not space here for a detailed discussion of these, but I shall attempt to address a number of them briefly in the form of question and answer.

Are concepts functionally defined, or even functionally describable?

One must differentiate between the mathematical and the teleological notion of 'function' here. It is my position that the original purpose of protoconcepts is teleofunctional in nature: namely, to track salient environmental properties. In human conceptuality, concepts may take on additional functions as well.

However, it is absolutely essential to the process of concept-learning that concepts *not* be functionally defined in the mathematical sense of 'function' usually associated with functionalism. Consider two issues: (1) the shaping of a concept through experience and learning trials, and (2) the ability to integrate a concept with additional sensory modalities and motor schemas. Since it is the original function of concepts to track salient environmental properties, and the realizing system is capable of adaptive learning, the concept must be capable of tracking the *same* property over time, and doing so ever more closely through learning. But through this process the input/output mappings change, and hence the math-functional description. Likewise, it is possible to learn to identify objects and properties through various sensory modalities—one can detect a rose by seeing one or by smelling one, for example. When I have seen roses in the yard, and then learn what one smells like, my concept of [rose] is enriched by taking on a new dimension, but it is not replaced by a completely new concept. (Consider, by contrast, if I were to learn that the things I saw in the garden were in fact tendrils of some horrible sci-fi creature.)

The right way to look at this, I think, is to see concepts as having multiple constraints, as previously suggested: I "pin down" the concept to a given property by way

of an ostensive element—I am trying to get clearer on the nature of *that sort of thing*—and then allow the criteria connected with perceptual schemas or the semantic connections to other concepts to vary. Arguably, it is only this sort of multivariate analysis of concepts that allows us to tell a coherent story about how one can have a single concept over time or across individuals.

Why are concepts fuzzy?

There are several factors contributing to this. First, the dynamics of some neural networks is such that their partition of a state space behaves more like a bistable system for states close to the paradigms on which they were trained than on states equidistant from paradigms. The paradigms act as attractors, in the technical sense employed in nonlinear dynamics, and the borders of state space lie in the orbits of more than one attractor. Second, in an organism that needs to learn to distinguish salient phenomena, it is not good that its conceptual system be *too* stable. The disonance provided by cases that occur in borderline cases provide internal impetus towards further refinements of conceptual space. Third, the degree of granularity and exactitude of a neural network is in part conditioned by the nature of its learning trials, and there is often a trade-off between *speed* and *precision*. [creation myth]

Why is there a tension between the nature of our concepts and logic?

Concepts were not designed to accommodate logic. They were designed to help us and our ancestors get by in our (changing) environments. The ability to think logically was added on very late in the picture. If one were starting out to design an animal precisely to be a logical reasoner, one would have made it with different design

specifications. One might, perhaps, have made it out of digital circuits rather than neural nets. However, then it would not be optimized for learning. We are discrimination- and learning engines that, against all odds, are capable of being shaped into logical reasoners.

Why are there failures of bivalence and why is this so troubling?

This is an artifact of the mismatch of aims of the conceptual system and the reasoning system. The reasoning system aims at the careful completion of valid inference. It operates as though the conceptual system were well-defined rather than fuzzy. However, in fact the conceptual system was designed for other ends, and concepts are more oftenthan not fuzzy, with the result that soemtimes there is no fact of the matter whether a given thing is an X or not. This is troubling because it indicates that we are not built with a consistent design. From the standpoint of logical reasoning, we have been equipped with the wrong conceptual parts.

Why do people think in terms of contrast pairs (and larger contrast sets) rather than in terms of the logical partitions of P/not-P?

Neural nets partition a state space into two or more sectors based on paradigms and feedback. The result is a set of distinctions centered around positive paradigms—mines and rocks, or animal/vegetable/mineral—rather than predicates and their complements. A partition of a state space into two segments is equivalent to a logical partition into P's and non-P's in terms of its extension, but not in terms of its lineage or function . Neither does it involve the *syntactic* aspect of a representation cojoined with a negation indicator. A partition into more than two sectors can be analyzed externally (or by another system within the organism) into any sector and its complement, but this is not a natural description of its dynamics as a neural network.

Why does our assessment of truth values sometimes depend upon context, as in the case of the train and relative motion?

Both concepts and judgements operate within the context of local models, and their standards of exactitude and background assumptions are specific to local models. The commonsensical models we use for things like navigation on the surface of the Earth treat the earth as stationary. This is a natural feature for the model given that it is primarily employed for practical purposes. It is also natural and desirable that we retain these practical models in spite of knowledge such as that there is in fact only relative motion. This sort of scientific model is far more cumbersome for calculations, because the more commonsensical model is at once simpler and very likely efficiently encoded into animal brains in the form of a special-purpose module. In practical terms, it would be disastrous if one actually "unlearned" the idea of a stationary earth. Moreover, the question of truth only arises once one has specified concepts and a local model. Given the relatively loose standards of accuracy called for in a model used for getting around on the face of the Earth, it *is* true that the speaker remained where he was—i.e., (a) within this model, rest is a possibility, (b) the speaker met the standards of rest within the model, and (c) the specifications of the case are not sufficient to force abandoning or refining the model.

The deep mistake comes in thinking that, when we learn something a new and fundamental as the impossibility of absolute motion, that it can or should permeate all of our inner models. Changing all of our inner models is *not* just like refining a few definitions in a formal system, because our mental life is not generated from a kernel of a few axioms, but rather is cobbled together from hundreds or thousands of more local

models that are crafted for pragmatic purposes. The purposes of science—of modeling the world as accurately as possible and allowing very particular sorts of interventions in it—are only one sets of purposes among many others. It would likely be disastrous for an organism like us to in fact abandon more rough-and-ready models, even if we could. But very likely we cannot—many of our commonsense models are literally *incommensurable* with scientific models, because they are optimized for different purposes and have different representational resources.

Perhaps worse for some familiar views of science is the possibility that some of our scientific models are incommensurable with one another in the same ways, and for the same reasons. Local scientific theories are developed with specific empirical questions and specific metatheoretical constraints in mind. They employ particular representational resources, and are more directly driven by the pragmatic reinforcement of experimental and explanatory success or failure than by integration with other local representations. The fact that a representational system is successful in modeling one aspect of the environment by no means assures that its formal structure will be easily integrated with other successful representational systems. (E.g., wave and particle theories of matter and energy.) It is an empirical question, not just about the world but about human psychology, whether any two successful theories can be "unified" in the sense of being accommodated in a single framework that serves as a "common denominator" for them both. (It is a much simpler matter, of course, to "accommodate" them both in a natural language: natural language will let in whole concpetual frameworks that are incommensurable with one another. Natural language, in short, is promiscuous.)

As a result, the disunities of science are a special case of mental modularity—in particular, of the "software" modularity of learned models. The *actual* disunity of science is surely an artifact of this feature of our minds. And whether an eventual unification of science is at all possible is likewise dependent upon the answers to empirical questions about our minds and their interface with the world.