10/1/04

The Structure of Thought

John L. Pollock
Department of Philosophy
University of Arizona
Tucson, Arizona 85721
pollock@arizona.edu
http://www.u.arizona.edu/~pollock

1. The Fregean Theory of Thought

According to what might loosely be called "the Fregean theory of thought"[1], when we think, we entertain propositions. These are abstract objects that possess truth values. The proposition entertained is the content of the thought. Furthermore, when we use language to communicate our thoughts to others, what we convey is propositions. So propositions play a dual role in providing the contents of thoughts and the meanings of linguistic utterances. They are both "objects of belief" and "products of assertion".

Language and thought both have structure. For instance, when I have the thought that the apple is red, the apple is "represented" in my thought and so is the property of being red. According to the Fregean theory, the proposition that I entertain has as a constituent a concept of the property of being red. So concepts play a role in describing the structure of our thoughts. Concepts were are supposed to play a role in language, providing the meanings of the predicates used in talking about properties. The Fregean theory of thought was traditionally coupled with what can be called "the classical theory of concepts", according to which all but some "logically simple" concepts have definitions that provide logically necessary and sufficient conditions for something to exemplify the concept.

The combination of the Fregean theory of thought and the classical theory of concepts provided an elegant framework within which analytic philosophy functioned throughout most of the twentieth century. However, during the last third of the century, various parts of this framework were called into question. The failure of ordinary language philosophy to discover definitions for philosophically interesting concepts led philosophers to search for an alternative to the classical theory of concepts, and the work of Kripke, Putnam, Donnellan, Kaplan, and others led to the conviction that the relationship between language and thought is more complex than the simple one envisioned by the Fregean theory. The general tendency in philosophy has been to retain at least the bold outline of the Fregean picture of the structure of thought, but give a more complex account both of concepts and of the way in which language is related to thought.

This seemingly sensible strategy requires us to investigate thought first, and then use the results to investigate language. To do that, we must consider what constraints a

[1] Frege's views were subtle and complex, and I do not claim that this crude characterization does justice to them.

theory of thought must satisfy. If we set language aside, the remaining role of thought is in rational cognition, both epistemic and practical. Thought is the process of deciding what to do or what to believe, and thoughts are mental objects manipulated in such cognition. The purpose of this paper is to investigate what might be called "the logical structure of thought" and its role in determining how thoughts function in cognition. When I began work on this paper I assumed I would be producing a theory of propositions and concepts, but I have been led instead to the surprising conclusion that although we can usefully distinguish syntactical and semantical aspects of thought, focusing on the cognition of a single agent provides no role for anything like propositions and concepts as traditionally construed. Thoughts have syntactic structure, and the constituents of thoughts have semantical properties that govern their use in cognition, but these semantical properties cannot be described in terms of proposition and concepts. Only when we return to the relationship between thought and language can something like the traditional notion of concepts be resurrected, and even there the notion we arrive at is importantly different from traditional philosophical views of concepts.

2. Thoughts

In cognition we entertain and manipulate thoughts. Thoughts are mental objects that come into existence as cognizers entertain them. Once they are entertained they are stored in working memory for awhile, and then go out of existence if the cognizer does not retrieve them from working memory and explicitly entertain them again. We know that we have thoughts because we introspect the having of them. But what are these mental objects that we call "thoughts"? This is one of the central problems of the philosophy of mind, but it is not a problem we have to solve now. For present purposes we can just take it for granted that we have thoughts and they are manipulated in cognition. However, I have recently presented a theory of mental objects like thoughts ("What am I? Mental states and virtual machines."), and it provides a useful perspective from which to think about some of the issues that arise in this paper, so let me sketch it briefly. What I argued is that mental objects like thoughts are datastructures, and as such they are virtual objects of our cognitive virtual machine. The word processor running on your computer is an example of a virtual machine. Virtual machines manipulate "virtual objects" like files, windows, and so forth. I gave an account of virtual machines and virtual objects, but the details are not important here. What is of more significance for present concerns is my claim that cognition can be viewed as the operation of a virtual machine implemented on our underlying neurophysiological architecture, and mental objects are datastructures (virtual objects) manipulated by our cognitive virtual machine. This view of thoughts is not crucial to the general claims of the present paper, but thinking of thoughts in this way will make some aspects of their role in cognition less mysterious than it might otherwise seem, so I will occasionally appeal to this theory as we proceed.

As mental objects, thoughts are individual things. In this respect they are like pains, tickles, percepts, and so on. Individual things can be classified in various ways, yielding type/token distinctions. There are many ways of classifying things, so the same object can be a token of many different types. If we turn specifically to thoughts, one way they can be classified is in terms of their very general role in cognition. Some thoughts are (occurrent) beliefs. But others are hopes, fears, desires, etc. These are said to represent different "propositional attitudes". What distinguishes thoughts from other mental objects is that they can be regarded as having (or at least purporting to have)

truth values. Thus a hope is a hope that something is true, a desire is a desire that something be true, and so forth. This is one uncontroversial way of classifying thoughts. Let us turn to some of the other ways thoughts can be classified.

2.1 Syntactical Typing

One way thoughts can be classified is "syntactically". The crucial observation is that thinking is "productive", in the same sense language is. We build complex thoughts out of simpler thoughts, using various compositional devices (connectives, mental adverbial modifiers (big mouse), etc.). By virtue of what does a thought have its syntactic structure? This can seem puzzling if we suppose that thoughts must be ordinary physical objects and their syntactic structure must derive from their low-level physical properties (as it does for sentences of a public language). But this is one place in which it is useful to think of thoughts as datastructures (virtual objects in our cognitive virtual machine). Viewed as datastructures, thoughts can have syntactic structure simply by referencing other datastructures. This is analogous to the observation that an object produced by a computer program can have a structure without that structure being directly reflected in its physical implementation. For instance, if we are programming in LISP and we create a list (a b) of two atoms a and b, the list has a structure that involves the atoms, but this cannot be discovered by directly inspecting the memory locations at which the list and the atoms are stored. The relationship is instead a functional one that depends on the fact that we are running LISP. That has the effect of making the list a datastructure that references the atoms (which are themselves simpler datastructures). In the same way, thoughts get their syntactic structure by being datastructures manipulated by our cognitive virtual machine, not by the low level properties of their physical implementations.

So we can think of thoughts as having syntactic structure and constituents. Sometimes the constituents of a thought are other (complete) thoughts. For example, a thought might be a conjunction of two simpler thoughts. But thoughts can also have constituents that are not themselves thoughts. For example, some thoughts can be viewed as attributing properties to objects. For example, I may have the thought that an apple I see is red. Such thoughts have constituents that represent the properties and objects, and it is by virtue of having those constituents that they attribute the property to the object. As I am using the term, thoughts at least purport to have truth values, so the constituent representing the apple and the constituent representing the property of being red are not thoughts. On the analogy of language, we might call these "subsentential constituents". Some constituents of a thought may themselves have syntactic structure and constituents. This is true even of subsentential constituents like definite descriptions. But eventually we reach constituents that do not. These atomic constituents of thought are like "mental words" (tokens, not types).

Even though they have no structure, we can classify two atomic constituents as being "the same or different" in a purely syntactical or lexical sense, analogous to recognizing two tokens of a word of English as being instances of the same lexical type. I will refer to this as lexical sameness, and say that the constituents are of the same lexical type. Lexical sameness can seem mysterious. How do we detect that two mental items are of the same lexical type? Not by somehow holding them before our minds and comparing them piece by piece. They do not have introspectible pieces. In fact, we do not do it by doing anything — we just do it. The ability to make such comparisons is built into our cognitive architecture. We do it by performing a primitive operation (a "basic act") analogous to our ability to move our effectors. Agency has to start somewhere, with some acts that can be performed "directly", and that is no less true of

mental agency. There is nothing difficult about building a cognitive system that can do this. For example, an implementation of LISP includes a primitive operation for comparing LISP atoms to see whether they are the same. This is just built into the architecture. In exactly the same way, the ability to compare mental items lexically is built into our cognitive architecture. The mental items are datastructures in a virtual machine, and the virtual machine includes operations for comparing them. Of course, at a lower level this may be accomplished by performing all sorts of finer-grained operations, but those are not things we do, they are things our low level neurological machinery does.

Given the notion of lexical sameness, we can define syntactical sameness by saying that two thoughts or constituents of thoughts are syntactically the same iff either (1) they are atomic constituents and they are lexically the same, or (2) that are not atomic, they have the same syntactic structure, and their corresponding atomic constituents are lexically the same. So lexical sameness is a special case of syntactic sameness. I will say that two thoughts or thought constituents are of the same syntactic type iff they are syntactically the same.

2.2 Grammatical Typing

In language, to be the same lexical type is to be the same word. Words of different lexical types can be of the same grammatical type, e.g., predicate, adverb, etc. It is similarly useful to say that different constituents of thoughts are of the same or different grammatical types. Some are representations of objects, others are representations of properties, and so on. Cognitive processing is sensitive to these grammatical types and not just to the lexical types. Object representations are used in some ways, property representations in other ways, logical operators in still a third way, and there may be other grammatical types that enter into the general description of cognition. The grammatical type of a constituent of thought is stored in the constituent (a datastructure) and accessed by cognition.

It is also useful to talk about the grammatical type of a thought. Let us take the grammatical type of a thought to be determined jointly by its syntactic structure and the grammatical types of its constituents. Interchanging constituents of the same grammatical type produces thoughts of the same grammatical type.

My terminology here differs from that of the linguist, who would probably use "syntactic types" as I use "grammatical types". I have chosen to use my present terminology to keep some distinctions clear that are obscured by the standard linguistic terminology. In talking about language, the linguist starts with sentence tokens and then talks about them being "tokens of the same sentence". This is to use "same sentence" to classify distinct tokens. So "sentence" here really means "sentence type". The trouble is, there are many types of sentence types. We can classify sentences in terms of what language they are sentences of, or in terms of whether they are interrogatory, declarative, etc., or in lots of other ways. The expression "same sentence" is being used to classify sentence tokens in a particular way that ignores "extrinsic features" like type face, but includes all those aspects of the "look" of the sentence token that are relevant to its linguistic behavior. It is this classification that is analogous to what I am calling "syntactical sameness" for thoughts. This is to appeal to syntax in the sense of philosophical logic. Linguists use "syntactical sameness" to refer to what I am calling "grammatical sameness". In order to follow their lead more closely I would have to talk about two thought tokens "being the same thought", thus using "thought" to refer to the type. That seems unnecessarily confusing, so I have adopted this alternative terminology.

The ability to determine the grammatical types of mental objects is built into us. And so is the ability to determine the syntactic structure of a thought. Thoughts and their constituents are mental objects, and hence datastructures in a virtual machine. Syntactic structure and lexical and grammatical type simply consist of the datastructures containing the appropriate data, where that data is accessible in particular ways to cognition.

2.3 Propositions and Semantical Typing

Thoughts can also be classified semantically, and it will turn out that there is more than one way to do that. The standard way of doing it is in terms of the propositions "expressed" by the thoughts. On the Fregean theory, propositions were taken to be abstract entities that are the "objects of belief". That is, they are "what are believed". Two propositions are the same iff to believe them is to "believe the same thing" or to "have the same belief". Hypostatizing propositions in this way strikes me as at least mildly objectionable. Types are automatically abstract entities, and we need them to talk about semantical classification. Thus propositions and (one kind of) semantical types go hand in hand. According to the Fregean terminology, two thoughts "have the same proposition as their objects" iff they are of the same semantical type. But it is somewhat more parsimonious to simply identify propositions with the semantical types. That will be my course here. I am not sure this makes any significant philosophical difference to anything. We can still allow ourselves to use familiar philosophical terminology, talking about "believing a proposition" (doxastically endorsing a thought of that semantical type), about a thought "expressing a proposition" (the thought being of that semantical type), and so on. Note that, allowing ourselves to talk in this way, propositions can be entertained without being believed. They can instead be feared-true, hoped-true, desired-true, or simply contemplated ("What if ...?").

Propositions have been supposed to play a role both in understanding thought and in understanding language. It will be useful to have notation that clearly distinguishes whether we are talking about sentences of public language, thoughts, or propositions. For language, I will use ordinary quotation marks, writing things like "The cat is on the mat" to refer to a sentence of public language. I will use corner-quotes to refer to thoughts typed syntactically. So I may refer to a syntactical thought type by writing ⌜The cat is on the mat⌝. Similarly, I will use French quotes for propositions, writing «The cat is on the mat». Although I use the English sentence in each notation, that is just a convenience. There is no implication that thoughts and propositions are related to public language in any particular way. That is an issue we will explore below.

Because it will turn out that there is more than one way to classify thoughts semantically, propositions constitute just one type of semantical classification — what I will call "propositional typing". What is it for two thoughts to be of the same propositional type, and what is that notion good for? Propositional typing is supposed to track the notion of believing the same thing or having the same belief. When thoughts are the same in this sense, I will say that they are instances of the same proposition. There are at least two necessary conditions for believing the same thing. First, a proposition has a fixed truth value (at a possible world). If a cognizer has the same belief twice, what is believed must have the same truth value. We might take a page from philosophical logic and define the truth condition of a thought to be a function from possible worlds to truth values. The requirement is then be that for two thoughts to be instances of the same proposition, they must have the same truth conditions. A second necessary condition for propositional sameness is syntactical

sameness. This is necessary because if two thoughts can be distinguished syntactically, then a cognizer can think one without thinking the other. Thus if he believes both he has two different beliefs — he could give up one without giving up the other. Hence propositions are very narrowly specified thought types. Logically equivalent propositions are true under the same circumstances, but they are distinct as long as it is possible to believe the one without believing the other.

It will be important to realize that although syntactical sameness is a necessary condition for propositional sameness, it is not by itself sufficient for propositional sameness. This is because there can be thoughts entertained at different times that are syntactically identical but have different truth values. This occurs for thoughts containing indexical designators like ⌜here⌝ or ⌜now⌝. These designators are atomic constituents of thoughts, and I will say more about them below. But for present purposes it suffices to note that if I have thoughts of the syntactic type ⌜It is now 3 PM⌝ on two different occasions, one can be true and the other false. Furthermore, if I believe one and not the other, I have not "changed my mind". These are two different beliefs. Thus the thoughts are propositionally distinct even though they are syntactically the same. It can be argued that the same phenomenon occurs for other reasons as well. For instance, I have argued in numerous places (references) that percepts are representations of objects and can play that role in thoughts (see below). However, the same percept (individuated in terms of the visual appearance of the perceived object) can represent different (visually identical) objects on different occasions, and hence the thoughts must be propositionally distinct. This has the somewhat surprising consequence that propositional typing is even more fine grained than syntactical typing.

2.4 What are Propositions Good For?

According to the Fregean theory of thought, the whole point of thoughts is to express propositions. But it is important to step back and consider this supposition more carefully. Why should we want to classify thoughts in terms of their propositional types? Does that kind of classification play any role in cognition? I am going to suggest the somewhat surprising conclusion that it does not. In defense of this conclusion, consider what roles it might play. Under what circumstances might cognition require two thoughts to be of the same propositional type?

Epistemic Reasoning

The most obvious suggestion is that propositional sameness is required for epistemic reasoning. For instance, if I have the thoughts P and (Q ∅ R), I can infer R from them by modus ponens when and only when P and Q are thoughts of the same propositional type. At first, this seems to be a place where propositional typing is important. However, for reasoning the thoughts must occur at the same time, being held in working memory simultaneously. We saw that syntactical sameness is a necessary condition for propositional sameness, but not in general a sufficient condition. The reason it is not sufficient is that thoughts can contain indexical items that represent different things at different times. However, if two thoughts contain indexical constituents that are lexically the same, like ⌜now⌝ or ⌜here⌝, it seems that the different indexicals must represent the same thing at any given time. If I simultaneously think (hold in working memory) ⌜It is now 3 PM⌝ and ⌜I am now sitting at my desk⌝, both occurrences of ⌜now⌝ represent the same time. Thus although syntactic sameness does not guarantee propositional sameness in general, the syntactical sameness of two thoughts held at the same time does seem to be both a necessary and a sufficient

condition for propositional sameness. If this is right, we need not appeal to propositional sameness to validate reasoning. To infer R from P and (Q ∅ R) by modus ponens, it is necessary and sufficient that P and Q be syntactically the same and entertained at the same time.

Practical Reasoning

Another possibility is that propositional sameness might be required for practical reasoning.[2] For instance, when we reason that we should adopt some course of action in order to achieve a particular desire, the desire is a thought — it is the desire that something P be true. We then employ beliefs to the effect if we do such and such, P is apt to be true, and so acquire a defeasible reason for performing the action envisioned. But for this to work, the P in the belief must match up with the P in the desire, in the sense that they must have the same propositional content.

However, reflection indicates that this appeal to propositions can be replaced by an appeal to syntactical sameness in exactly the same way that was done for epistemic reasoning. So again, there is no need for appeal to propositions.

Memory

If propositional sameness is to play a role in cognition that cannot be played equally well by syntactical sameness, that role must concern thoughts entertained at different times. In order for an earlier thought to play a role in current cognition, it must be stored in memory and retrieved in the form of a current thought. Consider how memory storage and retrieval works. A simplistic view of memory would have it that thoughts (individual mental objects) are simply put away somewhere and taken out again later for further use. We might call this the "bucket theory of memory". On this view, when we remember correctly we have literally the same thought (a mental particular) as we had earlier. I doubt that memory ever works this way. Memory stores information, but probably not in the form of discrete thoughts. Even in the most straightforward cases, it seems that the thought we get as a result of remembering is a different thought (individual mental object) than the one we had originally. This is true even if they have the same propositional content. To defend this claim convincingly, we would need an account of how memory storage works. Various partial accounts have been proposed, but there is no generally accepted theory. Most theories of memory do, however, have the feature that information is stored in a more compact form than simply as a bucket or list of thoughts. Memory retrieval "reconstructs" the original thought, which is just to say that it constructs a new thought appropriately like the original thought.

If we suppose that memory produces new thoughts rather than retrieving the same mental particular that gave rise to the stored information, this might provide a role for propositional sameness in cognition. Perhaps for the new thought to be "appropriately related" to the original thought is for it to be propositionally identical. The received view in epistemology has been that you can only (correctly) remember something if you previously knew it. So it seems that the propositional content of thoughts produced by memory retrieval should at least purport to be the same as the propositional content of the thoughts giving rise to memory storage. Discovering that you did not previously have a thought with the same propositional content as what you now seem to remember would be a reason for rejecting your purported memory.

---

[2] I owe this suggestion to Justin Fisher.

Thus propositional content would have a role to play in cognition.

Unfortunately, the received view of memory is wrong. We are often able to correctly remember things that we did not previously know. For instance, I may remember that there will be a solar eclipse today on the strength of knowing yesterday that it would happen the next day. These thoughts are syntactically different, and hence propositionally distinct. Even more simply, I can remember now that something happened last year on the basis of knowing then that it was happening at that time. Similarly, I can remember that I did something on the basis of previously knowing that I was doing it. There was no time I could have known that I previously did it without relying upon memory. That is something I can only know on the basis of memory, so event memory automatically produces thoughts that are syntactically different from the thoughts produced by direct observation. In each of these cases, I am believing the same thing about the same time, but I am thinking of that time in a syntactically different way and temporally locating myself with respect to it differently, so my thoughts are syntactically different, and it follows from that they are also propositionally different.

There is a second important reason why we often remember things that we did not previously know. For example, I can remember that I am 6 feet tall. What memory gives me is a thought about my current height. However, all I previously knew was my height at some earlier time. Most of the simple facts I know about myself, like my gender, my height, or my hair color, I know by remembering, but what memory provides is the belief that I now have these properties, not just that I previously had them. This is equally true of many of my thoughts about other objects. For example, I remember that Tucson is in Arizona, that my daughters live in California, and so on. In all of these cases the thought I have as a result of memory retrieval is a different propositional type from the thought initiating memory storage. On the other hand, unlike event memory, the thought I get from these memories is plausibly the same syntactic type of thought that initiated memory storage. In these cases, memory stores thoughts containing the temporal indexical ⌜now⌝, and retrieves thoughts with the same syntactic form, but the result is a thought of a different propositional type. As I observed in my (1998), this is an efficient way of implementing temporal projection without requiring the cognizer to perform explicit inferences. So this seems to be a case in which memory tracks (or tries to track) syntactical sameness even when that conflicts with propositional sameness.

Two important lessons can be learned from these examples. First, memory is not a passive process of reproducing earlier thoughts. At least in many cases, memory is a constructive process that produces new thoughts of propositional (and sometimes syntactic) types never previously entertained. Second, there does not seem to be any role for propositional sameness in memory cognition.

2.5 Propositions, Syntax, and Semantics

I can see no role that propositions might play in cognition, and on that basis I question the usefulness of the notion of a proposition for studying cognition. There are, however, other possible uses for propositions. In particular, they might turn out to be essential for constructing a theory of language. But I will argue below that this is not true either, so I am left wondering what propositions might be for.

The conclusion that propositions play no role in cognition is a simple corollary of another principle that is commonly believed. According to this principle, cognition consists of computational processes that compute exclusively on the syntax of thoughts. Fodor has recently taken this to be definitive of the computational theory of mind, and

numerous authors have endorsed it (e.g., Stich, and probably me). If we could be confident that this principle is true, there would be no temptation to look for a role for propositions. However, this principle is fairly obviously false. For instance, most of our reasoning is defeasible. When reasoning defeasibly, if we retract a belief we should also retract conclusions we drew on the basis of it. To do that we must keep track of the basis of a belief, and our reasoning makes use of that information in belief update. However, the basis of a belief is not part of the syntax of the thought. It consists of a different kind of information that is stored with the thought (the thought being a datastructure). So reasoning is sensitive to more than the syntax of thoughts. How much more remains an open question.

Once the observation is made, it is obvious that cognition appeals to extrinsic relations between thoughts as well as to the intrinsic syntax of thoughts. Relations like the basing relation are not syntactical relations. Are they semantical? At this point the syntax/semantics distinction becomes unclear. Appeal to propositions and truth conditions constitutes what we might call "objective semantics". However, I am inclined to think that an appeal to truth conditions is generally unexplanatory. The truth condition of a thought is supposed to consist of a general specification of the circumstances under which the proposition would be true, and then a proposition is true at a world iff the truth condition is satisfied at that world. But I question whether this really makes sense. If a proposition can be given an informative logical analysis, one can use the analysis to state the truth condition. Then it can be informative to talk about truth conditions. But most propositions do not have logical analyses. Then the most natural way to describe the truth conditions of a thought is disquotationally, by appealing to the proposition it expresses. For example, the truth condition of «snow is white» is that snow is white. If this is the only way of understanding truth conditions, then an appeal to truth conditions is essentially vacuous. Does objective semantics, in all its vacuity, exhaust semantics? Semantics is supposed to govern how we should decide whether thoughts are true or false. But that is the same thing as deciding whether we should believe them. So it is not much of a stretch to view semantics as being about how we evaluate thoughts, i.e., how we decide whether we should believe a thought. The process of deciding whether we should believe a thought is a matter of engaging in reasoning involving the thought. If we can construct a good argument having the thought as its conclusion, then it becomes at least defeasibly reasonable to endorse the thought, taking it to be true. If we can construct a counterargument, that gives us reason to think the thought is false. So it looks like semantics should tell us how to reason with thoughts, and otherwise employ them in cognition. Semantics should be about "how thoughts work in cognition". Let us call this "procedural semantics". Procedural semantics includes much more than (usually vacuous disquotational) statements of truth conditions that constitute objective semantics. At least, it gives semantics something to do.

The way in which we employ a thought in reasoning is determined in part by what thought it is, i.e., by its syntactical structure and its atomic constituents. For instance, the way we employ the thought ⌜The apple is red⌝ is determined by more general rules of cognition regarding, on the one hand, definite descriptions like ⌜the apple⌝, and on the other hand, property representations like ⌜red⌝. So the semantics of thoughts are determined in part by the semantics of the subsentential constituents of thoughts. Numbered among the constituents of thoughts are representations of objects, representations of properties, various kinds of connectives, and perhaps more. In section three I will discuss representations of objects, and in section four I will discuss representations of properties.

I have cast doubt on whether propositions have any role to play in cognition, and I have questioned whether the notion even makes sense. However, insofar as it does make sense to talk about propositions, we can regard propositions as inheriting syntactic structure from the thoughts of which they are types. I have argued that for two thoughts to have the same propositional type, they must be syntactically identical, so all thoughts of the same propositional type have the same syntactic structure, and we can attribute that structure to the proposition derivatively. It follows that where the thought has a constituent representing an object or a constituent representing a property, the proposition does too. The constituents of the proposition can be regarded as semantical types for the constituents of the thoughts. For example, if the semantical type of a thought ⌜The apple is red⌝ is the proposition «The apple is red», we can regard the property representation ⌜red⌝ as being of the semantical type «red» and take that to be a constituent of the proposition. These semantical types for property representations are one natural interpretation of the term "concept". Similarly, the semantical types for object representations are what I have called "logical designators" (1983, 1984).

Talk of concepts and logical designators categorizes subsentential constituents of thoughts in the same way talk of propositions categorize thoughts themselves. But it is important to realize that if we are understanding semantics liberally as consisting of the rules for how thoughts and their constituents function in cognition (i.e., as procedural semantics), it could turn out that these rules do not partition representations of objects or representations of properties into equivalence classes at all like traditional conceptions of concepts and logical designators, and so do not generate useful notions of a concept or a logical designator. This should remain an open question. Hence our fundamental question should be how these mental representations work in thought, not how corresponding constituents of propositions work in thought. There may be no useful way to recast the former question in the latter terms.

3. Mental Designators

When a thought is about a particular object, this is by virtue of its containing a mental representation of that object. Let us call mental representations of objects (construed broadly) mental designators. The corresponding constituents of propositions are logical designators. When a cognizer has a thought about an object the thought contains a representation of the object, and the propositional type of the thought is a proposition containing a logical designator that is the semantical type of the mental designator contained in the thought. So logical designators are one kind of semantical type for mental designators.

We can distinguish between several different kinds of mental designators, each subject to different rules for use in cognition. These include definite descriptions, mental indexicals, percepts, de re designators, and perhaps more. I will discuss each in turn.

3.1 Definite Descriptions

We sometimes think of objects under descriptions. In entertaining a thought, a cognizer is thinking of an object under the description D only if, necessarily, in entertaining that thought, the cognizer is thinking about whatever object uniquely exemplifies D. It was once popularly believed that definite descriptions are the only kind of mental designator. Kripke and Donnellan were perhaps the first to realize that this is false, for a variety of reasons. At the very least, we are almost never in a position to construct a description that picks out an object uniquely unless the description

contains another designator that is not a definite description. For example, I can think of my mother as "the mother of me", but this only works because I am employing another designator for thinking of myself. Definite descriptions are almost always parasitic on other designators in this way, so they cannot provide our fundamental way of thinking of most objects.

Thoughts involving definite descriptions may well be analyzable in accordance with Russell's theory of definite descriptions. However, even if this is true, thoughts involving definite descriptions are syntactically distinct from their Russellian analyses so they are different thoughts. We must recognize definite descriptions as a genuine category of mental designator. Still, if we accept Russell's analysis (which I am inclined to do), the logic of our reasoning with definite descriptions is fairly simple. This is one of the few cases in analytic philosophy in which logical analyses have actually been found.

3.2 Mental Indexicals

We employ the mental indexicals ⌜I⌝, ⌜here⌝, and ⌜now⌝ for thinking about ourselves, our present location, and the present time, and it is now generally recognized that these are not analyzable in terms of other designators.[3] They are a sui generis kind of mental representation. To these must be added the designators ⌜up⌝, ⌜down⌝, ⌜right⌝, and ⌜left⌝. It turns out that there are purely computational reasons why a sophisticated cognizer must have such indexical representations built into its cognitive architecture. This is discussed at length in Ismael and Pollock (2005), and I will not go into it further here.

These mental indexicals are a fixed feature of our cognitive architecture — a feature shared by all human beings. As Castaneda and Perry demonstrated, when we have beliefs containing them we cannot have the same beliefs, or even logically equivalent beliefs, that do not contain them. It follows that they have unique semantical types «I», «here», «now», «up», «down», «right», «left» that are logical designators and constituents of the corresponding propositions. Propositions containing «I» are often called de se propositions, and «I» is referred to as the de se designator. One of the most interesting characteristics of de se propositions is that they are "logically idiosyncratic" — only the person who is the designatum of the de se designator can entertain thoughts of that propositional type (Pollock 1981, 1983). This is because only I can think of myself in terms of a de se designator. If someone else employs a de se designator, he is thinking of himself, not of me.[4] So even if talk of propositions makes sense, de se propositions cannot be shared. Most of our simple thoughts are actually de se. For instance, vision locates perceived objects in a reference frame having the cognizer at the origin. Thus such thoughts cannot be shared or conveyed to others by linguistic utterances. This is a problem for the Fregean theory of thought, which assumes that the role of language is to convey our thoughts to others. I will discuss this further below.

3.3 Percepts

Vision apprises us of the state of the world around us, and it does so by representing the world in various ways. As argued in Pollock and Oved (2005), the visual image is best viewed as a rich but transitory database of mental representations, and among those representations are "percepts", which are object representations. The

[3] Castaneda, Perry.

[4] This was noted already by Frege.

point of calling them representations is that they can be moved from the visual image into thoughts about the world that are justified on the basis of having the image. When I believe, on the basis of vision, that the apple on the table is red, I see the apple by having a percept of it, and the thought I have to the effect that the apple is red employs the percept as its representation of the apple. I literally think of the apple in terms of the percept. The solution to the problem of perception lies, in part, in the fact that when we hold beliefs on the basis of perception there is no inference required to generate a belief with a thought that is syntactically unrelated to what goes on in perception itself. Instead, the syntactical constituents of the thought are drawn directly from the perceptual experience (the visual image). The role of visual processing in cognition is the construction of these mental representations.

Percepts are mental objects. If we are to classify our perceptual beliefs by talking about propositions, we must have propositional constituents corresponding to percepts. We might call these perceptual designators. So percepts are mental designators, and perceptual designators are logical designators. As noted above and discussed further in Ismael and Pollock (2005), perceptual beliefs are invariably de se, locating perceived objects in a reference frame with the cognizer at the origin. They are logically idiosyncratic for that reason. But even if this were not so, it seems that propositions involving perceptual designators would have to be logically idiosyncratic anyway. Percepts are just datastructures in our cognitive virtual machine. It makes no sense to ask whether another cognizer has a datastructure of the same (unique) semantical type. Other cognizers can have datastructures that "work the same way" in a generic sense, so they are of the same general semantical type, but there is no apparent way to identify the percepts of different cognizers as being of the same specific semantical type in such a way that this would be sufficient for them to entertain thoughts of the same propositional type. For instance, suppose I see an apple through a half-silvered prism that splits the light into two images, with the result that I have two identical percepts of the apple in two different parts of the my visual field. These are two distinct mental representations. If I were unaware of the way the images were produced, I could accept a thought involving one as veridical while rejecting the corresponding thought involving the other. Consequently, the two thoughts are propositionally distinct. Now suppose someone else sees the apple from the same perspective, but without the intervention of the prism. He has a qualitatively identical percept of the apple that figures in a corresponding thought. Is his thought propositionally the same as one of my thoughts about the apple? If so, which? There is no way to decide this question, so his thought must be propositionally distinct from both of mine. Thus no matter how similar the percepts of two cognizers are, they cannot be contained in thoughts that are of the same propositional type. In other words, perceptual thoughts are logically idiosyncratic.

3.4 Skolem Designators

A natural presumption is that in deductive reasoning, we should only infer thoughts that are logical consequences of the premises of the reasoning. However, when we teach students to construct derivations in the predicate calculus, we often teach them rules that do not have this characteristic. Specifically, when reasoning with existential quantifiers, we teach a form of existential specification that allows the reasoner to move from $(\exists x)Fx$ to Ft where t is a "special" singular term that is used like a free variable. The rules are designed in such a way that although conclusions containing t need not follow logically from the premises, conclusions inferred from Ft but not containing t do follow logically from the premises. The precise explanation for why this works has to

do with the Skolemization of first-order formulas and the introduction of Skolem terms, but we need not go into the details of that here.[5]

I think it is generally supposed that existential specification and the use of Skolem terms is a human invention that makes deductive reasoning more convenient, not a characterization of the form of human thought. I, at least, used to suppose that. But I know think that the innate structure of human thought has a similar form. Suppose I come to believe that there is something (not necessarily just one thing) that is F. So I start with an explicitly existential belief (∃x)Fx. Suppose I believe that all F's are G's, and also all F's are H's, and in addition I believe that anything that is both a G and an H is a J. Then I should be able to infer that something is a J. The way I would naturally perform this reasoning can be transcribed linguistically as follows. I note that something is an F. Because all F's are G's, "it" is a G. Because all F's are H's, "it" is also a G. But anything that is both a G and an H is a J, so it is a J. Thus something is a J.

Imagine thinking this through in your head — not writing it down or saying it to someone else. This is about thought, not about language. For instance, suppose you believe that someone is in the garden. But no one is supposed to be there now, so he is not supposed to be there. The jewels are hidden in the garden, so he is in the garden when the jewels are there. You believe that anyone who is improperly in the garden when the jewels are hidden there is probably a thief who is trying to steal the jewels. So you infer that he is a thief. Hence, you conclude that there is a thief in the garden. In this reasoning, you draw two intermediate conclusions: (1) "He" is not supposed to be in the garden; (2) "He" is in the garden while the jewels are hidden there. From these you conclude that "he" is a thief who is trying to steal the jewels, and hence that there is a thief in the garden. The "he" in this reasoning is functioning as a singular term. Your thoughts contain what we may be tempted to call a mental designator of the person in the garden. However, putting it in this way suggests that you are thinking of a single individual. In fact, you may not be making the assumption that there is just one person in the garden. There may be several. Thus your thought contains a mental item that plays a syntactical role similar to that of a mental designator, but it does not purport to designate a unique object. We might say that it purports to "indifferently" designate each of a set of objects, but that is not real designation. Even though these are not real designators, it is convenient to call them "Skolem designators" because of their similarity to Skolem terms in first-order reasoning.

What I want to conclude from these considerations is that our thoughts do contain Skolem designators as syntactic constituents, and that these designators are lexically simple. Although the designator is in a sense "derived from" the existential belief (∃x)Fx, the property representation F is not a syntactic part of the designator. One might resist these claims by insisting that my observations are about language, not about the structure of the underlying thoughts. Thus far, my claims are based only on instrospection of what I am thinking. We can never get away from introspection altogether. After all, that is our ultimate access to our thoughts. But I can give logical arguments to buttress my introspective claims. First, I claim that our thoughts contain a lexical item that functions syntactically like a mental designator. The only apparent alternative is to insist that the thoughts that are our intermediate conclusions cannot retain existential quantifiers throughout, and the appeal to "he" and "it" in describing the reasoning is just a linguistic convenience. However, it is not difficult to see that our

[5] In fact, the details are quite complicated. See my "Natural Deduction II".

thoughts do not retain the existential quantifiers. Consider the form our reasoning would have to take in order for that to be the case. From (∃x)Fx, (∀x)(Fx → Gx), and (∀x)(Fx → Hx) we can infer (∃x)Gx and (∃x)Hx. But from the latter and (∀x)[(Gx & Hx) → Jx] we cannot infer (∃x)Jx. To draw the latter conclusion we need the more complex intermediate conclusion (∃x)(Gx & Hx) rather than the two separate conclusions (∃x)Gx and (∃x)Hx. To get (∃x)(Gx & Hx) from our premises, we would presumably first have to infer (∃x)(Fx & Gx), and then because (∀x)(Fx → Hx), infer (∃x)(Gx & Hx). So the structure of our reasoning would have to be something like this:

(∃x)Fx
(∃x)(Fx & Gx) because (∀x)(Fx → Gx)
(∃x)(Gx & Hx) because (∀x)(Fx → Hx)
(∃x)Jx because (∀x)[(Gx & Hx) → Jx]

At this point we have to rely upon introspection again, but it seems clear that our thoughts are not this complicated. I simply think, "It is a G", not "Something is both an F and a G". And I never form the conjunctive belief that something is both a G and an H. Instead, I have the separate beliefs that "it" is a G and that "it" is an H. So my reasoning has the following form:

(∃x)Fx
Ft (introduction of Skolem term)
Gt because (∀x)(Fx → Gx)
Ht because (∀x)(Fx → Hx)
Jt because (∀x)[(Gx & Hx) → Jx]
(∃x)Jx

Our intuitive reasoning is logically correct if we understand our thoughts as having syntactic forms like what we get from employing Skolem terms, but there is no way to make it logically correct if we insist that it retains the existential quantifiers. The reason is simply that from Gt and Ht you can infer (Gt & Ht), but from (∃x)Gx and (∃x)Hx you cannot infer (∃x)(Gx & Hx).

At this point the observation that our thoughts contain Skolem designators is more of a curiosity than a profound philosophical conclusion. My reason for interjecting it into the present discussion is to soften the resistance to some similar observations to be made in the next section.

3.5 De re Representations

Thus far I have talked about familiar kinds of mental designators. Now I want to argue that what is perhaps the most pervasive kind of mental designator is not of any of these familiar forms. Consider thinking about an object with which you are familiar. If you know a lot about it, you might be able to think of it in terms of a definite description, but we rarely do. When I think of one of my daughters, thinking that she lives in California, I could construct definite descriptions that I would regard as picking her out uniquely,[6] but I could discover of any such description that she does not

[6] They would not be purely qualitative. They would most likely describe her in terms of her relations to me, where I think of myself in terms of a de se designator.

actually satisfy it, or that she is not the unique person that satisfies it. It follows that I am not thinking of her under that description. Instead, I think of her in another way and, while thinking of her in that way, I have the contingent thought that she satisfies the description uniquely. Just to have a label, I will call these "de re thoughts". I will attach further theoretical baggage to this label shortly.

3.5.1 Syntactic Considerations

What is the mental representation that I am employing when I think of my daughter in this way? Waving vaguely towards theories of proper names in the philosophy of language, it might be tempting to propose that my thought is "directly referential". But it is not clear what that would even mean in the present context. What we want to know is how the thought comes to be about a particular object. Thoughts are mental objects. My daughter cannot literally be part of my thought about her. One might have a theory of direct reference for propositions that took objects to be constituents or propositions,[7] but that makes no sense as applied to individual thoughts (mental particulars). The thought must contain another mental object as a constituent, and be about that object by virtue of the second mental object representing my daughter. In other words, the thought must contain a mental designator for my daughter. Let us call this a "de re representation".

It might be supposed that although de re thoughts are not directly referential, the propositional type of a de re thought is directly referential. It is not clear what we want to say about the structure of a directly referential proposition, but any construal of this would require either that it is impossible to have two different de re representations for the same object, or that when I have de re thoughts involving different de re representations that represent the same object, they are propositionally identical. Examples like Hesperus/Phosphorus seem to be clear counterexamples. If I am unaware that the Morning Star and the Evening Star are the same object, I will very likely think about them in terms of two different de re representations. While employing those representations I may have the belief that the Morning Star is not the same thing as the Evening Star, but refrain from believing that the Morning Star is not the same star is the Morning Star. Thus these are two different beliefs, and hence the propositional types of my thoughts are different. Hence there is no sense in which the propositions are directly referential.

The de re representation I employ in thinking about my daughter is not any kind of simple definite description. But it might be suggested that it is a logically complex "Searle-type" definite description. Roughly, such a definite description would characterize my daughter as the unique thing that maximally satisfies $P_1,...,P_n$, where the latter are all the beliefs I have about her. This is analogous to a theory of proper names that Searle once proposed. Here we have to distinguish between two proposals. The weaker proposal would be that this describes the logic of the mental designator that I employ in thinking about my daughter. I will return to this proposal shortly. The stronger proposal would be that I literally think of my daughter under this definite description. The stronger proposal is easily rejected on the grounds that my thought is syntactically distinguishable from a thought involving any such long and complex definite description. I do not think anything that complicated when I think that my daughter lives in California. In fact, it seems introspectively that my mental representation for my daughter is devoid of syntactic structure — I just "think of her", I

[7] Bertrand Russell had such a view at one time.

don't think any syntactically complex description that picks her out. I will discuss this issue further below. In my (1980) and (1983), I introduced the term "de re designator" for the propositional designators corresponding to these mental representations, and I am now calling the mental representations themselves "de re representations".[8] My suggestion is that we employ such a class of syntactically simple representations in our thoughts about the world.

Let us consider more carefully the claim that de re representations are syntactically simple. This strikes me as very plausible on introspective grounds alone, but I would prefer not to rest my case exclusively on introspection. The data of introspection can be buttressed with an argument. The argument has two parts. First, it must be argued that de re thoughts really do have subject/predicate form with the subject being the de re representation. Here we can appeal to the same argument I gave above for Skolem representations. The only apparent alternative would be that what I am calling "de re thoughts" are complex constructions involving quantifiers. Once I come to think de re thoughts about an object, I can think many different such thoughts about the same object, and they are literally different thoughts, entertained at different times and interspersed with many other thoughts about other matters. It follows that we cannot regard them all as being constituents of one big complex de re thought wrapped in some combination of quantifiers. Thoughts, as mental objects, have to exist at particular times. They cannot be built out of constituents that exist at different non-overlapping times. This is analogous to the observation regarding Skolem designators that I think the separate thoughts Gt and Ht, not the complex thoughts $(\exists x)(Fx \,\&\, Gx)$ and $(\exists x)(Gx \,\&\, Hx)$. I may have the thought that my daughter is currently in California, and then later, after thinking many other thoughts about other matters, have the thought that I will spend Thanksgiving with her. There is no single thought that is the logical compound of all the individual things I think about her. Thus my de re thoughts about her do not form a single thought wrapped in quantifiers. The alternative is that each individual thought involves quantifiers rather than a syntactically simple de re representation. But here we encounter the same problem we encountered for Skolem representations. I can combine my different thoughts about my daughter and draw inferences from several of them to new conclusions about her. If my thoughts were simple existentially quantified thoughts, that would not be logically valid.

There is just one way my de re thoughts could be existentially quantified and still allow inferences from combinations of separate de re thoughts. This would be if they are more complex quantified beliefs taking the form of Russellian definite descriptions. That is, from

$$(\exists x)[(\forall y)(Fy \times y = x) \,\&\, Gx]$$

and

$$(\exists x)[(\forall y)(Fy \times y = x) \,\&\, Hx]$$

we can infer

$$(\exists x)[(\forall y)(Fy \times y = x) \,\&\, (Gx \,\&\, Hx)].$$

So the only way my de re thoughts about my daughter could be existentially quantified and still play the kind of role they do in reasoning is if they are really definite descriptions.

Might de re representations be (or work like) definite descriptions? Consider what

[8] See also Kent Bach.

definite descriptions they might be. Here we can rehearse familiar arguments taken from the philosophy of language and aimed at a similar views of proper names. De re representations cannot be simple definite descriptions. For example, when I think of my daughter Erika I am not thinking of her as "my first daughter", because it is always possible that, without knowing it, I had another daughter before Erika. This objection can be raised against any simple definite description. The alternative is very complex definite descriptions, like Searle-type definite descriptions. But I have already argued that my de re thoughts are introspectively distinguishable from thoughts involving any such complex definite descriptions. So de re representations cannot be identified with any definite descriptions. They are a distinct kind of mental designator and play an important role in thought.

So de re representations are a primitive kind of mental designator. We can go further and observe that they have no syntactic structure — they are syntactically simple. The reason for saying this is that there is no role for syntax to play in de re representations. The only mental designators that have syntactic structure are definite descriptions, and the reason definite descriptions have syntactic structure is that the built-in property representation dictates how we reason with a definite description. I have argued that no property representation can play a similar role for de re representations, because when we think of an object in a de re way, it is always possible for us to discover that the object lacks any particular property we might have thought it had uniquely.

3.5.2 A Schematic Semantics for De Re Representations

Thus far I have argued that de re representations are a unique syntactically simple kind of mental designator, and they constitute the way in which we most commonly think of individual objects. The harder question is how de re representations work. That is, what determines the representatum of a de re representation, and how can we employ them in reasoning? An answer to this question describes the semantics of de re representations.

I argued above that de re representations are not literally Searle-type definite descriptions, but the weaker proposal that de re representations "work like" Searle-type descriptions is not implausible. We might call this the "belief satisfaction model" of de re representations. However, it fails for the same reason this view fails as a theory of proper names. Suppose you point someone out me, and I begin thinking of him in a de re way. Then you begin telling me interesting stories about him. If I am sufficiently gullible, this might go on for years, despite the fact that everything you tell me about this person is pure fabrication and you had never seen the person before you pointed him out to me. I am still thinking about the person you pointed out, but the huge preponderance of my beliefs about him are false. On the other hand, they might, purely by chance, be largely true of someone else. This would not make it the case that I am thinking of that other person who just accidentally satisfies most of my beliefs. My de re representation remains a representation of the original person. Thus the belief satisfaction account cannot be correct.

If the representatum of a de re representation is not determined by the cognizer's beliefs, how can it be determined? The only thing left that might play a role in determining the representatum would seem to be the representation's history. In my (1980) and (1983) I proposed that one must always begin with some other way of thinking of the representatum — another mental representation — and then the de re representation is "anchored by" that initial representation. We continue to think about the object using the de re representation and may subsequently lose the ability to think

of the object in terms of the original representation. For example, my initial contact with an object is often through perception. I see an object and then begin thinking about it in a de re way. When I no longer see the object I cannot continue to think about it perceptually, so the de re representation gives me a way of continuing to think about the object. Thus we can say that the de re representation is anchored by an initial percept. However, this is not the only way I can get started using a de re representation. I think of Aristotle and George Washington in terms of de re representations too. In those cases my de re representations are anchored by a different kind of representation.

This suggests a schematic semantical theory for de re representations. The schematic theory has two parts. First, a de re representation must be "initialized" by anchoring it to another mental designator. Second, subsequent occurrences of de re representations of that syntactic type continue to represent whatever was represented by the initial anchoring representation. Notice that this theory attaches a semantics to mental items in terms of their syntactic types. Each mental object (a lexical item) that is of the grammatical type "de re representation" gets a representatum that it shares with all other mental objects of the same syntactic type.

3.5.3 Anchoring a De Re Representation

I will ultimately replace this simple schematic theory with one that is a bit more complex. But first we must consider how to fill out the details of the theory. For that purpose we need an account of how de re representations get anchored. For this, let us consider a variety of cases.

Perceptual Anchoring

The simplest form of anchoring is perceptual anchoring. Upon seeing an object, one can immediately begin thinking about it, and one can continue to think about it even when one ceases to see it. Perception computes a representation of the object, and the simplest view is that the percept thus computed literally is a de re representation. If this is right, perceptual anchoring is a special case of anchoring which does not after all consist of anchoring the de re representation to a prior mental designator. Instead, the anchoring is provided by the act of perceiving the object.

Descriptive Anchoring

Sometimes, a de re representation can be anchored by a definite description. Examples of this are difficult to come by, but they do exist. For example, in a marvelous series of tongue-in-cheek thrillers, Elizabeth Peters introduces Amanda Peabody, married to the British archeologist Radcliffe Emerson. Together they solve murder mysteries against the background of late-19th century Egypt. A central theme of the books is that at a certain point Peabody hypothesizes that there is a "Master Criminal" behind a series of archeological thefts, and dubs him "Sethios". She is not sure Sethios exists, although in fact, he does. In trying to confirm her theory, she thinks about Sethios, thinking about what he must be like if he exists. In this, she really is thinking about Sethios, despite her uncertainty that he exists, and it seems pretty clear that she is thinking about him in a de re way. Despite the fact that she begins with a definite description like "the master criminal behind these crimes", for the reasons given in section 3.5.2, she is not subsequently thinking of him under that description. For instance, it could turn out that one of the crimes was the work of a separate gang of thieves. Peabody employs a syntactically simple designator to think about the putative villain. So this is a case in which a de re representation is initialized by thinking of an object in terms of a definite description, and it is anchored by the definite description.

Linguistic Anchoring

Perceptual anchoring and descriptive anchoring are easily understood and not particularly surprising. What is more surprising is that a large proportion of our de re representations are not anchored in either of these two ways. This results from the fact that most of our de re representations that are not anchored perceptually are initialized by verbal communication with other cognizers. Suppose Alfonse starts telling me about an auto accident he witnessed. He says, "As soon as the light changed, this white car started to go. But as it pulled into the intersection, a blue car came out of nowhere, ran the red light, and hit the white car broadside. The driver of the blue car got out and ran down the alley. The police found him later hiding behind a garbage can. The white car was totalled, and when the police examined it they discovered the trunk was full of illegal drugs, so they arrested the driver." When I hear this story, I immediately begin thinking about the white car, the blue car, and the drivers of both, and it seems to me introspectively that I think of them in a de re way. This introspective datum can be confirmed by appealing to now familiar arguments based on the observation that I do not just think one big thought in response to hearing the story. Instead, I think a series of separate thoughts, but they are about the same individual objects. Consider how these de re representations get initialized.

All Alfonse tells me is that there existed four things that mutually satisfied the various constraints recounted in his story. But my mental designators are not just Skolem designators derived from an existential thought. As a result of hearing his story, I acquire mental designators for thinking about a particular four things, namely, the ones Alfonse was talking about. For example, I can ask where the white car was taken when it was towed away. I can then go to the wrecking yard and find it. So it is a particular white car I am thinking about. How are my de re representations anchored? Certainly not perceptually — I did not see the accident. It is initially plausible to suppose they must be descriptively anchored, but what are the definite descriptions that anchor them? The content of what Alfonse told me does not provide enough information for me to construct definite descriptions on its basis. There has probably been more than one such accident in the history of the world. What is unique about this accident is only that it is the one Alfonse was telling me about. In making his statements, he was thinking about particular blue and white cars and their drivers, and it is by virtue of this that I come to think about them. This suggests that my de re representations are anchored by definite descriptions, but the definite descriptions make essential reference to the communicative act and not just to its contents. Perhaps at the time the de re representations are initialized I think of the blue car as "the blue car Alfonse is telling me about", or more generally as "the object that played such-and-such a role in the events Alfonse is describing".

But consider a more muddled case of verbal communication. I am hiding in a closet and overhear a conversation in which several speakers are talking about the accident. I do not know the identity of the speakers, and cannot hear them very well over the background noise and through the thick closet door. I am not even sure how many speakers there are. Still, as they talk I form de re representations of the objects they are talking about. If my de re representations are initialized by definite descriptions, the descriptions cannot make reference to the speakers, because I am unsure who is speaking. It cannot make reference to a precise role in the story, because I am hearing only a garbled version of the story. Might the initializing description for the blue car just be something like "the blue car that the speakers I am now listening to are talking about"? I think not. Suppose the first car belonged to the Bloo exterminating company, and the speakers were actually saying "the Bloo car", but I misheard them. My de re

representation still represents the car they were talking about. I am just wrong in thinking it was blue. So the de re representation cannot be initialized by a definite description containing "blue". Similar variations on the example seem to rule out definite descriptions based on any other concrete parts of what I take the story to be. But I still have de re representations representing the objects being discussed, and from hearing the conversation I form some true beliefs and some false beliefs containing these de re representations.

Of course, I am hearing a real conversation, so I might employ a description that appeals to that conversation and not just to what is said. But what would the description be? A number of different objects are being discussed in the course of that conversation, and each of my de re representations represents a particular one of those objects. We have seen that my definite descriptions cannot discriminate between these objects in terms of properties recounted in the conversation. Could it appeal instead to the properties I took the speakers to be attributing to the objects? No, because I cannot take a speaker to be attributing a property to an object unless I can already think about the object, and it is the initiation of my way of thinking of these objects that is at issue. Might my description instead appeal to my thinking that the speakers were talking about some unique object or other and attributing certain properties to it (it was blue, it was hit by the white car, it carried illegal drugs, etc.)? That does not work either, because as soon as I hear the first mention of the blue car, I start thinking of it in a de re way, and then as I hear more of the story I acquire additional thoughts involving that de re representation.

The preceding observation suggests another account. As I hear the speakers talk, I automatically identify certain parts of their speech as utterance of referring expressions. Thus I understand the utterance bloo kär as purporting to refer, and my de re representation is initialized by the definite description "the thing to which that utterance refers". But this cannot be right either. The difficulty is that although I do, in some sense, identify utterances of referring expressions, this is not something that I do explicitly or consciously. The observation has often been made that the visual image is "transparent", in the sense that we see the world "through it", without thinking about the image itself. The same thing is true of language. Language is basically a tool for manipulating the thoughts of the audience, getting them to think appropriate thoughts in response to the speaker's utterances, and at least in straightforward cases it does this without requiring the audience to think about the utterances they are hearing. Their thoughts are the thoughts they get from hearing the utterances, not thoughts about the utterances. So in most cases it just isn't true that I will have the thought that an utterance of bloo kär just occurred and it purports to refer. Instead, in response to hearing the utterance of bloo kär I will just automatically get a thought involving a de re representation and attributing the property of being a blue car to its purported representatum.

What this illustrates, I believe, is that language plays a special role in thought, in some ways analogous to perception. This is a topic that I will take up in more detail towards the end of the paper. What is important for now is that verbal communication seems to play a special role in initiating de re representations — a role that cannot be reduced to descriptive anchoring. Somewhat like perception, understanding a linguistic utterance involves our cognitive virtual machine computing appropriate thoughts, and part of that process may involve computing new de re representations, just as perception involves the computation of a new percept. I will refer to this as linguistic anchoring. Of course, there are cases in which communication does not go smoothly and we have to think explicitly about what the speaker is saying and reason about what

he is trying to convey, and in those cases de re representations may actually receive descriptive anchoring. But I take linguistic anchoring to contrast with this. By definition, linguistic anchoring is not descriptive anchoring.

The anchoring relation is supposed to determine what a de re representation designates. Linguistic anchoring accomplishes this by anchoring a de re representation to another cognizer's referring utterance, and hence ultimately to some mental designator employed by that cognizer to think about whatever object he is talking about. If the speaker's putatively referring utterance fails to refer to anything, then the hearer's de re representation fails to designate anything. For example, in my (1980) I gave the example of Dilapedes. Richard tells me about a pre-Socratic philosopher named "Dilapedes", whose view was that everything is broken. By coincidence, there really was such a person but Richard does not know there was and is just making it up. On the strength of what Richard says, I acquire a de re representation, but does it designate Dilapedes? It does not seem so. That is because this is a case of linguistic anchoring, not descriptive anchoring. My de re representation designates whatever Richard was talking about in using the term "Dilapedes", but in fact he was talking about no one, so my de re representation fails to designate.

Linguistic anchoring makes thought a partly social phenomenon. I will suggest later that this is true for other reasons as well. But for now notice that linguistic anchoring constitutes a kind of "historical connection" account of de re representations. Perhaps this is what lies behind the appeal of historical connection theories of proper names.

Reconsidering the Anchoring Relation

The schematic theory with which I began this section supposed that de re representations are initiated by anchoring them to prior mental designators of other kinds. That may be an accurate way of thinking of descriptive anchoring, but it is not an accurate description of either perceptual or linguistic anchoring. In both of those cases, the de re representations are the first mental designators we have of the objects we are thinking about. They are not anchored in earlier designators, but instead in the causal processes that produce them. Is there a philosophical story to be told about the anchoring relation? Probably not. From a first-person point of view, it makes little sense for me to ask what determines what I am thinking about. I am thinking about what I am thinking about. I may only be able to think about the referent of a mental designator disquotationally — by using the designator. If I see a bird high in the sky and you ask me, "Yes, but what are you seeing?", I may only be able to answer in puzzlement, "That", thinking of the bird in terms of the very percept you are asking about. Similarly, if a speaker tells me about some object of which I have no previous knowledge, and you ask, "Yes, but what are you thinking about?", my only answer may be to think of the object in terms of the very de re representation I got from the speaker's utterances and say, "That". Of course, this won't satisfy you, but there is no reason I must be able to give a better answer in order for my de re representation to represent.

From a third-person perspective, if we ask what determines the representatum of a cognizer's percept, we are asking a different question. In that case we are asking how visual cognition works, and that seems to be an entirely empirical matter. There is no a priori philosophical analysis that will tell us what object in the world is being perceived when perception computes a particular percept. That depends on the details of how perception works, and can only be revealed by empirical investigation. Note that this investigation cannot be pursued just be examining the experimental subject. We must know not only what is occurring in his perceptual system, but also what he is seeing. Ultimately, we can only know the latter by assuming that he is like us and knowing

what we would be seeing under those circumstances. In other words, to carry out such an empirical investigation we must employ both the first-person and third-person perspectives. We know what we would be seeing in various perceptual situations, although the identification of the objects of perception is disquotational. That is, we think of the perceived objects in terms of the percepts we have of them. We assume that experimental subjects will see the same thing we see under normal perceptual circumstances, and then by varying the circumstances we can investigate how their perceptual mechanisms work.

I suggest that something similar is true of language comprehension. The computation of de re representations in response to hearing linguistic utterances is an automatic psychological process — not something we control voluntarily. When we understand an utterance, the construction of de re representations is part of the process of understanding, just as the construction of percepts is part of the process of seeing. If the hearer understands the utterance correctly, his de re representations will represent whatever the speaker was talking about. So an account of what the de re representations represent depends on an account of what it is to understand an utterance. Whether this can be completely elucidated by giving an a priori philosophical analysis is at this point an open question, although I am skeptical about the possibility of that.

The remaining case is descriptive anchoring. Here it might seem initially more plausible to suppose that there is a philosophical analysis to be given. This analysis would provide an a priori analysis of what is represented by a de re representation that is initialized by descriptive anchoring. But I am not sure that this is determined a priori. The de re representation represents whatever the anchoring definite description represents, but the process of initializing the representation is a psychological process, and not something that we do intentionally. It just happens to us as we think. The conditions under which it happens are no more subject to a priori analysis than are the conditions under which we acquire percepts representing objects in our vicinity. It will take an empirical investigation of the appropriate aspects of cognition to determine when such initialization takes place. So it seems there is no a priori analysis to be given of the relationship between a de re representation and a definite description which determines that the latter anchors the former, and so no a priori analysis to be given of what the de re representation represents, even given that it is anchored by some definite description or other. This problem is compounded by the fact that once the initialization has occurred, we do not seem to have any privileged access to how the representation was anchored. For example, I know other members of my profession and think of them in a de re way, but in many cases I cannot now remember whether I saw them before I heard about them from others. So I am unsure whether my de re representation was initialized by perceptual anchoring or linguistic anchoring. Thus it seems to me that an account of the anchoring relation is basically a psychological problem, not a problem that can be solved by doing armchair philosophy.

3.5.4 The Semantics of De Re Representations

Now let us return to our fundamental question about de re representations. We want to know how they work in cognition. A theory of this describes the semantics of de re representations. We began with a schematic semantical theory according to which a de re representation is initialized by anchoring it to some other representation, and then the proposal was that de re representations that are syntactically the same as the anchored de re representation will henceforth represent whatever the anchor represents. We have now seen that this schematic theory must be modified, because

anchoring is not usually anchoring to another representation. That is only true of descriptive anchoring, which is the least common case. We can instead understand anchoring to attach a de re representation directly to a representatum. The modified theory then says that a de re representation represents an object x iff either (1) it is the first token of a newly constructed syntactic type and it is anchored to x, or (2) it is of the same syntactic type as some earlier de re representation that was anchored to x. The basic idea is that de re representations get their representata by one of the three kinds of anchoring we have discussed, and then once they are anchored, subsequent tokens of the same syntactic type have the same representatum.

This theory has the consequence that, in effect, representata attach to syntactic types of de re representations, not just to the individual representations of those types, and the representatum of a syntactic type of de re representation cannot change. Unfortunately, there are examples that seem to show that this is false. Consider an example taken from a novel (whose title I have forgotten). A man meets a woman and falls in love with her. Then he leaves on a business trip. While he is gone, the woman is murdered, but the murderers were mistaking her for her twin sister. So as not to reveal that she is still alive, the twin takes the place of her sister. The man returns to find the twin in the place of the first sister, but does not realize it. He "continues" his love affair, which becomes a stable relationship that extends over a period of years. Only much later does he learn about the switch, but at that point his relationship is with the second sister. Initially, he was thinking of the first sister in terms of a de re representation. When he returned, he still thought of her that way but mistakenly identified the second sister with her. Subsequently, whenever he encountered the second sister he was led to form beliefs involving the de re representation, but the beliefs were false because by virtue of containing the de re representation they were beliefs about the first sister. But surely a time comes when he is thinking about the second sister by employing the de re representation. So the representatum of the representation can change over time.

In the twin sisters case, the representatum changes in response to subsequent perceptual acquaintance with the twin. But the same thing can happen when the subsequent information is linguistic. Suppose Keith points out a man at a diplomatic reception. I thereby acquire a perceptually anchored de re representation of him. The man is a bit tipsy, and his companions find his behavior hilarious. As we are leaving, Keith tells me in an offhanded way that the man we were watching is the President of the small island nation of Aegelia. In this he is mistaken, although the man we saw does look a bit like the President of Aegelia. Because of his humorous behavior, I remember this man, and continue to employ my (syntactic type of) de re representation in thinking about him. Among my beliefs about him is the belief that he is the President of Aegelia. I then read about the President of Aegelia in the newspaper, and because I believe he is the man I saw, I form beliefs containing the perceptually-derived de re representation. The President of Aegelia is transported into a position of international prominence when he undertakes a peace-making role in the Mideast, and I continue to form beliefs containing tokens of this de re representation. At some point I may completely forget about having seen the man or his having been pointed out by Keith. As time passes, whenever I read about the President of Aegelia, I form beliefs containing tokens of the de re representation. This may go on for years as the President of Aegelia becomes Secretary General of the United Nations, wins a Nobel Peace Prize, etc. Surely, at some point, when I have thoughts containing tokens of the de re representation I am thinking about the President of Aegelia and not about the man at the reception.

The Dilapedes case can be modified to generate similar results. There, I get my de re

representation from Richard's made-up story (by linguistic anchoring), and at that point I am thinking about the real Dilapedes. But suppose having once acquired the syntactic type of representation in that way, I go on to read about Dilapedes in historical texts about the pre-Socratics. As I read, I form beliefs containing tokens of the de re representation initiated by Richard's tale. Surely, at some point I am thinking about the real Dilapedes in this way, even though initially I was not. When I first encounter an historical reference to Dilapedes and form a belief using my de re representation, that belief is false, because it is a belief about the person Richard was purportedly talking about. But as time passes and I form more beliefs in this way, it seems that the initially false beliefs (i.e., the syntactically individuated thoughts, not the propositions they express) become true. In other words, the syntactically individuated thoughts containing the de re representation come to be of a different semantical type and become true as their semantical type changes.

Here is another example with some similar features but an interesting difference. Knowing little about the history of philosophy, I read a bit about Aristotle, learning something of his philosophical views but missing the important detail that he died two thousand years ago. In another context, I read about Aristotle Onassis and mistakenly think that he is the same man. I have a single de re representation that occurs both in beliefs about there being a philosopher with certain views and beliefs about a Greek shipbuilder who married the widow of an American president. Who, if anyone, am I thinking about when I employ this de re representation? As I first read about Aristotle the philosopher, I was at that point acquiring true beliefs about the philosopher. When I subsequently read about Aristotle Onassis and confused him with Aristotle the philosopher, I was acquiring false beliefs about the philosopher. But as time passed and my beliefs involving the de re representation derived equally from reports of Aristotle the philosopher and Aristotle Onassis, it does not seem right to say that all my beliefs involving this de re representation are about Aristotle the philosopher. Nor does it seem right to say that I have many beliefs, some true and some false, about Aristotle the philosopher, but I have no beliefs at all about Aristotle Onassis. Instead, it seems that I have a bunch of confused thoughts that cannot be sorted out as being about one of these people rather than the other. Notice that if I eventually discover my confusion, I cannot go back and sort out my beliefs, saying that some are about Aristotle the philosopher, and others are about Aristotle Onassis, and all are true except the belief that Aristotle the philosopher is the same person as Aristotle Onassis. Instead, upon learning that these were two different men I acquire two new de re representations, and can then try to sort out my beliefs by forming separate beliefs about the two men on the basis of the original confused mix of beliefs. I do not, in this way, reaffirm any of my earlier beliefs. I just use them as raw material for constructing new beliefs involving the new de re representations. One thing this example illustrates is that a de re representation can lose its representatum without at the same time acquiring a new one.

In all of these examples, I start with a de re representation anchored in one way and then have contact (perceptual, linguistic, or descriptive) with a different object or person that would have led to the initiation of a new de re representation anchored in a new way if I did not mistakenly believe that the new object is the same object as the previous representatum. We can refer to these as "potentially anchoring contacts". The first potentially anchoring contact results in the de re representation being anchored. The de re representation may retain its representatum in the face of some further conflicting potentially anchoring contacts, but it seems that if there are enough of the latter they can eventually detach the de re representation from its original

representatum and re-anchor it to a new representatum. Or, as in the Aristotle example, they can detach the de re representation from its representatum but fail to re-anchor it because there is not sufficient agreement among the new potentially anchoring contacts.

These seem to be two different questions: (1) what was I thinking about when I was confused; (2) how would I correct my beliefs if I discover the confusion. We can think of the first question as being about the objective semantics of de re representations, and the second as being about the procedural semantics. Consider the twins case. It is clear that I am thinking about the second sister. I would reject beliefs that were originally about the first sister. Similarly, in the Dilapedes case I should reject beliefs that I got from Richard, unless they were later confirmed by what I read. This is because they are not about the current representatum of the de re representation. So although these are two different questions, they are connected. In general, I should reject beliefs that were originally about a different representatum unless I subsequently acquired reasons to hold them regarding the current representatum.

Putting it that way gets things slightly backwards. We do not want to have to decide what our thoughts are about before deciding what we should believe. Rather, we think of our current representatum disquotationally, and when we have the belief that an earlier thought containing the de re representation was originally about something different than the current representatum (so we are thinking about the earlier thought), that is a reason for ceasing to believe it. Note that we have to think about the original thought, not think the thought, because if we think the thought it is about the new representatum. We cannot think about the original representatum by thinking the thought now.

Let us note in passing that de re thoughts, like de se thoughts and perceptual thoughts, are logically idiosyncratic. That is, no one else can have a de re thought that is of the same semantical type as one of my de re thoughts. The only way to compare my de re representation semantically with a de re representation employed by another cognizer is in terms of their functional properties. But a single cognizer can have two different de re representations that are functionally the same. For example, I might begin with a definite description and acquire a de re representation that is anchored in it. If I subsequently forget that the object I think about in that way satisfies the definite description, I may later acquire a second de re representation anchored in the same definite description. Another person could similarly have a de re representation anchored in that definite description (provided the definite description contains no logically idiosyncratic constituents), but there would be no way to decide which of my semantically distinct de re representations is semantically the same as his de re representation. So there just isn't any way to make sense of the idea of two different cognizers having de re thoughts that are semantically the same. In particular, they cannot entertain the same de re propositions.

4. Concepts

- What are concepts and theories of concepts for?

(1) categorization

- This is a matter making judgments about objects, judging them to exemplify concepts.
- Note that I am distinguishing between concepts, designators, propositions, logical operators, etc. Concepts are constituents of thoughts (better, propositions) that are used in categorization. In other words, they are semantical types of property representations. It is important to make this distinction at least initially, because it is not clear without an argument (and it is probably false) that all of these logical or mental items work in the same way.

(2) constituents of propositions

- If we are careful to make a type/token distinction here, we will say that concepts are constituents of propositions. The corresponding constituents of thoughts are property representations. (Concepts are not "in the head" — representations are.)
- Concepts are used for more than categorization (property attribution). They are also used for thinking collectively about all exemplars of a property.
- Note that this is to use "concept" more narrowly than some people do (Fodor, Jackendoff).

(3) determine extensions

(4) used in formulating a theory of linguistic meaning

- It is most commonly assumed (following Frege) that belief types and statement types are the same, viz., propositions. But this seems to be false. Consider "I" (Pollock 1979, 1983). Propositions and concepts may still be used in a theory of meaning, but they must be used in a somewhat indirect fashion. It may or may not be true that the meanings of some lexical items are concepts. It will follow from the discussion below that in most cases they are not.

(5) pick out properties, where properties play a role in scientific laws.

- If scientific laws are propositions, then it should be concepts rather than properties that are their constituents. If they are not propositions, we cannot believe them. And if they are truths, what kind of thing are they? It might be better to say that they are "about" properties by virtue of containing concepts of representations of properties, and the sense in which they are about properties is that different concepts picking out the same property can be interchanged in nomic generalizations. That in turn is equivalent to requiring that nomically equivalent concepts express the same generalizations.
- properties are coarser grained than concepts. E.g., being lightning and being a certain kind of electrical discharge are supposed to be the same property. One might suppose that properties are something like equivalence classes of nomically equivalent concepts. (This does not handle the case of counterlegal properties.)
- Properties are supposed to be "ways things can be". I use this formulation to indicate to the reader what I am talking about, but both logically and ontologically properties are just as philosophically problematic as concepts, so it will not help to try to analyze concepts in terms of properties.

(6) explaining psychological phenomena, like typicality effects.

- Although this has been influential in the development of psychological theories of concepts (e.g., prototype theory and exemplar theory), it probably does not show anything about the structure of concepts. E.g., "Even number" exhibits typicality

effects.

- The most fundamental role of concepts seems to be as constituents of propositions, in which role they pick out properties. This is how I will henceforth think of concepts. How a proposition is constructed out of concepts, designators, etc., determines (2a) how we can reason about or with the thoughts of that type, and (2b) when the thoughts are true.
- So I assume that two propositions are the same iff they have the same structure and same constituents. A necessary (but not sufficient) condition for the identity of propositions is that they are true under the same circumstances. This in turn generates a necessary condition on the identity of concepts and other propositional constituents.
- Propositions are "semantical" thought types. Let us say that two thought constituents are of the same semantical type iff interchanging them in a thought would produce thoughts that are tokens of the same propositional type. In this terminology, property representations are the same semantical type iff they are tokens of the same conceptual type. That is, concepts are semantical types for property representations. I will say that a representation expresses the concept that is its semantical type.
- The relationship between (2a) and (2b) is parallel to that between (1) and (3), and can be taken to subsume the latter.
- Arguably, (2a) does not completely determine (2b) unless (2a) takes the form of definitions. Similarly, (1) may not determine (3). It may be that truth and extension can only be understood disquotationally. (Putnam denies this, but see below.)
- One way to argue this is to think of an agent "from the outside". All we observe is the agent's reasoning, and we can form hypotheses about how it works. In particular, we can form hypotheses about how the agent categorizes things (i.e., the reasoning involved). But given that he does not put some things either in or out of a category, why should we think that there is a determinate matter of fact about whether they are in the category? Why think that his cognition involves categories as opposed to just categorization? Of course, he thinks it does, because he can reason disquotationally, but that is just more reasoning.
- Causal theories (Fodor) try to get around this. But the formulation of a causal law requires appeal to either concepts or properties. If a causal law is a proposition, it must reference a property via a concept. (I have also suggested that properties must be cashed out in terms of concepts.) For example, these theories allege that what makes ⌜horse⌝ representations be of the semantical type «horse» is that there is a lawlike regularity to the effect that horses tend to elicit ⌜horse⌝ representations. But aren't horses just things that exemplify the concept «horse»? This seems to make causal theories circular.
- We can give exactly the same argument about truth. Why think that it is a determinate matter of fact whether an agent's beliefs are true? From the inside, the agent can reason to the conclusion that the belief is either true or false, but from the outside, that looks suspicious.

5. Propositions, Concepts and Designators

- Propositions that are predications are constructed out of (i.e., they syntactically reference) concepts and designators. But concepts and designators are individuated somewhat differently, and hence so are the propositions containing them.
- Lexically identical object representations can designate different things at different times without the world changing. This can result, instead, from the cognizer's place in the world changing. The most obvious example of this is the indexical ⌜now⌝,

or for definite descriptions containing an implicit ⌜now⌝. Propositions containing these designators at different times can correspondingly have different truth values without the world changing, so they are different propositions. Assuming that the identity of the proposition is determined by its structure and constituents, because the propositions are different but have the same structure, we must say that the designators are different despite the lexical identity of the representations. So propositional typing is finer grained than syntactical typing.

- The extension of a concept at a time is the set of all things exemplifying it at that time. Concepts can have different extensions at different times, while remaining the same concept. If on two different occasions I have the thought ⌜Electrons are negatively charged⌝ (where this is about all electrons at all times), I am thinking the same thing twice. That is, my thoughts are of the same propositional type despite the fact that there exist different electrons at those two different times. This is because the thoughts are not about electrons at a time — they are about electrons at all times. Clearly, although the extension of my property representation ⌜electron⌝ has changed, that does not automatically require that the thoughts are of different semantical types, i.e., that propositions expressed by the thoughts are different. If the propositions are the same, so are the concepts that are constituents. Thus when the extension changes because the world changes, this need not lead to a difference in the concept or the proposition containing it.
- In philosophical logic, concepts are sometimes identified with their intensions — functions from possible worlds to extensions. However, if a concept can have different extensions in the same possible world at different times, that will not do. One might suggest that we should instead understand the extension to be the set of all objects, throughout the history of the universe, that exemplify the concept. But that won't do either, because an object may exemplify the concept at one time and not at another. If no object exemplifies a concept for all time, but every object exemplifies it at some time, then the concept and its negation would have the same extension on this construal.
- One might suggest that the extension should be viewed as a function from a time to the extension at that time. Let us call this the temporally indexed extension of the concept. This function doesn't change with changes in the world. For formal purposes, one might then take the intension of a concept to be a function from possible worlds to temporally indexed extensions. The "logical role" of concepts as constituents of propositions is to determine what the propositions are about by having a fixed intension that determines an extension at any given time in any given possible world. This intension is supposed to be an essential feature of the concept.
- One could similarly take the temporally indexed designatum of a designator or object representation to be a function from a time to the designatum at that time. So the temporally indexed designatum of ⌜now⌝ is a function from a time to itself. However, unlike property representations and their corresponding concepts, the designator expressed by an object representation can change without the temporally indexed designatum of the representation changing. In particular, the propositional types of thoughts containing ⌜now⌝ change without the temporally indexed designatum of ⌜now⌝ changing. So the representation ⌜now⌝ is unchanged, but the designator «now» must change. Hence we cannot take the logical behavior of the designator to be characterized by an intension defined to be a function from possible worlds to temporally indexed designata. Logically, designators and concepts work differently.
- What about other indexical representations, like ⌜here⌝? Let us take eternal propositions to be those that are not temporally indexed to the present time. An

example would be, «Columbus stood right here in 1492». If I have the thought ⌜Columbus stood right here in 1492⌝ twice, while standing in two different places, one thought may be true and the other false, so they are of different propositional types. Still, the temporal designatum (the function from a time to my location at that time) of ⌜here⌝ (in my thought) is unchanged, so again thoughts that are syntactically identical can be of different propositional types, and this is attributable to the designator. This is another example of a lexical type of object representation that expresses a different designator when its current designatum changes.

- We can reason similarly about de re designators (e.g., «She once stood where Columbus stood» in the twins case).
- Definite descriptions can change designatum over time if they embed an indexical «now». Then the propositions containing them change too, so it follows that the designators expressed by the descriptions (mental representations) change. Again, the representations are the same syntactic type but express different designators (i.e., are of different semantical types).
- One way to think of this is that the intension of a concept assigns a temporally indexed extension to a possible world, but the intension of a designator assigns a single designatum to each possible world. In the latter, variation over time is not allowed.

6. Classical concepts

- Arguably, there are several different kinds of concepts. Some concepts have definitions, and the definitions may determine extensions, truth conditions, etc.
- We can distinguish between syntactically simple and syntactically complex concepts. The latter build in a definition — they "wear their definitions on their sleeves".
- Don't confuse syntactically complex concepts with syntactically complex verbal expressions.
- Even when syntactically simple concepts have definitions, it is not entirely clear how they are connected with the concept. E.g., «triangle» might be defined as «three-cornered plane figure», but most people would instead give the definition «three-sided plane figure». It seems that concepts can become detached from their definitions, and sometimes acquire new definitions. We will encounter this again when we talk about «electron» below. It is a puzzling phenomenon in light of the classical assumption that the definition of a concept is an essential property of it.
- We can think of a property in terms of a definition (i.e., our thought contains the syntax of the definition explicitly), but that is different from thinking of it in terms of a concept "having" that definition. Perhaps definitions are never essential features of syntactically simple concepts.

7. Framework concepts

- Epistemology suggests that there is a core of central concepts that have fixed schemas for reasoning attached to them. They provide the epistemological framework into which other concepts and reasoning fit, and they are plausibly an innate feature of our cognitive architecture. Possible examples include perceptible properties, time, causation, probability, nomic implication, agency, value-theoretic concepts, moral obligation, etc.
- It is crucial to the functioning of these concepts that we can employ them in defeasible reasoning. The schemas for defeasible reasoning seem to be necessary features of the concepts. These cannot be derived from definitions.

- Some concepts are projectible, and others are not. This seems to be an essential feature of the concepts. It is about whether we can attribute the concept to things by employing the statistical syllogism. So this is also about defeasible reasoning in connection with the concept.
- This may be about more than framework concepts. It seems to hold for natural kind concepts (see below) as well. Are all syntactically simple concepts projectible?
- One might plausibly say that framework concepts are individuated by their place in the framework, which is to say, by their functional role in reasoning.
- I argued against this for concepts like «red» on the grounds of rational isomorphism, but I no longer want to count «red» as a framework concept, so I will have to think about that argument further.
- It seems unlikely that the rules for reasoning about framework concepts are sufficient to determine their extensions. They are the basis on which categorization proceeds, but categorization is something like "the pursuit of the extension". It tells us how to reason (defeasibly) about whether things are in the extension, but doesn't determine the extension.
- Do these concepts have extensions? Yes, in a disquotational sense. Concluding that something exemplifies a concept is equivalent to concluding that it is in the extension of the concept. But insofar as there are indeterminate cases of concept exemplification, there are also indeterminate cases of extension membership.
- Putnam talks about "intra-theoretic" and "extra-theoretic" notions of truth and extension, but I am not sure what that distinction comes to. Perhaps the "inside/outside" distinction I alluded to above.

8. Sortal concepts

- These are very general concepts like number, substance, condition, object (in the sense of being a denizen of space-time), action, event, etc. These are also framework concepts, but different from those characterized by defeasible reasoning schemas. Defeasible reason-schemas tend to govern reasoning about instances of a sortal, but they tend to presuppose that it is already determined whether the items reasoned about fall under a particular sortal. For instance, in reasoning about whether the apple is red, we presuppose that the apple is an object and red is a color. We do not employ reason-schemas for reasoning defeasibly to these conclusions. As a general rule, the way we have of knowing of their existence already categorizes them as falling under a sortal.
- Are there any logical reason-schemas unique to a sortal concept that we use in attributing the concept to something? You have to already know what sortal something falls under before you can reason about it in any interesting way. (But not in any way — consider the game 20 Questions.)
- On the other hand, if you know that something is divisible by 2, you can infer that it is a number. But you cannot know that it is divisible by 2 unless either you already know it is a number or you are using reason-schemas not unique to the concept of being divisible by 2 (e.g., you believe it on the basis of testimony).
- Sortal concepts do have built-in defeasible reasoning schemas, but they are concerned with identifying things as the same instance of a sortal rather than determining that something is an instance of a sortal. It seems to be de re necessary that something is of the sortal type it is.
- Perhaps this can be explained by saying that certain kinds of mental representations build in sortal concepts. E.g., you cannot see a number, so a percept cannot represent a number.

9. Natural kinds

- Natural kind concepts like «electron» or «cat» have even less in the way of predetermined logical features than do framework concepts.
- Perhaps there is always a sortal concept that goes with any concept to indicate what kinds of things are potential exemplars. (Jackendoff)
- Otherwise, it seems nothing is essential to a natural kind.
- Consider ⌜horse⌝.
- Consider ⌜electron⌝:

At the end of the 19th century, there was no generally accepted model of the atom. Most physicists believed that the atom was indivisible, although the discovery of radioactivity cast doubt on that in the minds of some physicists. At the same time it was generally believed that electric charge, like mass, was infinitely divisible. For example, James Clerk Maxwell urged that electric charge represented a "strain in the electromagnetic ether". However, there was also some evidence, deriving from Faraday's studies of electrolysis, suggesting that charge might come in indivisible units. (Faraday himself did not believe this.) In 1891, G. Johnston Stoney introduced the term "electron" to describe this smallest unit of negative charge (if there was such a thing). Let us call these "Stoney-electrons", for reasons that will become apparent below. J. J. Thompson proposed that Stoney-electrons were "negative corpuscles detached from atoms". At that time neither he nor anyone else had any clear idea of what the structure of an atom might be or how Stoney-electrons, if there were such things, might be related to atoms.

Although these ideas were familiar by the beginning of the twentieth century, they resisted experimental confirmation until Robert Millikan performed his oil drop experiment in 1909. This experiment was taken to demonstrate that there is a smallest unit of negative charge, i.e., that Stoney-electrons exist. Millikan's experiment consisted of suspending negatively charged oil drops in an electric field. By measuring the strength of the field that was required to suspend a particular oil drop, he could measure the charge on the oil drop. He discovered that all of the oil drops had charges that were integral multiples of 1.6×10^{-19} coulombs, and concluded that Stoney-electrons existed and he was measuring their charge. Millikan received the Nobel Prize in physics for this work.

Let us consider these discoveries from the perspective of contemporary physics. Stoney-electrons were, by definition, the smallest unit of negative charge. According to contemporary particle physics, these exist, but they are not electrons — they are negatively charged quarks, which have 1/3 the charge of an electron. Millikan thought that he was measuring the charge on Stoney-electrons, but in fact, he was not. He was measuring the charge on electrons. He took his experiments to confirm that Stoney-electrons exist. We can suppose that he, and the physicists that followed him, were justified in believing this, and they were right, but they were right by accident because what Millikan was measuring was not Stoney-electrons. In other words, this is a Gettier example. The physicists had justified true belief that Stoney-electrons existed, but it was not knowledge.

Now let us return to our brief survey of the history of atomic physics. Having argued that Stoney-electrons exist, it remained to explain how they are related to atoms. J. J. Thompson proposed the "plum pudding" model according to which an atom is a sphere of uniform positive charge with Stoney-electrons somehow embedded in it. Note that this was just a hypothesis. There was no experimental evidence for the

correctness of this model. The hope was to use this model to explain the patterns of spectral lines emitted from hot gases, but the model was not successful in doing this.

Between 1906 and 1918, Rutherford conducted experiments with the scattering of alpha particles passing through thin metal foils. This led him to propose the "nuclear model" of the atom, according to which the atom contained a very small positively charged nucleus surrounded at some distance by a cloud of Stoney-electrons. The evidence for this model pertained to the size of the nucleus that was required to explain the observed scattering, and did not directly pertain to Stoney-electrons. The Stoney-electrons were just there to balance the positive charge on the nucleus so that the overall charge on the atom was zero. In 1920, Rutherford introduced the term "proton" to refer to the nucleus of the hydrogen atom. The scattering experiments allowed him to compute that the charge on a proton was equal to the charge Millikan had (purportedly) measured for Stoney-electrons, and so he inferred that in a hydrogen atom there was just one proton and one Stoney-electron. This was a change from previous views, which often supposed that atoms contained thousands of electrons.

Of course, all this time people were using the term "electron", not "Stoney-electron", so let us continue in that vein. Thus far physicists believed (1) that electrons exist, i.e., they are the smallest unit of negative charge, and (2) that the atom consisted of a small positively charged nucleus surrounded by a cloud of electrons. The question remained how the nucleus and the electrons were tied together to form the atom. Progress towards answering this question came when Neils Bohr proposed the Bohr atom as the model of atomic structure. On this model, electrons are arranged around the nucleus in discrete shells, and energy is emitted or absorbed when electrons change orbit. The evidence for this theory was the data on spectral emission lines that the plum pudding model tried unsuccessfully to explain.

Subsequent advances in atomic physics led to the discovery of the neutron by Cavendish in 1932, and the development of quantum mechanics by Schrödinger and Heisenberg in the 1930's. At that point all matter was supposed to be composed of electrons, proton, and neutrons — the elementary particles. However, subsequent work led to the discovery of a multitude of additional elementary particles, including pions, positrons, neutrinos, muons, and most recently, quarks and gluons. Quarks and gluons explain the "strong" or "nuclear" force that holds the nucleus together. Negatively charged quarks have a charge 1/3 that of an electron, and positively charged quarks have a charge 2/3 that of a proton. There are also anti-quarks whose charges are reversed. With the advent of the quark, neutrons and protons were downgraded from the status of elementary particles. They are now believed to be composed of quarks.

Now let us reflect upon this history. Something remarkable happened. The term "electron" was originally introduced to mean "Stoney-electron", i.e., "smallest unit of negative charge". But according to contemporary particle physics, electrons are not Stoney-electrons. If we grant that when it was introduced, the term "electron" referred to quarks, and now it refers to electrons, it follows that the extension of the term has changed. How can this happen? It can only happen when a mistake is made and perpetuated by the relevant scientific community. In this case, the mistake happened early on. Millikan thought he was measuring electrons (i.e., Stoney-electrons), but he wasn't. Others accepted his value for the charge on the electron, and upon observing that the hydrogen nucleus has a positive charge of that same value, this led Rutherford to conclude that the hydrogen atom consisted of a single proton and a single electron. Accepting this, subsequent physicists were led to new conclusions about electrons. Eventually a huge body of beliefs about electrons built up and became entrenched.

When Millikan's mistake was eventually discovered, the result was to maintain the bulk of the entrenched beliefs and reject the belief that electrons are Stoney-electrons.

This is a particularly clear and well-documented case of a quite general phenomenon. For another example, consider "atom". Atoms were originally supposed to be the elementary and indivisible constituents of matter. This is the way the term "atom" was originally introduced. But now we believe that atoms have complex structure. As originally defined, "atom" referred to elementary particles, but now it refers to atoms. What this illustrates is that the extension of a scientific term can migrate, including different things at different times, without the world changing. In other words, the temporally indexed extension can change, not just the current extension. Thus far, this is an observation about words. What should we conclude about concepts? If we take the words to express concepts, then we must conclude that either concepts change their temporally indexed extensions, or the words change concepts.

It might be argued that "electron", as used by Millikan, never referred to quarks. It got its extension ostensively as the things Millikan was measuring. However, Millikan's conclusion was not that whatever he was measuring existed. That would be uninteresting. His conclusion was that Stoney-electrons existed. That is what people found remarkable, and that is why they gave him a Nobel Prize for his work. Note also that at that time the term "electron" had no other theoretical baggage, so it could not be claimed that the belief that electrons were the smallest unit of negative charge was just one belief among many and could be given up while retaining the others. The only thing people believed about electrons was that they were the smallest unit of negative charge. Electrons had not yet found a place in the rest of atomic physics, because atomic physics did not yet exist. In fact, all Millikan really discovered was that the charges on oil drops were quantized — they were integral multiples of a single small value. Had he announced this as his conclusion, people would have found it interesting, but they would not have considered it nearly so exciting. He would not have been claiming to have discovered something about the fundamental constitution of matter, and he would most likely not have gotten a Nobel Prize for his work. On the other hand, that more cautious conclusion would have been sufficient for the subsequent development of atomic physics. When Rutherford measured the positive charge on the hydrogen nucleus and discovered it was the same as the value Millikan found on his oil drops, Rutherford would have been in a position to postulate that there was a particle of that charge orbiting the nucleus, without assuming that it was a Stoney-electron (although he might well have mistakenly supposed that). Thus Rutherford would have been the true discoverer of the electron, not Millikan. But things did not work out that way. People thought that Millikan had confirmed the existence of Stoney-electrons, and so Rutherford concluded that it was a Stoney-electron that was orbiting the nucleus of the hydrogen atom. This was a mistake that was perpetuated until the discovery of the quark.

The main lesson I want to draw from this example is that the concept «electron» has no built-in necessary conditions (except, perhaps, the very general sortal concept «object»). If anything were necessary for being an electron, it would be the initial definition «particle with the smallest negative charge». But that turns out to not even be true of electrons, much less necessarily true. So I conclude that natural kind concepts like «electron» have no necessary conditions in the sense of conditions that are simply built into the concept. Let us call these "analytic necessary conditions".

10. Natural Kinds and Ostensive Definitions

One frequently hears the claim that natural kind terms like "electron" are

introduced "ostensively". This is supposed to explain how Millikan could have been talking about electrons rather than quarks. Ostensive definitions are supposed to be analogous to introducing a proper name by pointing to its denotation. What makes the latter possible is that the denotation of the name is a single object, and hence you can literally point to it. But nothing similar is possible for concepts, because unlike the case of designators, you can never get direct access to the entire extension of a concept. For example, you cannot point to all the lions in the world, so how can you define "lion" ostensively? Suppose I point to a single lion, who happens to be large, dark-colored, and male, and say "This is a buto". What does my ostensive definition of "buto" ostend? Animals? Felines? Lions? Male lions? Dark-colored male lions? Large dark-colored male lions? The word might be intended to pick out any of these. Ostensive definitions of natural kinds make no sense. We could introduce "buto" by saying it refers to animals that are the same species as some specified animal (or the same genus, or the same family), but that is a definition in the ordinary sense, not an ostensive definition. Similarly, we could introduce it by saying that it refers to animals that are the same species and color (and size, gender, etc.) as some specified lion. But all of this involves more than ostension.

On the other hand, although you cannot define a natural kind by simply pointing to an object and saying "that kind", there is something a bit like this that does go on and needs explanation. Suppose I see an unfamiliar animal and note some of its characteristics. Later, I see another similar animal. I may judge spontaneously, "That is the same kind of animal". If I later see some more that I also judge to be the same kind, I have acquired a new natural kind concept, and I can go on to investigate further properties of animals of this kind. For instance, something like this happens when an entomologist discovers a new kind of insect. This is a case in which it seems I really could introduce "buto" by saying something like, "I have seen several animals like that — let's call them butos," without inviting the question, "Which kind are you talking about?" It seems that we naturally type animals along certain dimensions, and naturally generalize about such types (an "inductive bias"). This is a matter that requires further discussion. (See Pollock and Oved 2004.) Notice, however, that although this sometimes works for animals, it would not work for electrons.[9]

11. How do we reason with natural kind representations?

[9] One might suppose that we are really just trading on a prior concept of species, and implicitly defining the kind to consist of animals of the same species as the one we observed. To some extent this may be right, although children can do this without being familiar with the word "species". A further difficulty for this kind of account is that the concept of a species is not all that well behaved. In particular, it is a philosopher's fiction that species can be characterized in terms of DNA. No one knows how to do that. For example, biologists argue about how many species of elephant there are. The answers range from 2 to 11. This is not a dispute that they would regard as resolved by an appeal to DNA. The problem is that there is a great deal of variation in the DNA of different members of the same species. For example, there is no particlar string of DNA that is necessary and sufficient for being human. It is possible that there might be some disjunctive set of strings of DNA such that an animal is human iff its DNA includes one of those strings, but no one has any idea what such a set of strings might be, and at this point there isn't even a good reason for thinking there is one.

If «electron» had the definition «smallest negatively charged particle», it would be impossible to discover that some quarks have 1/3 the charge of an electron. So although we may have originally introduced the term "electron" in that way, this was not a definition of a concept. It seems that a natural kind like «electron» has no essential features. But then what governs or licenses the inferences we make about electrons?

Consider another example. A young child living on a small island encounters a family of skunks living in the woods, and befriends them. She creeps into the woods without her parent's knowledge, and the skunks rub against her legs for attention, curl up in her lap as she pets them, etc. She judges them to all be the same kind of animal, and she comes to think of this kind of animal in a certain way. That is, she acquires a natural kind concept (or at least, representation). But she has no word for it. For the sake of neutrality, let us say that her representation is the ⌜furry-friend⌝ representation. It is a quite specific lexical type of representation and it represents the natural kind property skunk. Then she moves with her parents to the mainland, leaving her furry friends behind. On the mainland she encounters cats, and mistakenly thinks they are the same kind of animal.[10] She goes on to learn all the usual things about cats, and takes what she learns to be true of furry-friends in general. In particular, she comes to call them "cats". As time passes, she forgets entirely about her childhood furry-friends on the island. Years later she returns to the island, encounters the descendants of her skunk family (without remembering them from before), and has the thought that they are not furry-friends. At that point, it seems clear that her furry-friend representation represents cats, not skunks. But originally it represented skunks.

This example illustrates that the behavior of the representation ⌜electron⌝ is not unusual. Like ⌜electron⌝, the representation ⌜furry-friend⌝ comes over time to pick out different objects. Its (temporally indexed) extension changes. This is a remark about mental representations, not concepts. Are there also concepts «electron» and «furry-friend» that change extension? We must either say that or say that the representations have come to express different concepts. It is not initially obvious which to say.

In talking about the representations ⌜electron⌝ and ⌜furry-friend⌝, we are not talking about individual token representations. Rather, we are talking about lexical types. There is a common lexical type of representation that is being used over an extended period of time to think of furry-friends or electrons. That much remains constant. What is at issue is whether there is a fixed concept that is expressed by all instances of that lexical type. If so, it must be possible for concepts to change extension. Alternatively, the lexical types could have become attached to different concepts.

It is at least peculiar to suppose concepts can have changing temporally indexed extensions. We have to ask what the semantical typing of representations is supposed to be for, and the standard view is that it picks out an intension, which assigns a temporally indexed extension to each possible world and thereby determines what things a proposition containing the concept is about. If we detach concepts from intensions in this sense, one is left wondering what the point of semantical typing is. This argument seems compelling, but as we will see, it is hard to reconcile with the way we actually employ natural kind representations in reasoning.

Consider the examples of ⌜electron⌝ or ⌜furry-friend⌝. It is not too difficult to describe how the reasoning goes that seems to lead to a change in extension. At any

[10] Alternatively, we might suppose that her ⌜furry-friend⌝ representation includes cats. This makes no difference to the example as long as we agree that it also includes skunks.

given time, we have lots of beliefs about objects of a particular natural kind, and we form new beliefs by reasoning in perfectly normal ways from the old ones. We have to start out thinking of objects of that kind in some particular way. This leads to the initiation of a new lexical type, which is the particular natural kind representation we are using (e.g., ⌜electron⌝ or ⌜furry-friend⌝). However, if our reasoning were perfectly standard, we would never get a recognized change in extension. If we discovered that we were attributing the concept to things it was not originally attributable to and refraining from attributing it to things it was originally attributed to, that would indicate that something has gone wrong and it would lead to the defeat of the beliefs responsible for these changed attributions. But in fact it doesn't. Reasoning about natural kinds does not work that way.

What actually seems to happen is that contingent beliefs about natural kinds "become entrenched", and when they do they become detached from the basis on which they were originally held. It is a bit like forgetting the basis, except that if we are scientists we may record all our reasoning and we don't really forget anything. Rather, entrenched beliefs acquire their own inertia, and can lead to the overturning of the beliefs we originally employed in thinking about the natural kind and arriving at the new beliefs. This is a kind of cognitive boot-strapping phenomenon.

Beliefs alone are not the whole story. There can also be contingent means of recognizing instances of a kind (e.g., cat-detectors) that do not function via explicit beliefs. They too can become entrenched and play a fundamental role in reasoning about the kind by playing a role in identifying instances of the kind. These recognitional capabilities can, in some cases, provide the starting point for a natural kind representation. This is illustrated by the ⌜furry-friends⌝ example. But although this might work for animal kinds, it does not work for electrons, because electrons are not individually recognized as such.

- This seems to be about logical reasons for identifying (grouping) instances of a sortal.
- Are some of the beliefs more central than others, and correspondingly more resistant to change? Some of them will have stronger degrees of justification than others, and that will automatically make them more resistant to change. We seem to get the right structure for the reasoning if we just regard all the entrenched beliefs and recognitions as prima facie justified to varying degrees.
- This makes sense of Burge's arthritis example. You may start out thinking of a kind like arthritis in terms of a public language word. Arthritis is what people are talking about when they use the word "arthritis" in certain ways. On this basis you form various entrenched beliefs about arthritis. In particular, I may believe that I can have arthritis in my muscle, but I believe more firmly that my doctor knows more about arthritis than I do, so when he tells me that is impossible, I believe him and reject my belief that I can have arthritis in my muscle. This illustrates that Putnam's division of linguistic labor is an automatic concomitant of the logic of natural kinds.
- Putnam defines a narrow psychological state to be one such that your being in it does not entail that anything else exists. He then gives the twin earth argument to conclude that knowing the meaning of a word is not a narrow psychological state. It might also be used to argue that having a natural kind concept is not a narrow psychological state, although I will question whether this argument works. (It turns out whether the kinds have the same temporally indexed extensions now as they did before the rise of chemistry.) However, even if this is a bad argument, the conclusion is clearly true. In general, having a belief, i.e., a thought of a particular propositional type, cannot be a narrow state. This is illustrated by the indexicals here, now, up, down, left,

right, and also by percepts. The syntactically same thought can be of different propositional types (and have different truth values) depending on your situation in the world.

- Still, this picture is inconsistent with Putnam's account of natural kinds.
- Putnam claims that natural kinds have ostensive definitions. I argued above that this makes no sense.
- In the twin earth example, Putnam claims that the extension of water in 1750 is the same as that today, on the grounds that scientific advance can only refine our knowledge of water, without changing extensions. Scientific advance includes finding new instances of a kind and investigating their properties. So suppose a worm hole opened between earth and twin earth in 1751 and remains open today, allowing people to travel freely between the two planets. A few years after that we would all have been calling both XYZ and H_20 water. With the rise of chemistry, we would have concluded that there were two kinds of water (like the two kinds of jade). Would we have been wrong? Surely not. Did the extension of water change when the worm hole opened? That seems very peculiar. (Nobody's theory would countenance that.) Thus if scientific advance cannot change extensions, then in this example the extension of water in 1750 includes XYZ. But the extension of water in 1750 in the example is surely the same as the extension of water in 1750 in the real world, so by Putnam's principle, it follows that the current extension of water includes XYZ. But it is universally agreed that it does not. Putnam's principle must be wrong.

12. Concept stability

- If we insist that a natural kind representation comes to express a different concept when its temporally indexed extension changes, we must ask under what precise conditions such a change occurs.
- It is not sufficient to take changes in the temporally indexed extension to be a necessary and sufficient condition for concept change, because the temporally indexed extension is not well determined. There is not, in general, any way to pick it out except in terms of the concept itself.
- It looks like we will be forced to say that the concept changes whenever the temporally indexed extension could change, and it looks like that happens whenever our entrenched beliefs change. We can make this argument precise as follows. For a concept C, let E_t be the temporally indexed extension of C at time t (i.e., the function from times to extensions, not the current extension of C at t). I have argued that this cannot change, i.e.,

(1) $\Box(E_t = E_{t^*})$

Let B_t be the set of all entrenched beliefs for C at time t. If B_t changes, then the temporally indexed extensions could change (and would change if the world behaved in certain ways), i.e.,

(2) $\Box[B_t \neq B_{t^*} \rightarrow \Diamond(E_t \neq E_{t^*})]$.

From (2) it follows by modal contraposition that

(3) $\Box[\Box(E_t = E_{t^*}) \rightarrow B_t = B_{t^*}]$.

From (1) it follows (I assume that logical necessity satisfied S4) that

(4) $\Box\Box(E_t = E_{t^*})$,

and from (3) and (4), by the principle that $\Box P, \Box(P \rightarrow Q) \vdash \Box Q$, we get

(5) $\Box(B_t = B_{t^*})$,

i.e., the set of entrenched beliefs for C cannot change. This makes natural kind concepts horribly unstable. They must change whenever their entrenched beliefs change.

- Here is a bad argument for thinking this conclusion must be wrong (Putnam gives this argument):
- If having different entrenched beliefs requires cognizers to have different concepts, then different cognizers will hardly ever have the same concepts, but then they would not be able to use the concepts in linguistic communication.
- Answer: language does not convey propositions. It elicits thoughts in the hearer that are "appropriately related" to those of the speaker, but the relationship is much coarser grained than "same propositional type".
- Here is a more compelling argument for thinking that this conclusion must be wrong:
- The entrenched beliefs are stored in long-term memory. They are not simply held in working memory. As such, they must be retrieved. What is retrieved is not literally the same thought token. It is a different token of the same syntactical type. But what is crucial here is that the retrieved beliefs must still be about the original concept in order to count as being remembered. But they must also employ the current concept if they are to play a role in its individuation. So the current concept must be the same as the original concept, i.e., the concept cannot have changed.
- This suggests that all that is required for concept stability is lexical sameness. There does not seem to be any intermediate position between this and saying that the concept changes every time the entrenched beliefs do.
- But consider the counter-argument:
- If the entrenched propositions change in a way that would lead to different attributions, even given perfect knowledge of all the relevant considerations, it is plausible to say that the temporally indexed extension has changed.
- I don't have an argument for that. All I can really conclude is that we would be warranted in believing that the temporally indexed extension has changed. But then we will be warranted in believing that some attributions have changed truth value, and this is without the world changing.
- This seems to imply that the concepts have changed identity. In particular, some of the attribution propositions have changed truth value, and it seems to follow that some of the entrenched propositions used in deriving these attributions have changed truth value, contrary to the first argument.

13. Memory Retrieval

As formulated, the memory argument turns on a premise that we have already rejected, viz., that what you remember must be the same proposition as the one your originally stored. We saw that there are a number of counterexamples to this principle involving change over time. E.g., I can remember my daughter's age, or that

something occurred earlier.

- Memory is not just the simple retrieval of previously stored thoughts (tokens). Thought tokens occur when you have them. As such, the same thought token cannot occur twice, any more than you can write the same sentence token on the board twice. That is part of what we mean by a "token". Rather, memory is "productive", in the sense that memory retrieval must result in the production of a new thought token appropriately related to the stored thought token. I.e., it must result in the production of a thought of an appropriate type.
- The preceding examples illustrate that the retrieved thought need not be syntactically identical to the stored thought. Retrieval can involve a computation that produces a thought of a different syntactical type.
- More important for present purposes, the propositional types can be different too.
- Here is another form of memory retrieval that does not preserve propositional type.
- We saw the importance of temporal projection.
- It would be computationally awkward, to say the least, to require a cognitive agent to always make an inference by temporal projection whenever it wants to reuse information previously obtained about an earlier time. Instead, we employ the mental temporal indexical "now" (linguistically it is often dropped) and simply retrieve the belief «P is true now» from the previously stored belief «P is true now». E.g., I remember that my car is white. These are the same syntactical type, but different propositional types. This is a kind of implicit temporal projection, which eases the computational burden on the agent.
- These examples suggest that the memory argument is wrong. Memory regarding natural kinds may work like implicit temporal projection. We retrieve a thought of the same syntactical type, regardless of whether the concept has changed, and let the propositional type fall where it may.

14. Penultimate Conclusions

- It is not clear that propositional identity across time has any use. Concept identity only seems to have a use in determining propositional identity, so concept identity across time also seems to be without a clear use. Neither propositional identity nor concept identity seem to be required for understanding cognition. That works in terms of representations and their lexical and grammatical types rather than concepts (i.e., semantical types).
- There are different kinds of concepts, and correspondingly different types of property representations.
- Syntactically complex ones, whose syntax sometimes encodes definitions.
- Framework concepts that have predetermined defeasible reasoning schemas. These do not change, and so concept identity does not seem to be an issue.
- Natural kind concepts. These are the ones that have caused most of the trouble for theories of concepts.
- It is fairly easy to describe how reasoning works with natural kind representations.
- But identity conditions across time for natural kind concepts is problematic. We seem to be forced to say that the concept changes every time the entrenched beliefs change.
- This has the consequence that the concepts hardly ever remain the same, and there is no reason to think that different cognizers will have the same natural kind

concepts.

- But this does not hurt anything, because identity over time does not seem to have any use anyway. (In particular, it cannot be what is required for linguistic communication.)
- Perhaps what is important is to be able to say that certain syntactically identical thoughts are not propositionally identical. This is to block certain kinds of reasoning. For instance, I cannot preserve an occurrent belief involving ⌜now⌝ or ⌜here⌝ in working memory just by rehearsing it. However, this is handled by knowing the grammatical types of the property representations. We need not know the semantical types.
- Perhaps what we should conclude is that if concepts are required to work the way philosophers have traditionally supposed they do (i.e., propositional constituents with fixed temporally indexed extensions), then there aren't any, but it is a myth to suppose that such things play a role in understanding cognition anyway. This probably stems from Frege's theory of language.
- What we really have are various grammatical categories of property representations, and different processing rules for each.
- We acquire new lexical types of property representations for the "natural kind" grammatical category, and when we do we talk about "learning a new concept". In this sense, to learn a concept is to acquire the capacity to have thoughts of a propositional type containing the concept. For natural kinds, this amounts to having an appropriate set of entrenched beliefs. But it does not, in any interesting sense, establish a relationship to a "philosophical concept". What we are really acquiring is a new lexical type of a particular grammatical category, not a new semantical type.
- There may be other uses of the term "concept" that are more useful. To be useful, concepts must be shareable and accessible over time. This is also true for whatever is involved in linguistic communication, so perhaps that is the best place to look for a useful notion of "concept". (Note my use of "notion" in the last sentence. What is that?)

15. Concepts, Social Cognition, and Language

- Apparently we do not need concepts and propositions for understanding first-person cognition. They play no useful role in describing what is going on.
- But things get more complicated when we consider how we understand the cognition of others. For instance, I may believe that Jones thinks that if he had a horse, he could ride it to Newcastle, and I believe he wants to get to Newcastle. I also believe that Jones thinks that Nelly is a horse. So I conclude that Jones would like to have Nelly.
- The form of the reasoning might derive from simulation, but where do I get the beliefs that instantiate it. I.e., how do I know that Jones has appropriate beliefs and desires? In particular, how do I know that Jones believes that Nelly is a horse, and how do I know that Jones believes that if he had a horse, he could ride it to Newcastle? I am attributing thoughts to him involving a mental representation ⌜horse⌝. For the reasoning to work, it is not actually important what mental representation that is, as long as it is the same in both thoughts. However, we don't simply believe that there is a representation x such that Jones thinks ⌜Nelly is an x⌝ and Jones thinks ⌜If I had an x I could ride it to Newcastle⌝. I have separate reasons for believing that Jones thinks Nelly is a horse and for believing that Jones thinks that if he had a horse he could ride it to Newcastle. Then I combine the two beliefs in my reasoning. So I must have some fixed way of thinking of horses that I can use in attributing thoughts to Jones. This seems to be a role for concepts.

• This is probably related to linguistic communication. It is to the extent that Jones and I share the concept «horse» that we can communicate about horses. We just cannot assume that these concepts we share and use in communicating and in attributing beliefs to others are "philosophical concepts", i.e., have fixed temporally indexed extensions and are constituents of propositions, where propositions have fixed truth conditions. On the other hand, if I don't think that Jones is thinking about the same things I am as "horses", what am I attributing to him when I believe he thinks Nelly is a horse? This is just the distinction between P-concepts and L-concepts that I mentioned above.

• Perhaps language plays a crucial role here? Language socializes cognition.

• What is my reason for thinking that Jones thinks Nelly is a horse?

• It cannot be simulation. I do not believe that Jones has the same natural kind concept «horse» as I do, because I do not believe he has the same entrenched beliefs.

• I probably believe that he thinks Nelly is a horse because anyone would under these circumstances — Nelly looks and acts like a horse, and he is in a position to be able to see all this. But to know that anyone would think that Nelly is a horse under these circumstances, it seems I must be able to generalize from similar judgments about Particular individuals, and those judgments must be made more directly. So it must be possible for me to know that Jones thinks Nelly is a horse without basing this on a generalization about others.

• I might know that Jones thinks Nelly is a horse because he says so. But how do I know that what he calls "horses" are the same things I do? I know this because I know that he is a speaker of English, and I know that in English, something is properly called "horse" iff it is a horse. We must ask how one can know this, but let us set that aside for the moment. Assuming that we can know such things, how can we use them to acquire knowledge about the beliefs of others?

• Where W is a word (or expression) of language and C is a natural kind concept for a person S, let us call S's belief «Something is properly called W iff it is a C» an L-anchoring belief.

• Note that this has to be formulated in terms of concepts C rather than representations, because these are stored beliefs, not occurrent thoughts. If a belief is not occurrent, there is no representation (token).

S's L-anchoring belief relates C to W. Let us say that W anchors C for S. If two different people both have L-anchoring beliefs about natural kind concepts, this establishes a relationship between their concepts. If their beliefs are both true, then their concepts have the same temporally indexed extension. Let us say that two people's concepts are L-equivalent iff they are anchored by the same word or expression.

• What is novel about natural kind concepts is that having an L-anchoring belief tends to make the belief true. If I have an L-anchoring belief by virtue of which W anchors C for me, then I will conclude that something is C when I believe that it is properly called W, and conversely if I believe that something is C I will conclude that it is properly called W. This enables me to communicate with other speakers about natural kinds. I can convey information to them, which they can interpret in terms of their own concepts, and I can get information from them and interpret it in terms of my own concepts. This is Putnam's "distribution of linguistic labor", and it is what makes language possible.

• Having my concept anchored by an expression of public language has an important stabilizing effect. It means that my beliefs are criticizable by others, and so helps prevent me from acquiring new entrenched beliefs that change the temporally indexed extension of a concept. Language makes the extensions of our concepts more

stable, because they are determined intersubjectively.

- Do concepts ever get detached from their linguistic anchors? Yes, it certainly happens in scientific contexts, although this tends to spawn a new sense of the original word, e.g., the scientific use of "fish" or "heat" or "water".
- How do different senses of a word get factored into this account? When words are ambiguous, we can normally differentiate the senses linguistically. E.g., "bank that deals with money" vs. "riverbank". Being able to make such a distinction does not automatically generate ambiguity. We don't want to say that "apple" is ambiguous because there are red apples and green apples. But I don't really need an analysis of ambiguity. L-anchoring beliefs can be about expressions, not just words.
- Anchoring beliefs provide a vehicle for knowing that Jones believes that Nelly is a horse. What I might believe is that Jones has a concept C that is L-equivalent to my concept «horse» and a designator N designating Nelly such that he believes «N is a C».
- There is a potential problem for this account. It requires that Jones and I must share a language in order for me to know what he believes.
- If Jones and I both have language, but possibly different languages, we can still relate our thoughts by appealing to chains of bilinguals. For this purpose we simply identify L-equivalence with its transitive closure.
- But even if there is such a chain of bilinguals, I may not know it. Or what if Jones speaks no language? Is it then impossible for me to know what he believes? Compare this to attributing beliefs to nonlinguistic animals. Can I really know that a chimp believes Nelly is a horse? Maybe not, but I am pretty sure that a Tibetan will have such a belief, even though I have no idea what his word for "horse" is.
- We might say that two people's concepts are potentially L-equivalent iff there could be a word or expression making them L-equivalent, i.e., iff they could share a language relative to which the concepts are L-equivalent. If the L-anchoring beliefs must be true, this seems to hold iff the concepts have the same temporally indexed extension. Looking at it functionally, it seems that this is possible iff they would ascribe their concepts to more or less the same things in the same circumstances.
- Thus far I have ignored a crucial problem. How do people become justified in holding L-anchoring beliefs? L-anchoring beliefs have the form «Something is properly called W iff it is a C».
- What does it mean to say that something is properly called "horse"? This might be taken to mean that it is correctly called "horse". But to know this it seems we would have to know the temporally indexed extension of "horse". Presumably, the temporally indexed extension of "horse" is determined by the temporally indexed extensions of concepts anchored by "horse" in different speakers. But then to know that Jones and I mean the same thing by "horse", it seems I would have to already know that our concepts have the same temporally indexed extensions. This would make knowledge of language impossible without a prior solution to the problem of how we attribute beliefs to others. But my suggestion is that we do the latter in terms of the former. So we need an alternative account of L-anchoring beliefs.
- I am proposing that language can provide a connection between my natural kind concept «horse» and Jones' concept. We both take our concepts to be "expressed by" the English word "horse". What does this come to? When we hear people using the word "horse" we believe ⌜They are talking about horses⌝. But what does it mean to say they are talking about horses? That seems to mean that they are using the word with that temporally indexed extension, but this would make the account circular. Perhaps we believe that if someone says of something "That is a horse", it is probably a horse.

- What do children learn when they are learning language? Initially, they learn correlations between certain behaviors (utterances) involving "horse" and things being horses. The nature of natural kinds is such that we can then use the utterances of others to correct our own concept attributions. But it works the other way too. We can correct the statements of others. In fact, the connection between my concept «horse» and the English word "horse" is quite tight. If the speaker and I disagree, I assume that one of us is wrong. It is a general feature of language that if one person attributes "horse" to something and another one denies it, we assume both that one of them has said something false, and that he has a false thought.
- We discover that others sometimes say "horse" intentionally in order to help us have true beliefs. We can then reciprocate. Primates can learn these things too, and learn to behave in this way. This need not be knowledge of language, in the sense linguists have in mind. It is just knowledge of signing or indicating (Grice's "natural meaning").
- The belief that others tend to say "horse" in the presence of horses is only a rough statistical generalization. To turn it into an exceptionless generalization, we form the belief that something is properly called "horse" iff it is a horse. This is the L-anchoring belief. We can abbreviate «Something is properly called W iff it is a C» as «C is expressed by W». (This has to be relativized to a language or linguistic community. What does that involve?)
- The concept of a concept being expressed by a word or expression cannot be a natural kind concept. If it were, we would have to have an initial way of thinking about it that is not a natural kind, and then add generalizations to our stock of entrenched beliefs. To have an initial way of thinking about it, we would have to be able to at least roughly define it in terms of other concepts we already have, but there does not seem to be any way to do this. It seems to follow that it must be a framework concept, and innate.
- As a framework concept, we can have built in defeasible reason schemas for relating words and concepts in this way. I suggest that all we can discover without using this concept is that certain types of auditory acts correlate with the presence of instances of a concept. This correlation can never be terribly strong, because, for example, people do not always comment on the presence of horses. But we can strengthen the correlation in various ways. When we are teaching the word "horse" to a child, we make sure they are attending to the horse, and try to ensure that they can see that we are attending to it. These are skills that are built in. Children are naturally able to track the direction of an adult's gaze. In effect, children are predisposed in various ways to learn language. In particular, they are predisposed to learn connections between words and concepts, and these enable them to conclude defeasibly that the concepts are expressed by the words.
- It seems that children must be able to recognize that a person is saying something. They must have an innate ability to recognize speech acts. A person's saying something must be treated as a defeasible reason for thinking it is a proper utterance, i.e., that they are using words to talk about things that are properly called by those words.
- That something linguistic is innate will not surprise most linguists and psycholinguists. But Chomsky's learnability argument is about efficiency. This is instead about the possibility of learning language. I am claiming that, as a matter of logical necessity, you could not have language without this.
- The upshot of all this is that the ability to ascribe beliefs to others goes hand in hand with the ability to learn language. You could not do either without the other, and both

require some undefinable framework concepts and defeasible reasons regarding them.

16. Social Knowledge

- Are the linguistic utterances of others a primitive source of knowledge, not reducible to simple inductive reasoning? Or can we simply discover inductively that others generally speak truly, and use that as evidence?
- We cannot discover that others generally speak truly without knowing what they are saying. And we could not discover what they are saying if most of their utterances were not proper utterances.
- This isn't quite right. We could fool a child into thinking a concept was satisfied (e.g., virtual reality presentations of earth animals to children raised on Mars) and teach her a word for the concept in those contrived circumstances. Or we can do the same thing with pictures of animals, etc. I learned "tiger" as a child without seeing any real tigers. In that case, though, we are applying the word to what the picture is a picture of, so I guess this is not a counterexample. On the other hand, we could use drawings of imaginary tigers. Still, the imaginary tiger is a tiger.
- There is something inductive going on. Children are learning correlations between utterances of "tiger" and the presence of tigers (surely not!!), but this is different from learning that people generally tell the truth when they say that something is a tiger.