Minds without Meanings

An Essay on The Content of Concepts

Jerry A Fodor
Zenon W Pylyshyn
Rutgers Center for Cognitive Science

Chapters

Epigraph

If you slip….
Pick yourself up
Brush yourself off
And start all over Again.

Jerome Kern and Dorothy Fields

Minds without Meanings

An Essay on The Content of Concepts

Book Draft, March 12, 2013

Chapter 1 Introduction: Working Assumptions

Most of this book is a defense and elaboration of a galaxy of related theses which, as far as we can tell, no one but us believes. There are various ways to formulate these theses: That tokens of beliefs, desires and the like are tokens of relations between minds and mental representations; that mental representations are `discursive' (which is to say, language-like); that reference is the only semantic property of metal or linguistic representations[1] that there are no such things as word meanings or conceptual contents; that there are no such things as senses… This list is not exhaustive; we'll add to it as we go along. And we'll argue that these claims, if true, have profound implications for cognitive science, linguistics, psychology, the philosophy of language and the philosophy of mind; all of which are, according to us, long overdue for massive revisions. We'll sketch accounts of mental representations and processes that embody such revisions and are, we think, compatible with a variety of empirical data.

We don't, however, propose to start from first principles. To the contrary, we will make a variety of quite substantive assumptions which, though they are by no means universally endorsed, are at least less scandalous than the ones that we primarily propose to defend. We start by enumerating several of these. In chapters to follow, we will add more and consider some of their implications.

First assumption: *Belief/desire psychology*

To begin with, we take it that behaviorism is false root and branch; in the paradigm cases, behavior is the effect of mental causes, and the paradigm of explanation in cognitive psychology is the attribution of a creature's actions to its beliefs, intentions, desires and other of its `propositional attitudes', which are themselves typically the effects of interactions between its innate endowments and such mental processes as perceiving, remembering and thinking. Likewise, though we assume that the mechanisms by which mental causation is implemented are, in all likelihood, neural, we don't at all suppose that psychological explanations can be reduced to, or replaced by, explanations in brain science; no more than we suppose that geological explanations can be reduced to, or replaced by, explanations in quantum mechanics. Confusions of ontological issues about what mental phenomena are with epistemological issues

[1] More precisely, reference and truth. But we'll have practically nothing to say about the latter. We suppose that, given a satisfactory notion of reference, the notion of truth can be introduced in more or less the familiar Tarskian manner, the basic idea of which is that if the symbol `a' refers to (the individual) *a,* and the symbol `F' refers to (the property) F, then the symbol `a is F' is true if and only f a is F. There are, to be sure, lots of hard problems in this part of the wood. But we thought we'd leave them for the logicians to work out.

about how mental phenomena are to be explained have plagued interdisciplinary discussions of how ---or whether--- psychology is to be `grounded' in brain science. Current fashion prefers `central state' reductionism to the behaviorist kind, but we think, and will assume, that the prospects for both are equally dim, and for much the same reasons. `Naïve Realism' is the default assumption in the psychology of cognition, just as it is everywhere else in science. Otherwise why does it so often turn out that the predictions that the theories endorse turn out to be true?

The paradigms of belief/desire explanation, to which we will advert from time to time, is what Aristotle called a `practical syllogism'.

Practical Syllogism:

- **A** wants it to be the case that **P**
- **A** believes that not-P unless **Q**
- **A** acts so as to bring it about that **Q**.

We take it that explanations that invoke practical syllogisms are typically causal (beliefs and desires *cause* actions). But we think it was a mistake for Aristotle to hold that the `conclusion' of a practical syllogism is an action. For one thing, what syllogisms are supposed to preserve is *truth,* and actions are neither true nor false. Also, intentions to perform actions are often thwarted by facts that the agent failed to include in his reasoning (you try to scratch your nose, but somebody stops you) .

Readers who wish think such objections are mere quibbles are, however, entirely free to do so; we think that the main point of Aristotle's story is perfectly sound: Typical explanations of creatures' behaviors take them to be the effect of mental causes; and so shall we. Also, we take for granted that which of a creature's behaviors are correctly explained by reference to its beliefs and desires is a fully empirical question; as is the question which of the behaviors of such creatures constitute actions. Such issues can't be settled from the arm-chair, much philosophy to the contrary notwithstanding.

Second assumption: *Naturalism*[2]

Mental states and processes are part of the physical world. That means, at a minimum, that the processes that cognitive science postulates must be ones that can be carried out by actual physical mechanisms, and the states that it postulates are ones that physical objects can be in.

But it is sometimes claimed that naturalism is inconsistent with the kind of cognitive psychology that takes believing, intending and the like to be content-bearing states that are also bona fide causes of behavior. Propositional attitudes are relations that creatures bear to abstract objects; indeed they are abstract objects (as are numbers, properties and the like). And abstract objects can't be either causes or effects. The number three can't make anything happen, nor can it be an effect of something's having happened (though, of course, a state of affairs that instantiates threeness, ---for example, there being three bananas on the shelf--- can perfectly well be a cause of John's looking for the bananas there or an effect of his having put the bananas there.)

Likewise, propositions can't be causes or effects; the proposition that it is raining can't cause John to bring his umbrella; it can't even cause John to believe that it's raining. But then, if

---

[2] It has been suggested that this construal of Naturalism is vacuous because which states and processes count as `physical' keeps changing as physics advances. (See Chomsky REFERENCE). But we don't think this objection is sustainable. The point is that it turns out, so far at least, that physics, and physics only, is basic science: Everything that enters into causal processes, and every causal process that anything enters into, has a true description in the vocabulary of physics (but not in, as it might be, the vocabulary of geology or botany or cognitive psychology.) The discovery that any science is basic in this sense is a major achievement of the scientific enterprise; one to which a viable cognitive science is prima facie required to conform. This is clearly the view that scientists themselves hold: If it turns out that our best cognitive science (or, for that matter, our best neuroscience or our best geology) turned out to be incompatible with our best physics, general panic would ensue. Hilary Putnam (reference) rejects the naturalist's program in psychology, but for a different reason than Chomsky's; viz that "as a rule, `naturalism is not defined (110)" In consequence, he accepts Brentano's thesis of the "baseness of intentional idioms and the emptiness of a science of intentional idioms. (14);" Putnam is certainly right that no one has defined `naturalism' `believes that' `means that' or, indeed, any of the key methodological or theoretical terms that intentional psychology has resort to. But we wonder why he thinks that shows that intentional idiom is baseless' or, indeed, anything else of much interest. In fact, neither the theoretical vocabulary of empirical sciences, nor the vocabulary in which methodological constraints on empirical sciences are formulated. is hardly *ever* defined (or, we expect, ever will be), *The notion of definition plays no significant role in either science or the philosophy of science* as Putnam has himself often, and illuminatingly insisted. No botanist has defined `elm tree', nor has any chemist defined `water'. Rather, what botany , chemistry, the other non-formal sciences seek is a sort of empirical metaphysics. *Water is H20*' doesn't *define* water; it says *what water is; what makes something* water. Likewise, empirical theories in the sciences often advert to such undefined notions like `observation'; `confirmation' `data', `evidence' and, for that matter\ 'empirical' and `theory'). That is *not* a reason to suppose that empirical theories in science are ipso facto empty. Even sophisticated people used to say that science consists of `observations and definitions'. But that was a long while ago, and sophisticated people don't say that any more. In particular, Putnam himself doesn't, *except when he's talking about intentional science.* Doesn't that seem a bit arbitrary?

propositions can't be causes or effects, and if propositional attitudes are relations that creatures bear to propositions, mustn't propositional attitudes themselves be likewise causally inert? It looks as though we must have been wrong either about propositional attitudes being causes of behavior, or about their being states that have propositional contents. After all, the difference between propositional attitudes is, often enough, a difference between the propositions that they are attitudes towards: The difference between the propositions *Venus is red* and *Mars is red* is, quite plausibly, all that distinguishes John's believing that Venus is red from his believing that Mars is. We want propositional attitudes to be causes of behavior; but naturalism wants propositions not to be causes of anything; so perhaps we can't have what we want. It wouldn't be the first time.

This is a metaphysical minefield and has been at least since Plato; one in which we don't intend to wander. We will simply take for granted that abstracta are without causal powers; only `things in the world' (including, in particular, individual states and events) can have causes or effects. The question is how to reconcile taking all that for granted with the naturalism of explanations In cognitive science, More presently.

Third assumption: *The type/token distinction.*

It helps the exposition, here and further on, if we introduce the `type/token' distinction: If one writes `this cat has no tail' three times, one has written three *tokens* of the same sentence *type.* Likewise, if one utters ` this cat has no tail' three times. Propositions, by contrast, are *types* of which there may (or may not) be tokens either in language or (according to the kind of cognitive science we endorse) in thought. Propositions types are causally inert; but the tokens that express them --- including chalk marks on blackboards, ink marks on papers, utterances of sentences and so forth---- are bona fide physical objects. The `Representational Theory of Mind' proposes to extend this sort of analysis to thoughts, beliefs, and other contentful mental states and events that play roles in the causation of cognitive phenomena: Mental causes that express propositions (or whatever units of content propositional attitude explanations may require) are, by stipulation, `mental representations'. It is convenient to suppose that mental representations are neural entities of some sort, but a naturalist doesn't have to assume that if he doesn't want to. We are officially neutral; all we insist on is that, whatever else they are or aren't, mental representations are the sorts of things that physics talks about. There may be a better way out of the puzzle about how mental contents can have causal roles, but we don't know of any.

Fourth assumption: *psychological reality*

It has sometimes been suggested, both by philosophers of language and by linguists (REFERENCES; Jackson; Somes, Devitt) that accurate predictions of the intuitions (modal. grammatical or both) of informants is the most that can reasonably be required in some of the cognitive sciences; linguistics in particular. We assume, on the contrary, that insofar as

intuitions are of interest, that is only because they are ontologically reliable, and they are ontologically reliable only when they are often effects of mental processes of the sort that the cognitive sciences study. If there really aren't any such processes, who cares what informants intuit?

Fifth assumption: *Compositionality of propositions*

The propositions that are the objects of propositional attitudes (intuitively, the things that go in for the blanks in such formulas as `John believes that…; John remembers that…; John hopes that..., etc) have semantic contents; indeed, they have their semantic contents *essentially:* propositions which differ in contents are ipso facto different propositions.

The semantic content of a proposition is the product of (what philosophers call) its`logical syntax' together with its inventory of constituent concepts. Suppose the question arises whether the proposition *John is an unmarried man* is identical to the proposition *John is a bachelor* (and hence that whether believing that John is a bachelor is the same mental state as believing that John is an unmarried man . By stipulation, it is if and only if UNMARRIED MAN and BACHELOR are the same concepts.[3]

We assume that propositions are structured objects of which concepts are the constituents, much as English sentences are structured objects of which words (or, if you prefer, morphemes) are the constituents. Some *concepts* are also syntactically structured (eg. the concept GRAY SKY) and some (including, perhaps, the concepts GRAY and SKY) are `primitive'. Such analogies aren't, of course, accidental; propositions are what (declarative) sentences *express,* and (excepting idioms, metaphors and the like), *which* proposition a sentence expresses is determined by its syntax and its inventory of constituents. So the sentence `John loves Mary' expresses a different proposition than is expressed by the sentence `Mary loves John' or the sentence `John loves Sally, Likewise the *thought* that *John loves Mary* express a different propositions than either the thought that *Mary loves John* or the thought that *John loves Sally*. That thoughts and sentences match up so nicely is part of why you can sometimes say what you thinking and vice versa.

Sixth Assumption: *Compositionality of mental representations.*

If one believes that propositions are compositional, it is practically inevitable that one must likewise believe that mental representations are too. The proposition John loves Mary is true in

---

[3] `Why are you telling me this? I am a cognitive scientist; I am not a metaphysician. So why should I care whether believing that John is a bachelor and believing that is an unmarried man are the same mental state?' Well, imagine an experimental psychologist who is worried about what concepts are, and some of whose subjects decline the inference from `Hilary is a bachelor' to 'Hilary is an unmarried man' but accept the inference from `Hilary is a teenaged auto mechanic' to `Hilary is an unmarried man'. Do such subjects have the concept BACHELOR or do they not? In particular, does the datum argue for or against the theory that concepts are stereotypes?

virtue of John's loving Mary because it contains appropriately arranged constituents that are names of (the individuals) John and Mary, and of (the relation) of loving. Likewise, the mental representation JOHN LOVES MARY expresses the proposition *John loves Mary* because it contains appropriately arranged constituents that are mental representations of (the individuals) John and Mary, and of (the relation) of loving. The assumption that the compositionality of mental representations mirrors the compositionality of the propositions they expres does lots of useful work. For example, it explains why what one thinks when one thinks about John's loving Mary is, inter alia, something about John and Mary.

Perhaps it goes without saying that this is all very sketchy so far; certainly nothing so far amounts to a theory of the semantics either of propositions or of concepts, or of mental representations . At this stage, we're just setting out some of the vocabulary that we think such theories will require. But the decision to set things up this way is not mere stipulation. That sentences and propositions are both `compositional' is part of the story about why there are indefinitely many of each. Likewise, it had better turn out that the causal consequences of tokening in one's head the proposition that John loves Mary has different consequences for subsequent thoughts and actions than tokening the proposition that Mary Loves John. See practically any 19th Century novel.

Eighth Assumption: *The Representational Theory of Mind (RTM)*

So far, then:

- Cognitive phenomena are typically the effects of propositional attitudes.
- Relations between minds and propositions are typically mediated by relations between minds and mental representations that express the propositions.

For expository purposes, we'll refer to the conjunction of these theses as `The Representational Theory of Mind (RTM).'

We do understand that RTM is a lot to ask you to swallow even as a working hypothesis. Still, we aren't going to defend it here; suffice it that we're pretty much certain that RTM will have to be swallowed if cognitive science is to be interpreted Realistically, viz as a causal account of how the cognitive mind works; and the idea that cognitive processes typically consist of causal chains of tokenings of mental representations isn't itself at all radical, or even particularly new. It is almost always taken for granted in both Rationalist and Empiricist philosophy of mind, and is at least as old as Aristotle, Ockham, Descartes, Locke and Hume. To be sure, our version of RTM differs in a number of ways from classical philosophical formulations. We don't, for example, think that mental representations are images (images have a terrible time expressing propositions; which thoughts do routinely). And we aren't Associationists; that is, we think that mental processes are typically causal interactions among mental representations, but not that such interactions are typically governed by the ` laws of association.` More on this as we go along.

Ninth Assumption: *The computational theory of mind (CTM).*

Since mental representations are compositional, they must be structured. In the sort of cognitive psychology that was typical of Empiricism, the assumption was that the structure of the mental representations of propositions was associative Likewise, the structure of mental representations of `complex `concepts; ie all concepts that aren't primitives). To a first approximation: Mental representations of complex concepts are associations of mental representations of primitive concepts; mental representations of propositions are associations among primitive or complex concepts (or both).

Various considerations, some of which will be discussed later in this book, have made the inadequacy of this association/empiricist account of conceptual structure seem increasingly apparent. In the sport of cognitive science that is currently favored., the structures of mental representations of complex concepts and of propositions are both assumed to be *syntactic*: In particular, both have *constituent structure* (see above under *Assumption 6*). This. in turn, offers the possibility of a new view of cognitive mental *process* (like, in particular, thinking) according to which they are *computations.* Roughly. that's to say that are cognitive processes are defined over the constituent structures of mental representations of concepts and propositions that they apply to; which they may supplement, delete or otherwise rearrange. Thus the suggested analogy, ubiquitous in both the popular and the scientific literature these days, between minds and computers. This transition from associative to computational accounts of cognitive processes has the look of a true scientific revolution; in particular, it has opened the possibility of assimilating work in connect cognitive psychology to work in logic, computer science and AI. In what follows, we will take for granted throughout what follows that some version of a computational account of cognitive processes will prove to be correct.

Tenth Assumption: *The priority of thought to language*

In the course of linguistic communication, forms of speech inherit their semantic contents from the concepts and thoughts that they express; *not vice versa.* The reason that English speakers often utter the word `cat' when they wish to say something about cats is that, in English, `cat' refers to cats.[4] We take this to be very nearly a truism.

Still, why not go the other way around? Why not say that the semantics of linguistic forms that's what's at the bottom of the pile; it's what explains the semantics of propositional attitudes, thoughts and the like? Call that the `thought first' view and ours the `language first' view. We say, to put it roughly, that forms of speech inherit their semantic properties from those of the thoughts that they are used to express. The opposition says, to put it equally roughly, that

[4] Some people are squeamish about saying that linguistic expressions refer; rather, they say, it's people who do so, albeit by uttering linguistic expressions. Our sensibilities, however, aren't that refined; we're **quite** prepared to speak either way.

thoughts inherit their semantic properties from the forms of words that they are used to express them; in effect you can think about cats because you speak a language in which you can talk about them. Which, if either, of these views is right?

There is, in both philosophy and psychology, an enormous literature according to which mental states have semantic content only because they are states of the minds of language-using creatures; but we don't believe a word of that; usually either radical Behaviorism or radical Empiricism (or both) is lurking in the background, and the credentials of those have long expired yet another. Yet another case where bad philosophy has blunted perfectly reasonable commonsense intuitions.

There are very plausible reasons for holding that learning and using `natural` languages (English, Bantu, Russian, whatever) itself presupposes complex cognitive capacities; certainly neither philosophy nor cognitive science has thus far offered a serious reason for doubting that it does. For example, a tradition in the philosophy of mind claims that John's saying X because he intended to do say X is *not* an instance of a mental event with a certain content causing a kind of behavior that has the same content . Rather. it's a matter of a certain kind of behavioral disposition being manifested by a behavior that has a certain kind of semantic interpretation . But that simply can't be right. John's being disposed to say such-and-such isn't a sufficient condition for his saying it; whereas it's tautological that whatever caused John's to say X *must* have been a sufficient condition for his saying it; the cause of an effect is *ipso facto* sufficient for bringing about the effect. That a vase is fragile *can't* causally *e*xplain why it broke. (References to Wittgenstein and Ryle)

Another plausible argument against `language before thought':. Language learning --- including, in particular--- first language learning) takes a lot of thinking on the part of the learner. So, if you have to be able to talk before you are able to think, it follows that you can't learn a first language. This seems to be an embarrassment since, in point of fact, many children do so. Wittgenstein in philosophy and Skinner in psychology (reference Investigations, Verbal Behavior') both held that first language acquisition is somehow the effect of `training' (socially-mediated mediated' reinforcement or whatever). But that can't be right either since it turns out that children generally neither need nor get much language training in the course of learning a first language. Still more to the point, there is (as far as we know) no serious suggestion anywhere in either psychology or philosophy as to how, in the case of first-language acquisition, training might work its putative effects. (Skinner assumed that Learning Theory would provide the required explanation, but that proved to be a dead end. Wittgenstein dispensed altogether with a theories of learning, offering instead the suggestion that what happens in first-language acquisition is that the child learns how to play certain `language games' (presumably along the lines of `Jane-says-`Slab'; Tarzan-brings-slab'.) No details are provided.

In short: acquiring a first language is, prima facie, a very complex cognitive achievement; so far, neither pigeons nor computers are able to do it; nor has cognitive science been able, so far, to

explain it. So it's hard to imagine how first-language learning could proceed in a creature that lacks quite a lot of prior conceptual sophistication.

Here, in case they are needed, are some further reasons for holding that, in the order of explanation, as in the order of cognitive ontogenesis, thought comes first and language follows after:

- Thought-first avoids having to claim, from the armchair, that neither animals nor pre-verbal infants can think. But whether they can is surely an empirical issue, and, to our knowledge, there is no evidence that they can't. The thesis that (excepting, perhaps, occasional reflexes) new-borns are cognitively inert has become increasingly unattractive since the collapse of the Piagetian program in developmental cognitive psychology.
- Thought-first may explain the intuition that we can't *always* say what we think (cf the notorious tribulations of lawyers, logicians and poets).
- Thought-first may explain the intuition that we can almost always think what we say (compare: `I can't tell what I think until I hear myself say it', which is either just a joke or patently false).
- Thought, first may explain why we can communicate much of what we think to people who speak our language.
- Thought-first may explain why (with translation) we can communicate much of what we think even to people who *don't* speak our language.
- Thought-first may explain why, even if the `Whorf hypothesis'[5] turns out to be true, much of the evidence suggest that the effects of one's language on one's thought, perception, cognitive style, and the like are pretty marginal. REFERENCES

Of course, none of that is even close to conclusive. But we think it's sufficiently persuasive to warrant taking the priority of thought to language as a working hypothesis and seeing where it leads; notoriously, the way to prove the pudding is to swallow it and see what happens.

Summary: Basic Cognitive Science (BCS)

So far, we've supposed that beliefs and intentions are typical causes of behavior; hence that belief-desire causal explanation in cognitive psychology is OK (at least in principle); and that you can't learn or speak a language, including a first language unless you can already think. The conjunction of these of these claims constitutes what we will call `Basic Cognitive Science`

[5] Whorf hypothesized that cognitive states and processes are profoundly affected by differences between languages. (Presumably languages that are very different can't be translated.) These days, however, even its defenders endorse Whorf's only in its `weak' version. viz that there are at least *some* cognitive consequences of which language one speaks. (References)

(BCS), which we embrace whole-heartedly, and which we recommend that you embrace whole-heartedly too. But we don't wish to suggest, even for a minute, that if you do accept it, all your worries about cognition will thereupon disappear. Quite the contrary.

For example, we remarked that it is characteristic of cognitive science to offer theories in which tokens of propositional attitudes figure as causes and effects; and that propositional attitudes have semantic contents. So the cognitive science (very much including BCS) raises two questions that more familiar sciences do not, and to which nobody has yet got a satisfactory answer: *Just what is semantic content* and *Just what role should it play in a cognitive a science that is, by assumption, naturalistic?* Both of these questions are very hard, and we think that the ways of answering them that cognitive science has thus far proposed are wildly unsatisfactory. The rest of this book hopes to, in the words of our epigraph, to `start all over again'.

Chapter 2 Concepts Misconstrued

Many of the working assumptions we endorsed in Chapter 1 imply, directly or otherwise, constraints on theories of concepts and on theories of the mental representations that express concepts. For example, we've assumed that mental representations expressing concepts are the constituents of mental representations expressing the objects of propositional attitudes; we've assumed that concepts are productive, and that some of them (the `complex' concepts) have internal syntactic structure but others (the `primitive concepts') do not; we've assumed the psychological reality and causal involvement of mental representations of concepts and their internal structures; we've assumed that typical cognitive processes are sensitive to the conceptual inventories and structures of mental representations that they apply to… and so forth. That is, to be sure, a considerable inventory of assumptions to impose on what purports to be an empirical theory of the cognitive mind. But so be it. `Qualia', `raw feels' and such aside, everything in the mind that has content at all is either a primitive concept or a construction out of primitive concepts. This isn't to endorse any version of the `Empiricist Principle' that there is nothing in the mind that is not first in sensation; we are, so far, uncommitted as to what primitive concepts there are or what structures they can enter into. For all we've said so far, the primitive concepts might include ELEPHANT or CARBURATOR).

It shouldn't be surprising that so much of the Cog Sci literature (and of this book) is devoted to the nature of concepts, conceptual structures, and conceptual content. Once a psychologist has made up his mind about concepts, much of the rest of what he says about cognition is a forced option. But it is one thing to propose constraints on theories of concepts; it's quite another to propose an empirical adequate theory that satisfies the constraints. We now turn to our attempt to do so. That will come in three parts: In this Chapter, we will offer a critical (and frequently tendentious) review of the main theories of concepts that are currently on offer in the Cog Sci literature. Chapter 3 will explain why we think that they are, without exception, irredeemably wrong-headed: They all presuppose a more or less `Fregian' account of conceptual content; and, as Frege himself cheerfully agreed (RFEFERENCE), his sort of semantics rules out even the possibility of an empirical psychology of cognition.

Concepts as mental images

The root idea is that thinking about a Martini is having a mental image of a Martini, and a mental image of a Martini is much like a photograph of a Martini except that it is `in the head' rather than on the mantelpiece. So, if you believe, as many psychologists still do, that people have generally reliable access to what goes on in their minds ('How could I doubt my belief that I now have a mental image of a Martini? Could my mind's eye be lying to my mind's mind?') you may well wish to say that when you entertain a concept, your mind makes a picture of whatever it is a concept of.[1]

This introspective plausibility led to other claims about images. For example, a widely held belief among psychologists claims that there are two distinct kinds of thought: `verbal` (i.e. language-like) and pictorial. This has been called the `Dual-Code Theory' of mental representation (DCT) to which we will return below. The impact that DCT has had on psychology should not be underestimated. In

[1] Visual imagery has received most of the attention in Psychology, but other modalities of thought have also been proposed, such as kinesthetic, musical and mathematical thinking. REFERENCE

the 1960s and 1970s, much of cognitive psychology was concerned with discovering the parameters that affect leaning and recall. Of these, `frequency of occurrence' was said to be the most important. DCT opened the door to the suggestion that it is less the exogenous *frequency* than the endogenous *format* of a mental representation that is the primary determinant of recall (although "format" in this context was limited to serial-verbal or sentential form, or to pictorial or imagistic form). This view, initially put forward by Allan Paivio (Paivio, 1971), is now widely endorsed.

We think that DCT did well to emphasize the similarities between certain forms of mental representations and linguistic representations; that theme will recur throughout this book. But DCT also says that some concepts are mental images, and we think the objections to that claim are virtually insuperable. We'll now have a look at a number of these.

First reason why concepts can't be images: *concepts apply to things that can't be pictured.*

Bishop Berkeley famously pointed out (in his *Essay Towards a New Theory of Vision)* that the image theory fails for `abstract' concepts. You can picture a triangle, but you can't image *triangularity* per se*,* i.e. the property that triangles have in common *as such.* But one surely does have a concept that expresses that property; the concept TRIANGLE does. Likewise you can picture a chair, but not the property of *being a chair*, which is what all chairs have in common as such; and so forth. The problem, in all such cases, is that the property shared by all and only things in the extension of the concept (i.e. the things that the concept applies to) is, ipso facto, a *property* and, you can't make a picture of a property. The best you can do is make a picture of something that *has* that property. Yet, we do think about triangles and chairs from time to time; as when we think: *some chairs are comfortable* or: *not all triangles are oblique*). The right conclusion is surely the one that Berkeley drew: At least some of our concepts *couldn't* be images. That, indeed, Is what Berkeley's arguments are generally said to have shown; mental images can't be concepts of *abstract objects* like properties*.*

But it's a mistake to think of Berkeley's point as applying *only* to concepts of abstract objects. Consider your concept JOHN. Just as the concept TRIANGLE is too abstract to resemble all triangles *as such*, so your concept JOHN is too abstract to resemble John *as such.* A photograph of John shows how he looks *now, from this viewpoint, under this lighting, and in this context, and so on,* much as an image of a triangle shows how *that* triangle looks. But the concept TRIANGLE applies to each triangle *as such*, and the concept JOHN applies to John *as such* (in other words it applies to every possible token of triangle or of John). This is a mildly subtle distinction, but it shows something important: *concepts don't work like images.* JOHN doesn't apply to John in virtue of resembling *him* any more than TRIANGLE applies to triangles in virtue of resembling *them.* Berkeley's arguments are not about how `abstract' concepts differ from `individual' concepts; *they're about how concepts differ from images.* Or, to put it slightly differently, they're about how the `expressive power' of iconic symbols differs from the expressive power of discursive symbols. Considerations of their expressive power suggest that concepts (more precisely, the mental representations that express them) might be a lot like words. As we go along we'll see lots of reasons to believe that in fact they are; that is the substance of the `Language of Thought' hypothesis (LOT). But concepts can't be much like pictures.

There is doubtless much else of interest to say about the implications that differences of expressive powers between iconic and discursive have for cognitive theory of theory. For just one example: It might be argued that information can be extracted from the former more readily than from the latter; in effect, there might be tradeoffs between the expressive power of a form of representation and the computational complexity of drawing inferences from them. In an intriguing paper (Levesque, 1986) showed that a certain highly constrained system of representation ---one

that that doesn't directly express negation, disjunction, or universal quantification--- can avoid the exponential growth of complexity with increasing size of the database. To evaluate the hypothesis that, in a certain domain, *all the squares are red,* it is therefore necessary to check *each* square, searching for one that *isn't* red; to confirm the hypothesis that that there is no square in a certain domain, it is necessary to check each item to determine that it *isn't* a square; and so forth. (This is referred to in Artificial Intelligence as "negation by failure"). The expressive power of what Levesque calls `vivid' representations can be similar to the limited expressive power of pictures; for example, neither can express `practical syllogisms' (see Chapter 1). But, in certain cases (e.g., suitably small domains) the computational complexity of processes that involve this kind of representation would not scale exponentially with the number of objects than more general representations with quantifiers (Levesque & Brachman, 1985). What other kinds of power-complexity tradeoffs there may be between different systems of representation remains to be investigated.

Second reason why concepts can't be images: `*Black swan' arguments*.

The kinds of points we've just been making are familiar from Philosophy. But other considerations, less widely recognized, also tend to the conclusion that concepts can't be images (mental or otherwise). Some of them are very important when issues about the relations between conceptualization and perception come to the fore.[2]

Perception interacts with prior (`standing') beliefs. All else equal, seeing a black swan reduces one's previous epistemic commitment to the belief that all swans are white. How does this work? The image theory offers what may seem, at first glance, a tidy suggestion: Believing that all swans are white consists in having a mental image that shows all swans as being white. Seeing a black swan causes the formation of a mental image of a black swan. If the two images match, you decrease your epistemic commitment to *all swans are white* (and/or decrease your epistemic commitment to *this swan is black*.)

But that account could not be right. To the contrary, `Black Swan Arguments' show that the mental representations corresponding to perceptions and beliefs must both have a kind of semantics that images do not have. Draw a picture that shows a black swan. Now draw a picture that shows all swans are white.[3] Now draw a picture that shows that these two pictures are incompatible. It can't be done; compatibility isn't a kind of relation that pictures can express. This is one of the respects in which, the expressive capacity of `discursive' representation much exceeds that of `iconic` representation, as does the expressive power of thought. Once again, it appears that cognition can't be much like photography.

---

[2] In the philosophical literature, the sorts of points we're about to explore arose primarily in epistemology. They trace directly back to Sellars (who thought, probably wrongly, that they show that there are no `sense data') and indirectly to Kant and Frege. REFERENCES

[3] `How on earth can I do that?' Good question. Unlike `discursive' symbols, pictures have no way to express quantification, so they can't distinguish `this swan is white' (or `lots of swans are white') from `all swans are white'. But minds can. Our's just did.

Third reason why concepts can't be images: *constituent structure*

Concepts can have constituents, but images only have parts. For example, the complex concept MARY AND HER BROTHERS, which refers to Mary and her brothers, has a constituent that refers to Mary and a constituent that refers to her brothers. A crucial problem (the problem of `compositionality') is: How do the referents of complex concepts derive from the referents of their less complex constituents? What makes this problem crucial is that, to the best of anyone's knowledge, solving it is, the only way to explain why concepts are *productive* (vide: MARY'S BROTHER; MARY'S BROTHER'S BROTHER; MARY' BROTHER'S BROTHERS, etc….). And the productivity of concepts is needed to explain why there are so many of them that one is able to have and so many thoughts that one is able to think A precisely analogous situation arises with respect to the phrases and sentences of natural languages: It's because `Mary's brother' and `Mary's brother's brothers' are among their constituents that `Mary's brother is tall' and `Mary's brother's brothers are tall' are both sentences.

Neither the productivity of thoughts nor the nature of the productivity of sentences fully understood. But enough seems clear to add to our list of ways that concepts are different from images. Consider, to begin with, the productivity of pictures. Presumably there are indefinitely many of those, just as there are indefinitely many concepts, indefinitely thoughts and indefinitely many sentences. That's because pictures have parts, and (arguably) the result of putting the parts of a picture together is also a picture, and it too has parts. A picture of Mary and her brother can be analyzed into parts, some of which are pictures of Mary (or parts of Mary) and some of which are pictures of Mary's brother (or parts of him). But it is also perfectly possible to make a picture that shows Mary's left arm and her brother's nose. (Think of all the ways in you could carve up a picture of Mary and her brother into a jigsaw puzzle). That's why you often can't put the pieces of a picture back together unless you already know (or can guess) what it is a picture of.

The difference between something's having constituents (as thoughts and sentences do) and its only having parts (as pictures do) now becomes apparent: If a symbol has parts that are constituents, there are rules (in effect, syntactic rules) for combining them in a way that is guaranteed to produce a more complex symbol that *is itself a constituent* and you can follow such rules even if you don't already know what that complex constituent means. If, for example , the grammar of English is in force, the set of words `of friend a John's' can be put together in only one way that makes a well-formed sentence. That, fundamentally, is why you can use English to communicate new news. Sam says to Bill, "a friend of John's died"; and if Bill knows English, he can figure out that Sam has told him that a friend of John's died, which may well be news to Bill. What a lovely Invention! You just can't do that sort of thing with pictures.

Pictures have parts but not constituents; so if concepts are pictures, they too must have parts but not constituents. Since, however, concepts have both, it follows that concepts aren't pictures.

One more point along these lines and we can then go to yet other reasons why concepts can't be pictures. It's important, in thinking about whether pictures have constituents, to keep in mind, not just the distinction between constituents and parts, but also the distinction between the parts of a picture and the parts of what it pictures. Arguably at least, things in the world have parts that may, or may not, be among their constituents. (Maybe the constituents of an automobile are its `functional' parts: if so, the drive chain of a car is among its constituents, but an arbitrary piece of the front seat cover probably isn't). This sort of distinction has implications for psychology: for example, `memory images`, when they fragment, tend to respect the constituent structure of what they are memories of, not just arbitrary collections of its parts (Pylyshyn, 1973). Thus, you may `lose' part of a memory-image of John and Mary, but it's likely to be a part that represents John or a part that represents Mary (rather than, say, a part that represents a combination of John's nose and Mary's

toes.) But this consideration *doesn't* tend to show that pictures of John and Mary have constituents; at most it shows that John and Mary do. So the question remains whether mental images (as opposed to the things that mental images represent) have constituents or only parts; if, as we're suggesting, they have only the latter, then mental images are, once again, very poor candidates for being concepts.

Fourth reason why concepts can't be pictures in the brain: *`Leibniz' Law' arguments* [4]

Real images (pictures, drawings) share some, but not all, of their properties with the things they are images of (since they derive from geometrical/optical projections of the objects depicted).[5] But it's hard to see how *mental* images could do so unless one assumes that they are identified with parts of the brain, such as the surface of the visual cortex. If so, then pieces of cortex can have (at least some of) the properties that mental images do. For example, some psychologists have managed to persuade themselves that mental images of big things (i.e. mental images of things that show them as being big) correspond to big pieces of cortex, mental images of little things correspond to little pieces of cortex, and so on (see, for example, Kosslyn, 1975; Kosslyn, 1994). Even so, there are lots of kinds of cases where it's just not possible to believe anything that's `in the head', cortex or otherwise, has the *same* properties that corresponding mental images seem to have. Suppose, for example, that one's mental images are indeed displayed on the surface of one's visual cortex. Still, it's not credible that the pieces of cortex on which the images are displayed have any of the properties that the image does. Surely it isn't true that red mental images (that's to say, mental images of red) correspond to red pieces of cortex. Cortex is (more or less) gray. Similarly a vertical (or horizontal) elongated figure is not represented in cortex as a vertical (or horizontal) elongated patch.[6] Nor do auditory images of loud sounds correspond to loud pieces of cortex. Perhaps they correspond to *excited* pieces of cortex but that's no use: there are no such things as excited *mental* images. The problem, of course, arises from confounding the form with the content of images.

Early empirical support for the role of mental images came from studies of learning (specifically, associative learning of lists and pairs, see Paivio, 1971). More recent support arises from reaction times interpreted as indicators of cognitive processes. The most frequently cited of such studies get their intuitive grip from correspondences between the times it takes to perform a certain operation on a mental image and the time that it would take to perform that operation on the thing itself. For example, imagine a dot traversing a mental image; the time it takes to move across the image

---

[4] Liebniz's Law is the principle of logic that says: If A is identical to B, then every property of B is a property of A and vice versa.

[5] There is a story about a soldier visiting France after the war who met Picasso. The soldier admitted that he didn't much like Picasso's paintings: ``They don't picture people realistically' `What do you mean by realistic?' asked Picasso?) The soldier pulled out his wallet and showed a picture of his girlfriend. `Like that', he said. Picasso looked at the picture carefully and said to the soldier `I'm sorry to see that your girlfriend is so small and thin and colorless'. Exactly.

[6] The belief that representations must have many of the properties of what is represented (e.g., orientation) caused a great deal of misdirected research in the 17th century. Johannes Kepler was able to break through a lot of misguided beliefs about how the light is focused by the eye's lens, but was unable to solve the problem, then taken very seriously, of how we can see an upright world when the image on the retina is upside-down (Lindberg, 1976). Current worries about mental imagery suffer from a very similar mistake about mental images.

increases linearly with the distance it is imagined to travel. Similarly, the time it takes to rotate an imagined shape so that it is congruent with an actual shape shown beside it increases with the angle through which the former is imagined to be rotated. Our point is that, however subjects perform such tasks, *nothing in their heads ---in particular, no part of their cortex, including a pattern of activation--- moves when they do* so. Moreover all the psychophysical data, as well as the phenomenology of mental imagery, suggest that if spatial operations (such as scanning or rotation) literally apply to the mental image, then the images and operations must be *in three dimensions.* In particular, they can't be operations on the two-dimensional[7] surface of visual cortex. Apart from needing to be in three dimensions, there are many empirical reasons why they couldn't be displayed on visual cortex. Some examples (many others are discussed in Pylyshyn, 2003, 2007) include: impaired mental imagery and impaired visual perception can occur independent of one another; activity patterns in V1 are retinotopic while mental images are in allocentric coordinates (they do not appear to move when you move your eye or head); a mental image superimposing on a visual scene fails to scale the image according to Emmert's Law (the further away the visual background is the larger the objects in the mental image ought to be – as they are when an *afterimage* is superimposed on a perceived scene); mental images are intensional objects -- unlike pictures, mental images are already interpreted and can't be *re*interpreted in the way that ambiguous pictures can. (Empirical support for these and other such important differences between mental images and displays in V1 can be found in (Pylyshyn, 2003, Ch. 4; and Pylyshyn 2007, Ch. 7).

Leibniz' Law arguments seem to show that *noth**in**g in* your head can have the properties that mental images are said to have; and if your images aren't in your head, where *could* they be? We don't say these kinds of objections can't be answered; as usual, a lot depends on what you're prepared to swallow. But we do say that, if they can't be, then the suggestion that there are mental images, neural or otherwise, makes no clear sense.[8]

We've just been seeing that some of the (alleged) properties of mental images are properties that nothing that is literally in the head has. Likewise, though Introspection may suggest that one inspects one's mental images by a process that is much like seeing (in effect, that your mind's eye *sees* the pictures in your head), But just as it can't be literally true that a mental image of something red corresponds to a red piece of V1, it's likewise hard to take literally the idea that introspecting a mental image is a kind of seeing it. It's worth spending a moment on this because some seems to attract many image theorists are attracted. Mental images can't literally be *seen* any more than they can literally be *touched.* If introspection is a kind of mental seeing in which the actual, physical visual system is engaged, then mental images must be kinds of things that can reflect patterns of light that the visual system can detect, But there isn't anything of that sort. Brains, unlike like mirrors', don't reflect a*nything*. It's very dark in your head.

---

[7] Of course we don't mean that the cortex is two-dimensional. But only the surface of the cortex maps the topology of the retina. Go deeper into the cortex and you don't find a representation of the third dimension of the perceived world.

[8] There is a story about a soldier visiting France after the war who met Picasso. The soldier admitted that he didn't much like Picasso's paintings: ``They don't picture people realistically' `What do you mean by realistic?' asked Picasso. The soldier pulled out his wallet and showed a picture of his girlfriend. `Like that', he said. Picasso looked at the picture and said to the soldier `I'm sorry to see that your girlfriend' is small, thin and colorless'. Exactly.

Concepts as definitions

This is perhaps the most familiar account of concepts. It says there are two aspects of a concept's content: its reference and its `meaning` (or 'intension' (with an `s'), or its `sense`). The *ex*tension of a concept is *the set of things that the concept applies to* (according to the view we're inclined towards, it's the set of actual or possible things that the concept applies to; see Chapter 3) The *in*tension of a concept is the property in virtue of which things belong to its extension. So, for example, if the intension of the concept CAT, (and hence the meaning of the word `cat') is supposed to be domestic feline then it expresses the property of being a domestic feline and all and only domestic felines belong to its extension.

That is probably more or less the semantic theory they taught you in grade school. No doubt, they did so because they believed that words express concepts and that concepts are something like definitions. That, in turn, is why your teachers your teachers kept telling you how important it is for you to `define your terms`, thereby making clear exactly which concept you have in mind when you use them. But, like so much else of what they taught you in grade school, the theory that concepts are definitions most likely isn't true. We'll presently say why it isn't, but let's start with some of its virtues.

- As we just saw, the theory that concepts are definitions seems to explicate the relation between their contents[9] and their extensions. Likewise, the theory that concepts are definitions suggests a prima facie plausible account of having a concept: To have a concept is to know its definition In some cases a stronger condition is endorsed: To have a concept is to know not just its definition but also to know how to apply its definition; (i.e. how to tell whether something is in the concept's extension). Philosophers of a verificationist turn of mind, in particular, often like to think of having a concept as knowing `ways of telling' (aka `criteria` for telling) whether something falls under it . The attraction of verificationism is that it seems to avoid the (crazy ) skeptical thesis that there is no way of telling `for sure' what kind of thing a thing is. At very least, if a concept is some sort of verification procedure, then one can tell for sure whether something is in its extension (assuming, of course, that one knows how to apply the verification procedure; a caveat that verificationists invariably ignore).
- Many philosophers, and more psychologists than you might suppose[10], have thought that semantics should underwrite some notion of analytic truth (`truth in virtue of meaning alone') thereby connecting semantic issues with issues about modality. If it's true by definition that cats are felines , then there can't be a cat that isn't a feline; not even in 'possible worlds' other than our own; that is, the extension of CAT, contains neither any actual nor any possible cats that aren't felines. So if you think that there couldn't be such cats, you may well think that the definitional account of linguistic (and/or conceptual) content is the very the semantic theory you require since it purports to explain

---

[9] Perhaps you object to speaking of the `meaning of a concept': 'Concepts don't have meanings,' you may wish to say, concepts are meanings'. So be it. In practice, when we wish to speak of `conceptual content` we will usually be referring to the concept's intension or its meaning or its sense. But, since we don't believe that there actually are such things as intensions, meanings or senses, we think that all that concepts have by way of semantic content is their extensions; our use of the other terms will be just expository.

[10] See, for a classic example (Bruner, Goodnow, & Austen, 1986 / 1956) who are pretty explicit in endorsing a definitional theory of conceptual content. Accordingly, they test whether a subject has a concept by determining whether he is able to apply the concept correctly. See also the tradition of `procedural' semantics in AI. REFERENCES)

why there can't be. (`Linguistic semantics' and `analytic philosophy' both have a vested interest here since both rely heavily on informants' intuitions of modality as data for their theories.)

- It is plausible, first blush, that the definition story about conceptual content accounts for the fact that concepts compose; i.e. that you can make up new concepts by putting old ones together . (By contrast, see the discussion above of the troubles that composition makes for image theories of concepts.) The putative explanation is that concepts compose because concepts are definitions and definitions compose. If you have the concepts BROWN and DOG (that is, if you know their definitions) you can compute the content (i.e. the definition) of the concept BROWN DOG; a brown dog is anything that satisfies both the definition of `brown` and the definition of `dog'. And if you have the concept BROWN DOG and the concept BARK), you can likewise figure out the content of the concept BROWN DOG THAT BARKS; by definition, a brown dog that barks is anything that falls under the concepts BROWN, DOG, and BARKS. And so on ad infinitum. That might answer the question how a merely finite brain can master an indefinitely large repertoire of concepts, so it begins to make the definition theory of conceptual content look sort of promising.

Nonetheless, concepts aren't definitions. Here are some of reasons why they aren't.

Most words just don't have definitions, which they surely would if the concepts they express are definitions. More than anything else, it's the lack of a robust supply of clear examples that has led to the recent steep decline in popularity of the definitional account of conceptual content in cognitive science.[11]

Still, there is a scattering of words/concepts that really do seem to have definitions (like BACHELOR = df unmarried man; and; maybe WANT … =df desire to have…' etc. These provide the standard examples in introductory linguistic semantics courses, where they have become dull with overuse. Then too, there are concepts drawn from specialist vocabularies. These are, often enough, the products of explicit agreements and stipulations. So, a yawl is a two-masted sailing vessel of which the after mast is stepped forward of the helm. Similarly, a square is a four sided closed figure, all the sides of which are straight lines of the same length that intersect at 90 degree angles; the Jack is the card that comes between the ten and the Queen in a standard deck... and so forth.

Such definitions often do specify more or less necessary and sufficient conditions for being in corresponding extensions, and they are typically learned by explicit instruction. But they clearly don't suggest themselves as plausible exemplars for a general theory of word or concept meaning. Here are some examples drawn from a dictionary that we happen to have at hand: a `game' is "a contest governed by set rules, entered into for amusement, as a test of prowess, or for money, or for other stakes." But, as Wittgenstein pointed out, skipping rope doesn't count as a game by this criterion. Nor does a discarded tin can count as `garbage' according to our dictionary, which says that garbage is "refuse from a kitchen, etc consisting of unwanted or unusable pieces of meat, vegetable matter, egg shells etc." Such `definitions' are too open to determine extensions (note the `etc'), nor are they

---

[11] Though these days most cognitive scientists are dubious about it, the definitional theory of conceptual content still has many adherents among linguists; they hold that word meanings are, in many cases, definitions (which is tantamount to holding that concepts are, assuming that the meaning of a word is the concept it expresses). It is, for example, frequently claimed that there are structural (syntactical) arguments that favor a definition of `kill' as CAUSE TO DIE, and likewise in the case of other `causative' verbs. There is, in linguistics, a very large literature on this; indeed, it is one of the characteristic theses of `linguistic semantics`. But its status remains much in dispute. For what it's worth, however: `x killed y' and `x caused y to die' are clearly not synonyms; they aren't even coextensive: Because `cause' is transitive, and `kill' isn't, you can cause someone to die without killing him (eg. by causing someone else to kill him.) REFERENCES

seriously intended to do so. You may have been taught in grade school that a good way to `define a term' is to look it up in a dictionary. But it bears emphasis that dictionary entries generally don't provide definitions in that sense of `definition'. Rather, dictionary entries are typically meant to be informal guides to usage for an intended reader, who can already know the language in which the defined term occurs and can thus pick up a new term from a scattering of examples (together, sometimes, with some more-or-less synonyms). That is generally just what we want a dictionary for when we consult one. But, as Wittgenstein also pointed out, it isn't only having the examples but knowing how to extrapolate from them --- knowing `how to go on'--- that does the work; and that kind of knowledge dictionaries don't provide or purport to.

Even if lots of concepts did have definitions, there couldn't, barring circularity, be definitions for every concept; so what is one to do about the semantics of the ` primitive' concepts in terms of which the others are defined? This question is urgent because, if it isn't answered, it is easy to trivialize the claim that the sense of a concept is the property that is shared by all and only things in its extension. It is, after all, just a truism that being green is the very property that all and only (actual or possible) green things have in common; namely, the property of being green. Likewise, there is a property that all and only actual and possible cats share: the property of being a cat; likewise, there is a property that is common to all and only Sunday afternoons: the property of being a Sunday afternoon; likewise, for that matter, there is a property that all and only Ulysses S. Grants share: the property of being Ulysses S. Grant. If the thesis that concepts are definitions is to be of any use, such vacuous definitions must somehow be ruled out. Presumably that requires deciding which concepts are primitive (hence undefined) and what makes primitive concepts primitive. But, in fact, nobody knows how to answer either question.

Here's one suggestion we've heard: The primitive concepts are such very general ones as, PHYSICAL OBJECT, EVENT, PROPERTY etc.[12] But this seems to beg the question at issue since, presumably, PHYSICAL OBJECT has an extension too, so the question arises: what do the actual and possible things in its extension have in common? (It seems not to help to say that the Intension of PHYSICAL OBJECT is something that is physical and an object since this raises the question what the intensions of those concepts are.) Nor does the dilemma's other horn seem more attractive; namely that PHYSICAL OBJECT can be defined and so isn't primitive. For example, perhaps something is a physical object in virtue of its having: a `closed shape`, and/or a `continuous trajectory in space', etc (Spelke reference). We've tried hard, but without success, to convince ourselves that the concept of a TRAJECTORY is more basic than the concept of a PHYSICAL OBJECT; isn't it true by definition that a trajectory is the path of an (actual or possible) physical object through space? We will tell you (in ch.3) what we think a physical object is; but what we'll offer isn't (and doesn't purport to be) a definition; and. very likely, you will find our account of it disappointing.)

It is to be said in praise of the Empiricists (Hume, for example) that they did offer a serious suggestion about how to deal with this complex of worries: According to their `Empiricist Principle` all concepts reduce, via their definitions, to sensory /experiential) concepts. Accordingly, the primitive concepts are the sensory/experiential ones. But , of course, this is a principled answer to 'which concepts are primitive' ? only if there Is a principled answer to 'which concepts are sensory/experiential?' `Classical' Empiricism thought that there is: Roughly, a sensation is a mental

[12] Notice, once again, how much of the work the `etc' is doing. Perhaps that wouldn't matter if the examples really did showed us `how to go on' from them; but they don't.

object such that, you have one if and only if you believe that you do.[13] When taken together with the claim that sensory definability is the general case for concepts that aren't primitive, that provided Empiricists with their purported refutation of skepticism;. if all your concepts are directly or indirectly sensory, and if your sensory beliefs about your current sensations aren't subject to error, then contrary to what skeptics say, it follows that at least some of your beliefs about things-in-the world (tables and chairs and the like) are certainly true.

But according to the Empiricists' kind of semantic theory, intensions determine extensions; and, these days, it's hard to believe that, in the general case, things fall in the extensions of concepts in virtue of their sensory properties (that is, in virtue of how they look, sound, feel or taste and so forth). Maybe that works for GREEN, (though that it does is capable of being doubted). REFERENCES But it quite clearly doesn't work for CAT or PLUMBER (or for PHYSICAL OBJECT, come to think of it.) Moreover, If it's true that there are very few bona fide definitions, there are still fewer bona fide sensory definitions. It is rare for things to fall under a concept because they satisfy a sensory definition (except, maybe, sensations). Being a cat isn't the same property as being cat shaped, cat colored, and /or being disposed to make cat-like noises. We wouldn't b cats even if we were all of those.

As for us, we're not at all sure that refuting skeptics is a project that's worth the effort. Why exactly does it matter whether or not it is, in some tortured sense of the term, possible that there are no tables or chars or elephants or trees, given that, as a matter of fact, there are perfectly clearly lots of each? In any case, nothing further will be said about skepticism in this book.

One more objection to the theory that concepts as definitions: If they were, they couldn't be learned. Consider the paradigm of a definition, `bachelors are unmarried men'. Presumably it entails that the concept BACHELOR is the very same one as the concept UNMARRIED MAN; so it entails that to learn BACHELOR is (at a minimum) to learn that bachelors are unmarried men. But, by assumption, BACHELOR and UNMARRIED MAN are the very same concept. So, on the assumption that leaning a word is learning its definition, it follows that you can't learn the concept BACHELOR unless you already have the concept UNMARRIED MAN (and, of course, vice versa.) So you can't learn the concept BACHELOR (/UNMARRIED MAN) at all. This consequence (`sometimes known as `Fodor's Paradox') is patently intolerable. (There is a way out of it, as we'll see in later chapters. But it is very expensive; it requires abandoning the notion that concepts have intensions.)

It is essential, in thinking about the line of argument just sketched to keep in mind the distinction between concept learning, which adds a new concept to one's prior conceptual repertoire, and word learning, which merely provides a new term with which to express a concept. It's easy to conflate the

[13] This definition of `sensation' is, of course, blatantly epistemological; and, unsurprisingly, it begs the question why your beliefs about your beliefs about your which sensation you are having are infallible when so many other kinds are notoriously not. By contrast, cognitive science generally prefers to talk of `transducer outputs' rather than `sensory experiences': roughly, transducer outputs are (directly) caused by mind-world causal interactions, and they (directly or otherwise) cause perceptual beliefs. We will ourselves adopt this sort of view; but, of course, it doesn't provide anything remotely like a definition of `sensory' since nothing remotely like a definition of `transducer' is available. And it offers no answer to even such vexed questions as what, if any, connections there are between being sensory and being conscious.

two, but doing so leads to disaster.[14] There's a heuristic that helps to keep the distinction between concepts and definitions clear: test for symmetry. Consider, once again, `bachelors are unmarried men'. Since the definition of a term is synonymous with the term, then if learning the definition of `bachelor' is unmarried man are, and learning that bachelors are unmarried men is the way one acquires the concept BACHELOR, it follows that learning that unmarried men are bachelors is a way of acquiring the concept UNMARRIED MAN. But, clearly, you can't learn UNMARRIED MAN that way unless you already have the concept BACHELOR which, according to the definition, is the concept UNMARRIED MAN. The moral: definitions are about words, not about concepts. What with one thing and another, the mainstream opinion in Philosophy and Cognitive Science these days is that the definitional theory of conceptual content isn't true.

Concepts as stereotypes

Just as the sheer scarcity of good examples was a primary cause of so many cognitive scientists abandoning the definitional model of concepts, it is a primary cause of the current enthusiasm for the suggestion (endorsed by many people, including Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976) that concepts are stereoptypes; there are plenty of good examples of those. (REFERENCES) Moreover, while there are relatively few cases of reliable effects of a subject's knowledge of definitions on the sorts of experimental tasks that cognitive psychologists like to use to measure cognition, such effects are easy to elicit when the manipulated variable is the extent to which a stimulus is `similar enough` to the stereotype of the category that the stimulus belongs to.[15] In effect, the difference between stereotype theories of concepts and definition theories is that while satisfying its definition is said to *be necessary and sufficient* for falling under a concept, similarity to its stereotype is only said to be sufficient.

There are , in fact, large effects of stereotypy on many standard tests of concept possession : the strength of associations (if you are asked for an example of an animal, you are much more likely to respond `dog' than `weasel. Very likely that's because DOG is a stereotype for ANIMAL and weasles aren't very similar to dogs. Likewise, there are large effect of stereotypy on reaction times in identification tasks (dogs are more quickly recognized as dogs than weasels are recognized as weasels); on `inter-judge reliability' (we are more likely to agree about' whether a dog is an animal than about whether a germ is); Stereotypy is a better predictor of how early a word is learned than the frequency with which it occurs in utterances; the probability that a property of a stimulus will generalize to others stimuli of the same kind is higher when the stimulus is a stereotypic instance of a kind than when it isn't (a subject who is told that sparrows have a ten year life span and then asked whether starlings do too, is more likely to guess `yes' than a subject who is told that starlings have a ten year life span and is then asked to guess whether sparrows do too.) Etc. These sort of effects generally persist when `junk variables' are controlled. In short, effects of stereotypy on subjects' behaviors are widespread and reliable; this finding is about as certain as the facts of cognitive psychology ever get. The open question,

[14] We think the experimental studies that Bruner, Goodenough, and Austin (op cit) take to bear on concept acquisition are really about how people acquire words for concepts they already have. In particular, they don't bear, one way or the other, on whether concepts are definitions.

[15] Of course everything belongs to more than one kind. This desk belongs to the kind `thing in my office' and to the kind 'thing purchased at a discount' and to the kind 'wooden thing' and so forth. It is, indeed, perfectly possible for a thing to be stereotypic of one of the kinds it belongs to but not of others. An example is given below.

then, is not whether subjects know which stimuli are stereotypic of their kinds; it's whether stereotypes have the properties that concepts need to have. And they clearly don't.

To begin with their virtues: stereotypy is a graded notion ; things can be more or less stereotypic of a kind. A dog is more stereotypic of ANIMAL than is a pig; and a pig is more stereotypic of ANIMAL than is a weasel; and a weasel is a much more stereotypic ANIMAL than a paramecium, and so on. Indeed, a thing that is pretty stereotypic of one of the kinds that it belongs to can be far less stereotypic of another. A chicken is a reasonably good example of something to have for supper, but it's only a so-so example of a bird. Like the relative ubiquity of stereotype effects, all this argues for the thesis that concepts are stereotypes at least in contrast to a theory that says they are definitions. It is something of an embarrassment for the definition story that weasels are less good examples of animals than dogs; either weasels satisfy a putative definition of ANIMAL or they don't; either they are in the extension of ANIMAL or they aren't. The definitional theory offers no account of why that should be so. (Likewise for vague concepts, marginal instances of concepts, and so on). Because stereotype theories prefer graded parameters to dichotomies, they have no principled problem with such facts.[16]

A question in passing: Standard semantic theory has it that intensions determine extensions. Does the stereotype theory agree with that or not? That depends, of course, on what `determine' means. A definition determines an extension in the sense that it is true of all and only the things in the concept's extension. If `bachelor' means unmarried man, then all and only unmarried men are in its extension. That , however, isn't the way stereotypes work. Rather, the extension of a stereotype is usually said to be the set of things that are `sufficiently similar' to the stereotype;[17] and it is left open what sufficient similarity consist in. Indeed, what sufficient similarity consists in is different from concept to concept. The way that kings are similar to one another other is quite different from the way that oboes or crickets are. That's why it is natural for stereotype theorists to speak of a stereotype as a (more or less definite) location in a `multi-dimensional' similarity space, where different concepts correspond to different locations on what may or may not be dimensions of the same space.[18]

In any case, concepts can't be stereotypes. The argument that they can't is surprisingly straight-forward: If concepts are stereotypes, stereotypes have to compose; if stereotypes don't compose, the productivity of concepts defies explanation. But (unlike definitions), stereotypes don't compose. Here's the classic example: there are stereotypic fish; trout as it might be. (Bass and flounders aren't, perhaps, very similar to trout, but let's assume they are adequately similar for the purposes at hand. Certainly trout and bass are more similar to each other than either is to a rock or a tree.) Likewise, there are stereotypic pets; dogs win by miles with cats a bad second. But, crucially, the stereotypic pet fish isn't either a dog or a cat, it's (maybe) a goldfish. It is utterly clear that there is no general way to compute the stereotype structure of compound nouns, adverbiably modified verbs etc from the stereotype structures of their constituents. (Compare, for example, PET FISH with SUGAR SUBSTITUTE: A pet fish is a fish, but a sugar substitute isn't sugar.)

---

[16] On the other hand, there are dichotomous categories, and they do embarrass the identification of concepts with stereotypes. It is presumably a conceptual truth that here is no such thing as a number that is more or less even; but subjects will tell you that two is a better example of an even number than 78 is. (Gleitman REFERENCES)

[17] Assuming that stereotypes can properly be said to have extensions. It's no accident that people who hold that concepts are stereotypes are also prone to hold that the extension of a stereotype is a `fuzzy' set.

[18] There is a variant of stereotype theory according to which there a single similarity space, on the dimensions which stereotypes are all located: These are said to include color, shape, sound, smell, texture and the like `sensory' properties (Reference, Churchland) . But this is just Empiricism all over again unless there is a serious attempt to make clear which properties are to count as sensory (see above).

The cognitive science literature offers various rebuttals to this line of argument; REFERENCES; but none of them strikes us as remotely convincing. A standard suggestion is that the function of stereotypes is heuristic: Suppose that you have the concepts F and G and your task is to understand the concept an F that is G. Pretty generally, it's safe to assume that an F that is G applies to, and only to, things in the intersection of F and G. But, as we've just been seeing, that doesn't hold if concepts are identified with stereotypes: paradigmatic pets aren't fish and paradigmatic fish aren't pets. This looks to us like a conclusive argument against the thesis that concepts are stereotypes, but there is a way out that is widely credited: Think of stereotypes as defaults; if you don't know what Fs that are Gs are like, then bet that they are like stereotypic Fs and/or stereotypic Gs.

There is, however, a lot wrong with that suggestion. For one thing, it recommends bad bets. You would be ill advised to put a lot on Martian fauna being very like ours fauna; or on the likelihood that paradigmatic Roman citizen commuted to work in much the same way we do (viz by bus or subway); or that Winter at the North Pole is much like Winter around here; and so on. In fact, experimental studies by Connolly et al show that more subjects are aware that they don't know much about Fs that are Gs, the less they are willing to bet that Fs that stereotypic Fs are Gs is the general case. (Subjects who are prepared to be that carrots are orange are measurably less willing to bet that Alaskan carrots are.) This is entirely sensible of them. `Stereotype` is a statistical notion; and, where the standard deviation of a distribution is large, the mode is a bad predictor.

In any event , we think that the compositionality problem has been quite generally misconstrued in the Cog Sci literature because it is regularly taken for granted (without argument) that the appropriate context for raising semantic issues is in theories about language --- in particular, theories about linguistic communication--- rather than in theories about thought. Consider a case where a pet fish problem might arise: Smith says that he is terrified of pet fish. Jones is puzzled: `Why should anyone be scared of them? Pet fish are generally (and correctly) supposed to be innocuous'. Various possiblities might suggest themselves: Perhaps it's something in Smith's childhood; maybe he was once bitten by a rabid goldfish. Or possibly Smith doesn't know which kinds of fish are kept as `pet fish'; he thinks that a lot of pet fish are piranhas. Or perhaps (less plausibly if Smith is a fluent speaker) he doesn't understand that the form of words 'pet fish' applies to all and only pets that are fish. In any event, Jones has to figure out what Smith means by `pet fish' and, in the last case, this is plausibly a fact about the semantics of English. Our point, however, is that the Smith-and-Jones case is very unlike the ones that arise in considering the psychology of thought. Suppose Smith thinks (in whatever language he thinks in) that is a pet fish, and decides, accordingly, to flee. What can't be going on is that Smith thinks (the Mentalese equivalent of) `that is a pet fish' and then asks himself `I wonder what I meant by `pet fish' when I thought that. For Smith-Jones communication to occur, Jones has to figure out what Smith was thinking of when he said what he did; but Smith doesn't have to figure out what he himself was thinking about when he thought to himself that's a pet fish'\. Acordingly, it's at least arguable that, in figuring out what Smith meant, Jones consults a semantics of English to which he and Smith both have access; Smith uses it to translate his thought into English, Jones uses it to reconstruct Smith's thought from Smith's English utterance. But, to repeat, that's about interpreting Smith's utterance. Smith doesn't interpret his thought; Smith just has it. And Jones doesn't interpret Smith's thought either; he only needs to interprets Smith's utterance: that is, he figures out which thought Smith intended the utterance to convey.

We think these sorts of reflections have implications for cognitive psychology. First: So far at least, the theory of linguistic communication may well require postulating an internal representation of English semantics as part of its story about how English speakers communicate in a language the share. But we doubt that there's any reason at all to suppose thinkers resort to a semantics of Mentalese when they use Mentalese to think in. Tokens of Mentalese are, to be sure `in the heads' of thinkers; and, Mentalese had better have a semantics in the sense that there had better be indefinitely many truths

about what a creature is thinking about when he thinks something in Mentalese . But it's far from obvious that an internal representation of Mentalese semantics needs to plays any role in the psychology of thinking. In the kind of referential-causal semantics we think is appropriate for Mentalese, It doesn't.

Data about how subjects construe English `pet fish' sentences make it clear that subjects use background information in interpreting tokens of them. But that hasn't any obvious bearing on whether the constituents of the subject's thoughts) are stereotypes. If that's right, then the experimental data that have generally been take to bear on what conceptual content is are simply beside the point: they invariably are data about how subjects interpret tokens of such forms of words as `pet fish', not data about how the conceptual constituents of their thoughts mentally represent pets or fish. Perhaps all that's going on in such experiments is that , when someone uses `pet fish' to say something about pet fish, it is likely to be typical pet fish (rather than stereotypic pet fish, or still less, each and every pet fish) that he intends to speak about. The moral such experiments point would then be that it's a mistake to confound a theory of communication with a theory of conceptual content. More on such matters in Chapter 3.

Digression

Concepts aren't the only kinds of mental representations; there are also thoughts (roughly, mentgal representations that express propositions). But, though our discussion has had many things to say about the one, it has thus far said very little about the other. That wasn't an accident. The tradition in cognitive psychology ---especially in associationistic cognitive psychology-- has consisted, in large part, of not noticing the concept/thought distinction; and it has suffered horrendous consequences for failing to do so. Before we proceed to consider `neural network' theories of conceptual content, we want to enumerate some of the differences between concepts and thoughts.

Suppose the content of one's concept WATER is (partially) constituted by the associative connection WATER → WET (that is, by the fact that mental tokens of the former regularly cause mental tokens of the latter.) Still, as Kant and Frege both very rightly emphasized, it wouldn't be possible to identify an event of thinking the thought water is wet with an event of first thinking the concept WATER and then (eg by association) thinking the concept WET. The reason it wouldn't is that, in the thought that water is wet, the property denoted by WET is predicated of the stuff denoted by WATER; `water is wet' has, as one says, the `logical form' of an attribution of wetness to water. Accordingly, that thought is true or false depending on whether the stuff in question (*viz,* water) has the property in question (viz being wet), which (putting ice cubes, steam and `powdered water to one side, it generally does.[19] By contrast,

[19] `The thought that water is wet predicates wetness to water' and `the thought that water is wet is true iff water is wet' may well just be two ways of saying the same thing. In any case, to speak of thoughts as having logical form is to invoke denotation (reference) and truth as parameters of their content. And, remember, the content of a thought is supposed to be inherited compositionally from the content of its constituents; so the fact that thoughts have logical form constrains facts about the content of concepts. All of that is overlooked if you think of the semantics of mental representations the way Empiricists and associationists practically always did; their theory was primarily a theory about the contents of thoughts, not about the content of thoughts. We've already seen why this was so: The content of one's thoughts is constituted not only by their conceptual constituents but also by the compositional structure of the relations of their constituents to one another. Confusing the structure concepts with the structure of thoughts gets you into a great deal of trouble, as we are about to see.

to associate WET with WATER is just to have tokenings of the concept WATER reliably cause tokenings of the concept WET. The bottom line is: EVEN IF THE CONTENT OF CONCEPTS

Associative ('semantic') Networks

Imagine a cognitive scientist who holds that the content of a concept is It's a position in a network of associations; perhaps he says that the concept DOG is the position in such a network that has the associative connections: Dog→ANIMAL,; DOG → BARK; DOG→BITE; DOG → CAT; DOG→DOGFOOD, and so on, Suppose he also holds that some such network of associative connections is (/determines) the content of the concept DOG. And suppose that, in his enthusiasm, he adds that the life of the mind ---in particular the thinking of thoughts--- consists in the excitation of such inter-conceptual associations. Such a theorist is what these days is called a `Connectionist'; and the point we're urging is that Connectionism, so construed, arrived pre-refuted. EVEN IF CONCEPTS ARE POSITIONS IN ASSOCIATIVE NETS, THINKING ISN'T THE ACTIVATION OF LINKS BETWEEN THE NODES IN SUCH NETWORKS. This is a problem for Connectionists because the question what conceptual content is and the question what thinking must certainly be very closely related: Concepts are the constituents of thoughts; they're what thoughts are made of, and thoughts get their semantics from t heir conceptual constituents (viz by composition.) If you go wrong about what concepts are, you also bound to go wrong about what thoughts are. And vice versa. (For a lot more on this topic see, Fodor & Pylyshyn, 1988)

We think that such plausibility as Connectionism may have depends on confounding associating with thinking. (And we think that Kant and Frege thought this too).

Still, it's conceivable that we're wrong that; so suppose, for the sake of the discussion, that we agree to 'bracket' thoughts. That being assumed, the residual question nonetheless can still be raised: `What can be said about the Associative Network theory of conceptual content if, turning our backs on what thinking is, we decide to just not worry about what a Connectionist theory of concepts would imply for a Connectionist theory of thinking. This question is important if only because, patently, Connectionism is just Empiricism with a degree in Computer Science. And, for kinds of reasons we've been raising (among others). For still more reasons see (Fodor & Pylyshyn, 1988), Empiricism is a debilitating kind of affliction.

So, theories about thoughts to one side, consider the question whether concepts are locations in a network of associations; hence whether one might represent a mind's `conceptual repertoire' --- the totality of concepts available to the mind at a given time, as a graph consisting of finitely many labeled nodes with paths connecting some of them to some others.[20] On the intended interpretation, the label on a node tells you which concept it stands for and the length of the lines between them varies inversely with the strength of the association between the concepts.

[20] `I thought you said that a mind at a given time has access to indefinitely many concepts . So how could a finite graph represent an indefinitely large conceptual repertoire?' Well, it depends on how you read `available'. For the purpose of discussing Connectionism/Associationism, it's easiest to take `the conceptual repertoire available to the mind at t' to mean something like the set of concepts that are constuents of thoughts that the mind has actually tokened up to and including t. The question how a Connectionists /Associationist might represent the indefinitely large set of concepts that would be constituents of thoughts if the mind could think all the thoughts that its conceptual repertoire allows for is generally not treated in the canonical Connectionist/Assciationst literature.

So, for example, imagine a mind that has the concept RED and the concept TRIANGLE but has never happens to think a thought about red triangles: Given it's conceptual repertoire, such a mind could think such a thought; but, for some reason, it happens not to have done so. Is the conceptual repertoire of that mind finite? Because Connectionists/Associationists don't ask this question; they haven't needed to face up to the fact that concepts are productive.

By convention, the lines in Connectionist graphs of conceptual networks express: relatively long ones correspond to relatively weak associations. Fig 1-1 is an (entirely hypothetical) representation of what might be a very small part of a Connectionist graph of the structure of the associative relations among concepts in someone' associative space'. It helps the exposition to assume (what is quite

---------- Fig. 1-1 about here; Fig 1-1 (to come) ----------

possibly not true) that the relation `associatively connected to' is transitive; so nodes may be connected by paths that only reach them via intermediate nodes. (c.f. the many discussions of `mediated association' in the psychological literature, REFERENCES the upshot of which seems to be that, if there are such associations, they are relatively rare and relatively weak.)

So, we're considering a theory that says the content of a labeled node is (or supervenes on) its connections. But its connections to what? If the theory is that the content of a labeled node is its connections to other labeled nodes. Since (again by convention) the label of a node expresses its content, are that would beg the very question that Connectionism claims to answer: `What determines that content (i.e. what label) a node has?' Rather, the idea must be that corresponding nodes in isomorphic graphs have the same content whatever the labels of their connected nodes may be. That would avoid the threatened circularity, but it surely can't be true. It is perfectly possible, for example, that the concept PUPPY has the same location on one graph as the concept KITTEN does on some isomorphic graph if that's the situation then , according to the present way of understanding the graphs, the KITTEN, node and the PUPPY node express the same content. So there's a dilemma: if the suggestion is that two nodes stand for the same contept iff they are connected to the same labeled nodes, then it's circular. But if the suggestion is that two nodes in isomorphic graphs stand the same concept iff they are connected to the corresponding nodes in isomorphic graphs, regardless of the labels of the nodes that they are connected to, then the suggestion is false (cf PUPPY and KITTEN). Quine was right when he warned (in his iconic article `Two dogmas of empiricism') that the familiar theories of meaning run in circles, hence that meaning should be viewed as a suspect notion for purposes of serious theory construction. That is a claim we will endorse in Chapter 3.

For all we know, it may be true that the node labeled A in one graph must have the same content as the node labeled B in another graph if they have the same associative connections to corresponding nodes in isomorphic graphs. But even if that were true, it would be of no use when the project is to explain what identity/difference of content is, because the notion of identity of labels just is the notion of identity of content, and a dog can't make a living by chasing its own tail. Since, as far as we know, Connectionists/Associationists have no other cards up their sleeves, they have no theory of conceptual content, very many advertisements to the contrary notwithstanding. It's our practice to begin each of our discussion of each theories of conceptual content, by enumerating its virtues. But we don't think that associationistic/connectionist accounts of content have any except that, since

associative and connective relations are both causal by definition, they have a head start over many other kinds of semantics in that they aspire to being naturalizable.[21]

Concepts as inferential roles.

The idea here is that the content of a concept is (or supervenes on) its inferential connections, where inferential connections are not assumed to be associative. Following the literature, we call this sort of proposal an `Inferential Role Semantics' (IRS). Versions of IRS are very widespread in current philosophical theories of meaning , where it is sometimes taken to accord with Wittgenstein's suggestion that meaning is use.

The modern history of IRS starts with Sellars' REFERENCE observation that the content of the ` `logical constants' can be specified by their roles in inference. AND, for example, is the concept whose inferential role is constituted by the `Introduction Rule' P,Q → P&Q and the `Elimination Rule' P & Q → P; P&Q →Q . It is, however, far from clear how the content of the concept AND might serve as a model for, say, the content of the concept TREE. In Classical versions of semantic theory, the content (i.e. intension) of a concept is what determines its extension, the set of things the concept applies to. But AND , unlike TREE, doesn't apply to anything; `the set of ands` is, to put it mildly, not well-defined. So it's a puzzle how the semantics of AND could provide a model of the semantics of TREE (or vice versa) ; in particular, how the content of TREE could be determined by the rules for its use in anything like the way that introduction/elimination rules might be thought to determine the content of AND. There is a tendency (very unfortunate, from our perspective) to force the analogy by saying that, in the case of TREE, the `rules of use' are procedures for applying the concept correctly ; viz for applying it to trees. So, on close examination, IRS often proves to be yet another kind of verificationism.

All the same, the idea of constructing a theory of content based on inferential relations rather than associative relations may seems prima facie plausible; on the face of it, association doesn't seem much like thinking; nobody could seriously hold that associating CAT with DOG is tantamount to thinking that cats are dogs since, unlike thoughts, associations are neither true nor false. There are also a number of other respects in which IRS might be able to cope things that give other kinds of semantic theories nightmares. Consider concepts that have empty extensions; these include , we suppose, not just AND (see above) but also SANTA CLAUSE, THE NEXT EVEN PRIME AFTER TWO, GOD, GHOST, SQUARE CRICLE and so on. But though they lack extensions, many empty concepts appear to have contents at least in the sense that there is a story about the sorts of things that they purport to apply to. And it's true, more or less, that people who have the concepts generally know the corresponding stories. Maybe, then, the contents of empty concepts can be identified with their inferential roles in the corresponding stories.

---

[21] One reason this sort of objection isn't more generally recognized is that Associationist/Connectist often turn Empiricist when the circularity threatens. See, for example, Churchland (19xx) REFERENCE) who suggests, in effect, that conceptual contents reduce to connectedness in conceptual space except for the content of primitive concepts, which is taken to be sensory. (See also the literature on `concept grounding' which quite often depends, in one way or other, on making the same mistake.) In this respect, much of contemporary cognitive science recapitulates Bishop Berkley. "Those who ignore history are doomed to repeat it" includes to the history of philosophy inter alia.

We think there is something right about this view of empty concepts, but that it offers no comfort to IRS. (More on this in Chapter 5).

For those and other reasons , IRS is often the semantic theory of choice for philosophers and cognitive psychologists who like idea that the content of concepts is somehow determined by their connectivity, but who understand that association can't be the kind of connectivity that's required. But, agreeable or otherwise, it won't do.

What's wrong with IRS? [22]

If you want to hold that conceptual content supervenes on inferential roles, you must find some way to say which contents supervene on which roles. There are, we think, only two principled options: You can say that every inference that a concept is involved in is constitutive of its content; or you can say that only some such inferences are. In the latter case, you are obliged to explain for each concept, the difference between the inferences that are constitutive and the inferences that aren't. We are convinced that both these options invite catastrophe; indeed, that both demand it.

The first option: *Holism*

Suppose that last Tuesday you saw a butterfly between your house and the one next door. Doing so adds a cluster (in fact, quite a large cluster) of new beliefs to the ones that you previously had: I saw a butterfly; I saw a butterfly between my house and the one next door; there was a butterfly visible from my house yesterday; there were butterflies around here yesterday; the total of my lifetime butterfly sightings has increased by one; there was something between my house and the one next door yesterday that probably wasn't there last January; if I hadn't been home yesterday, I likely would not have seen a butterfly ... and so forth, and so on and on. And each belief adds a corresponding new rule of inference to those that you were previously wont to apply in the course of your reasoning: If yesterday was the fourth of the month, infer I saw an insect on the fourth of the month; since butterflies are insects, infer that there was an insect visible to me the on the fourth of the month... and so on, depending on how far the effects that adding a new belief to one's stock of standing beliefs directly or alters the inferences to which one is committed. So, if holism is true, and every inference that a concept is involved is meaning constitutive, then the content of one's concepts alters as fast as one's beliefs do; which is to say, instant by instant.

A plethora of crazy consequences follow: Suppose yesterday you and your spouse agreed that butterflies are aesthetically pleasing; and suppose that one or other (but not both) of you comes to believe that he/she just saw a butterfly. Conceptual holism says that you can't now so much as think the proposition about which you used to agree about, since you no longer share the concept BUTTERFLY that you used to agree about it with. That sort of thing can be very hard on relationships.

The natural reply to this objection is, of course, that, though the content of your concepts (hence of your beliefs) changes instant by instant, it doesn't usually change very much. But how much is tht? And in what direction do ones concepts change when one's beliefs do? If the day before yesterday you believed that the Sun is a considerable distance from New Jersey, and if yesterday you came to believe that you saw a butterfly, what belief does the belief about the distance to the Sun change into? This is a morass, from whose boundaries no traveler has so far returned. We very strongly recommend that you stay out of it.

---

[22] For a more extended discussion of issues raised in this section, see Fodor and Lepore, Holism (19xx) [REF]

The second option: *analyticity*.

The truth values of propositions are connected to one another in all sorts of ways. For example, if P and P →Q are true, so too is Q. If this is a butterfly is true, so too is this is an insect. If my house is near your house is true, so too is your house is near my house. (On the other hand, even if my house is near your house , and your house is near John's house, my house may not be near John's house.) If this glass is full of H20 is true, so too is this glass is full of water. If all swans are white is a law of nature, then if this had been a swan, it would have been white is true. But nothing of the sort follows from all the swans I've seen so far are white. And so on.

For all sorts of reason, it is often of great interest to know which propositions have truth values that are interdependent; sometimes because our well-being nay rest on it, but often enough just because we're curious. Accordingly, one of the things our cognitive processes permit us to do is trace such connections; having gotten hold of one bit of the web of beliefs, we can follow it to other ones to which it's inferentially connected. We can, of course, never know about all of the inferential connections there are; nor would we conceivably wish to. But, from time to, we can know about some of the ones that our well-being depends on, or that our curiosity leads us to look into. That's what logic and science and mathematics and history are for. That's what thinking is for.

If we can't, even in principle, know all the connections among beliefs, maybe we can know all of the kinds of connections that there are? Or all of the important kinds? In effect, the Empiricist tradition made two epistemological suggestions about the structure of the web of beliefs, both of which have seemed plausible in their time and the first of which Chapters 4 and 5 will explore: that propositions whose truth values are accessible to perception have a special role to play in finding ones' way through the web; and that all propositions have their truth values either in virtue of their content alone or of their content together with facts about the world. The second iof these is of special epistemological interest because, if there are propositions that are true/false just in virtue of their content , and however the world is, there are at least some fixed points in the tangle of connections of which the web is constituted. If (or to the extent that) the content BACHELOR fixes the truth value of if John is a bachelor, John is unmarried , then we can always and everywhere rely on that inference being sound, whichever of our other beliefs we may have to alter. That's to say: if there are propositions that are true in virtue of their meanings alone, and if the content of a proposition is its inferential role, then holism is false.

Call beliefs whose content doesn't depend on one's other beliefs `analytic'. Convenient as a substantive notion of analyticity might be as an antidote to holism, there are an increasing number of reasons why most philosophers have ceased to believe that such a notion can be sustained.

1. The first worry is that, by assumption, analytic beliefs can't be revised without changing the content o (some or all of) their conceptual constituents; i.e. they can't be changed without equivocating. But, there don't seem to be any beliefs that one can't (reasonably) revise under sufficient pressure from data and background theories. The trouble is that belief change is conservative: If enough rests on a belief, and if there is some replacement for it waiting in the wings, then any belief may be rationally abandoned; even the ones that are allegedly analytic.
2. The second worry is that you can't use the notion of analyticity to explicate the notion of meaning, content, intension, or any of the others that that are central to semantics, on pain of begging the questions that semantics is supposed to answer; otherwise, since analyticity is itself a semantic notion par excellence, you end up in a circle.

Both these lines of argument were spelled out in Quine's iconic paper `Two dogmas of empiricism'; (reference) nor, to our knowledge, has either been seriously rebutted.

The moral of this chapter is that all the available accounts of conceptual content (or, anyhow, all the ones we've heard of) seem to be pretty clearly not viable; we think that those who to cling them do so mostly in despair. At a minimum, it seems to us that the arguments against the available theories of content are sufficiently impressive that it would be unwise to take senses, intensions, or the like for granted in any semantic theory of content that you care about, theories of the semantics of mental representations very much included. So, what now?

***

Why , after all these years, have we still not caught sight of the Loch Ness Monster? Of course, it might be that we've been looking in the wrong places. We're told that Loch Ness is very large, very deep, and very wet; and we believe all of that. But, as the failures pile up, an alternative explanation suggests itself: The reason we haven't found the LNM is that there is no such beast.

Likewise, we think, in semantics: The reason that nobody has found anything that can bear the weight that meaning has been supposed to bear ---it determines extensions, it is preserved under translation and paraphrase, it is transmitted in successful communication, it is what synonymy is the identity of, it supports a semantic notion of necessity, it supports philosophically interesting notions of analytic necessity and conceptual analysis, it is psychologically real, it distinguishes among coextensive concepts (including empty ones), it is compositional, it is productive, it isn't occult, and, even if it doesn't meet quite all of those criteria, it does meet a substantial number--- the reason that meaning has proved so elusive is that there is no such beast as that either. We think that, Like the Loch Ness Monster, meaning is a myth.

The rest of the book is about how it might be possible to construct a semantics for mental representations that is reasonably plausible, sufficient for the purposes of cognitive science, and compatible with naturalistic constraints on empirical explanations, but which dispenses with the notion of meaning altogether: It recognizes reference as the only relevant factor of content. Accordingly, although there are plenty of extensions, they aren't determined by intensions. We don't claim to know for sure that any such theory is viable; and, even assuming that some or other is, we don't claim to know, in any detail, how to arrive at it. But maybe we can point in the right general direction.

APPENDIX: Semantic pragmatism[23]

We have disagreed, in one way or another with each of the theories of conceptual content discussed in this chapter. But they all share a premise that we fully endorse: Concepts are identified, at least in part, by their relations to thoughts. We think this is true in a number of respects: Concepts function as the constituents of thoughts; the same concept (type) can be a constituent of indefinitely many different thoughts; minds that share a concept may disagree about the truth (/falsity) of indefinitely many of the thoughts of which that concept is a constituent, and so on. One might suppose that, if anything in semantic theory is `true by definition', that is. So why do so many philosophers (and so many others) reject it?

At least since Wittgenstein advised that we ask for the use of symbol s rather than their meanings, semanticists have often been attracted by the more or less Pragmatist thesis that there is an intrinsic connection between a concept's content and its role in the integration/causation of behavior, Prinz and Clark (henceforth P&C) put it this way: to `sever… putatively constitutive links between thinking and any kind of doing is … a doctrine that we should not seriously contemplate, on pain of losing our grip on what concepts are for, and why we bother to ascribe them (57)." We don't know whether this anxiety is justified because, of course, nobody ever has suggested the severing of concepts from actions; certainly not the present authors. What has been widely suggested ---and what we think is true--- is that the connection between concepts and actions is invariably contingent; in particular, it is not constitutive of conceptual content . Rather, the connections between a the concepts that a creature's has and the behavior it performs is invariably mediated by what ithe creature thinks and what it wants . Ever since Aristotle invented the `practical syllogism' (see above), and excepting only behaviorists, the consensus view has been that what creatures do depends on their beliefs and desires.

That view is, of course, fully compatible with taking concepts to be individuated by reference to their roles as constituents of its propositional attitudes: You can't want to eat an apple unless you have the concept APPLE. Representational theories of mind say that's because wanting to eat an apple involves being related, in the appropriate way, to a mental representation of how things would be if you got what you want. Notice that that claim is not the one that P&C (or other neopragmatists) dispute. Rather Pragmatism is about whether the concept APPLE is somehow constituted by its role in "action-oriented" behaviors, among which eating apples would presumably be included. According to Aristotle, that's because lots of people who have the concept APPLE believe that they are good to eat; were it not that they think apples are edible, they would cease to eat them. Notice, however, that, Aristotle's view, unlike the pragmatist's, it is extremely plausible that one's APPLE concept would survive abandoning one's belief that apples are edible , even though apple-eating- behaviors would not. That is part of the explanation of why an argument between someone who thinks that apples are poisonous and someone who doesn't , isn't, as one says, `merely verbal'. If you think eating apples makes one sick and I don't, our disagreement isn't about the concept APPLE (or the word `apple'); it's about whether eating apples makes one sick. It was a discovery that apples are good to eat, not a semantic stipulation. Ask Eve.

Now P&C have, of course, every right not to believe that. They might hold --- perhaps on Quine-Duhem grounds--- that there just isn't any principled way to distinguish what one knows in virtue of having the concept APPLE from what one knows in virtue of knowing what apples are good for. Our

[23] This a reply to Prinz and Clark, *Mind And Language 19*, 1 Feb (2004), which is in turn a reply to `Fodor, *Language 19*, 1 Feb (2004).

point, however, is that you can't (and Quine-Duhem don't) defend such claims by saying that unless EDIBLE is constitutive of APPLE, the connection between thinking about apples and eating them is somehow `severed'. One doesn't need to turn pragmatist is avoid severing such connections; all one need is to do is hold that they are causal (hence contingent) rather than semantic (hence necessary).

We do agree that there's a problem about how to distinguish beliefs about the intensions of concepts from beliefs about the things in their extensions. As will presently become apparent , that's one of the reasons we think that individuating concepts by their intensions is a mistake. But, whether or not we're right about that, pragmatism doesn't help to solve the problem of concept individuation since if identifying concepts by their relation to thoughts would sever their connection to actions, identifying concepts by their relations to actions would sever their connection to thoughts; and severing the connections between concepts and thoughts would be a disaster since concepts are the constituents of thoughts. The concepts APPLE is a constituent both of the thought that apples are good to eat and of the thought that they aren't.

Apple-thoughts are, quite generally connected to apple-actions by more or less complicated chains of inference, among the premises of which there are, almost invariably, some that are contingent. (Among these, in the present case, is the proposition that eating apples makes you sick.) But how could that be so if, as pragmatists hold, the connection between conceptual contents and actions is constitutive of the concept's identity? P&C never do try to cope with this (perfectly standard) kind of objection to pragmatism. Instead they suggest that we consider the concept of a 50 pound egg. Very well, then; let's.

P&C tell us that, unless we keep it in mind that the mothers of 50 pound eggs are sure to be very large and likely to be very fierce, we are at risk of having serious trouble with the creatures that lay them . But though that is certainly true, we don't think that keeping it in mind is a condition for grasping the concept of a 50 POUND EGG. Rather, we think that, the content of the concept 50 POUND EGG is just: egg that weighs 50 pounds. This is compatible demanded by the view, which we likewise endorse that (idioms and like aside),that the content of complex concepts determined, by composition from the content of their constituents. So, which of its constituent concepts do P&C suppose contributes the content dangerous to the content of the concept MOTHER OF A 50 POUND EGG? Is it MOTHER, perhaps?

What leads us to give creatures that lay 50 pound eggs a wide berth is: putting the content of the concept together with what we know (or surmise) about a 50 pound egg's likely parentage (and with a lot of other stuff as well). If that weren't so, we wouldn't be able to so much as contemplate the possibility that mothers of such very large eggs have been the victims of slander and are, as a matter of fact, as gentle as lambs. If that isn't at least conceptually possible, what on earth could Sendak have been up to?

So why, in virtue of all the old questions that it fails to solve and all the new questions that it gratuitously invites, are there so many converts to semantic pragmatism both in philosophy and in cognitive science? To be sure, pragmatism shares with Marlboro adds a certain macho air of cutting through the frills and getting to the bare bones of doing things. ("In the real world thinking is always and everywhere about doing." P&C say) But, on reflection, that strikes us as pretty obviously untrue and, in any case, unduly dogmatic. Who knows what, if anything , thinking is always about? Or indeed, whether there is anything that thinking is always about? And anyhow, to the best of our knowledge, `The Real World' is just the world; people who rest a lot on a presumed difference between the two are almost always just bluffing.

References cited in Chapter 2

Bruner, J. S., Goodnow, J. J., & Austen, G. A. (1986 / 1956). *A Study of Thinking*. New Brunswick, NJ: Transaction Books.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*, 3-71.

Kosslyn, S. M. (1975). The Information Represented in Visual Images. *Cognitive Psychology, 7*, 341-370.

Kosslyn, S. M. (1994). *Image and Brain: The resolution of the imagery debate*. Cambridge. MA: MIT Press.

Levesque, H. J., & Brachman, R. J. (1985). A fundamental tradeoff in knowledge representation and reasoning (revised version). In H. J. Levesque & R. J. Brachman (Eds.), *Readings in Knowledge Representation* (pp. 41-70). Los Altos, CA: Morgan Kaufmann Publishers.

Lindberg, D. C. (1976). *Theories of vision from al-Kindi to Kepler*. Chicago: University of Chicago Press.

Paivio, A. (1971). *Imagery and Verbal Processes*. New York: Holt, Reinhart, and Winston.

Pylyshyn, Z. W. (1973). What the Mind's Eye Tells the Mind's Brain: A Critique of Mental Imagery. *Psychological Bulletin, 80*, 1-24.

Pylyshyn, Z. W. (2003). *Seeing and visualizing: It's not what you think*. Cambridge, MA: MIT Press/Bradford Books.

Pylyshyn, Z. W. (2007). *Things and Places: How the mind connects with the world (Jean Nicod Lecture Series)*. Cambridge, MA: MIT Press.

Rosch, E. H., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*, 382-439.

This page deliberately left blank

Chapter 3 Contrarian Semantics

As we understand the jargon of Wall Street, a `contrarian' is someone whose practice is to buy when everybody else sells and to sell when everybody else buys. In effect, contrarians believe that what more or less everybody else believes is false more or less all of the time. This book is an essay in contrarian semantics. Chapter 2 reviewed a spectrum of theories about the content of concepts (i.e. the intensions of concepts; the meanings of concepts; the senses of concepts ….etc) which, though they differ in all sorts of other ways, all agree that the content of a concept is whatever determines its extension: Concepts C and C' have the same content only if (C applies to a thing iff C' does.)

However, sticky questions arise at once. For example, is the content of concept C to be defined relative to the things that It *actually* applies to or to the things that it *would* apply to if there were such things? Philosophers love this kind of question; but in cognitive science they aren't often explicitly discussed. Our impression, for what it's worth, is that usually (but not always), it is the `modal' reading that cog sci has In mind. Assume, for illustration, that there are no green cats. Still, the question arises whether the extension of CAT would include a green domestic feline if there chanced to be one. If your view is that the content of a concept is its *definition,* then your answer is `yes'. If, however, your view is that concepts are stereotypes, the answer is less clear: It would depend on (among other things) the ratio of green cats to cats. Note that only *actual* cats need apply since there is, presumably, no bound to the number of *possible* green cats that there could be if there could be any. We think this disagreement about modal force is the heart of the difference between stereotype theories of conceptual content and definition theories, but that's not why we raise the issue here; rather, it's to illustrate the kinds of difficulties that arise in deciding just what Intensionalism amounts to; what exactly it is that the friends of intensions are required to assert and their antagonists are required to deny .

We propose to stipulate: Intensions are defined by their relation to *extensions*; in particular, by the following principles:

1. Intensions *determine* extensions; that is, intensionally identical representations are ipso facto coextensive.
2. If coexstensive representations are not semantically equivalent, they must be intensionally distinct.

This chapter will argue that there is nothing that satisfies both i and ii, hence that there are no such things as intensions.[1] Now, in the cognitive science community at least, this view is not just contrarian, it's heretical; it's simply taken for granted that intensions, (meanings, senses, etc), satisfy (i) and (ii) above. There is, however some indication, that the tide has turned in Anglophone philosophy, where versions of `causal theories' of reference are currently fashionable. We will ourselves offer one in Chapters 4 and 5; but, for all sorts of reasons, it differs from the kinds that are in vogue at present. (For an Anglophone philosopher who cleaves to the `intensions determines extensions' tradition, see, eg., Frank Jackson) (reference)

We start with what we will call `Frege arguments' (though we admit they may have at most a tenuous connection to arguments that the historical Frege actually endorsed).

---

[1] We won't, however, discuss the kind of intensionalist theories that identify conceptual contents with `sets of possible worlds'. That's because, naturalism wants conceptual contents to affect causal powers, and we can't imagine how sets of merely possible worlds could do that. *Merely* possible things don't make anything happen.

Frege arguments

The paradigm is: `If meaning didn't determine reference, then co*ex*tensive words/concepts would be co*in*tensive (in particular, coreferential words would be synonyms). But there are, according to several standard and reasonable criteria, coextensive words that are *not* synonyms. So the content of a word must consist of something more (or other) than its extension; and its intension is the only other option The same for concepts, *mutatis mutandis.*

We want to meet the Frege arguments head on, so we'll just grant that, for example, `George Washington' (hereinafter GW) and `Out First President' (hereinafter OFP), though coextensive, *aren't* synonyms. What shows they aren't is a line of argument that Frege relies on very heavily: Substitution of one synonym for another preserves the truth/ falsity of sentences and thoughts that contain them (in the philosophical jargon, they substitute `salve veritate' .) But substitution of coextensives need not preserve truth/falsity in propositional attitude (PA) contexts. 'John doesn't believe that GW was OFP' does *not* imply `John doesn't believe that GW was GW'. So coextensive aren't, per se, synonyms.

Fair enough; let's suppose that such considerations show that coextension isn't sufficient for identity of content. The question is: what follows from that? In particular, how does it show that there are intensions?

There is, to begin with, more than a whiff of missing premise here. To make this sort of argument work, an intensionalist must assume not only that `GW' and `OFP' are coextensive (to which historical consensus testifies) but also that *if it isn't coextension that guarantee synonymy,* (hence substitution *salve veritate) it's got to be cointension that does.* But that follows only if it's also assumed that coextension and cointension are the only candidates for identification with conceptual content. What entitles intensionalists (Frege included) to assume that?

We think that Frege's sorts of examples *do* show that *something other than mere coextensivity* is required for a viable notion of conceptual (or lexical) content; but why concede that co*intension* is the mandatory alternative? This question is pressing if you think, as we do, that there is no plausible account of what an intension *is*; and (to repeat) that even if there were, nobody has the slightest idea of how intension, synonymy and the like might be naturalized.[2] (Nauralizability, as the reader may recall, was among the `working assumptions` we set out in chapter 1).

Consider, once again, John, who believes that GW was GW but does not believe that GW was OFP, so that, if you ask John who GW was (and if he is feeling compliant) he will confidently and sincerely and correctly answer that GW was GW; whereas, if you ask him who OFP was, he will confidently and sincerely and correctly answer that he doesn't know. It seems that, if coextension implies cointension we must either hold that the substitution of cointensive terms does affect truth value, or that (contrary to a claim to which naturalistic PA (Propositional Attitude) psychology is surely committed one's beliefs aren't causes of one's behavior, question-answering behavior included. Prima facie, if Frege's sorts of examples show what intensionalists think they do, then they show either that PA psychology isn't true or that, if it is, it can't be naturalized.

If we read him right, Frege himself thought something like this: John doesn't believe that GW was OFP *when he (GW) is* described as (when he is `presented under the guise' Our First President. But John *does* believe that GW was OFP when he (GW again) is described as GW. That's a perfectly coherent view

[2] It bears mention that if, as Quine argued, there is a circle of interdefinition among notions like *intension, meaning, definition* etc, then *extension* is in that circle too. According to the usual understanding, the extension of a representation is the set of things that it applies to. But what could `applies to' itself mean if w what it means isn't something semantic?

as far as it goes; but the naturalization of belief/desire psychology requires that believing should be understood in terms of *nonsemantic* notions, and `satisfied` and `described as' and the like don't qualify. *Describing something as* such and such is, surely, a species of *saying that it is such and such.* And `saying that' is quite as semantic as `believing that`, So we're in a circle of interdefinitions again. So belief-desire psychology can't be naturalized.

The sound you now hear is that of indefinitely many hard-headed cognitive scientists replying that they couldn't care less who wins these word games. What they care about is, for example, how (or whether) John's beliefs cause his behavior. One can't but sympathize; but it just won't do. John does *say* that he doesn't know who OFP was, and we're taking it that he says so sincerely. Nor is this a case of, for example, behavior being driven by *unconscious* beliefs in the way that verbal slips are sometimes thought to be. So, if coextension entails cointension, John really *does* believe that GW = OFP; and his sincerely saying that he *doesn't* is a prima facie counterexample to BD Psychology in a way that his verbal slips are not. Looked at this way, the problem that John raises for belief-desire psychology isn't that it is forced to say that the belief that GW was FOC is in John's head but somehow unavailable to his introspection. It's that , for all either his introspection or his behavior shows, the belief that GW was OFP *just wasn't in there*, either consciously or unconsciously. Even analyzing his dreams won't bring it to light. For the purposes of psychological explanation, it might as well not be there.

Frege thought the right thing to say is that John believes that GW was OFP when GW is described as OFP, but not when GW is described as GW; a view that (as we mentioned above) is nonnaturalistic in that it relies on semantic notions (*described as*) to explicate a psychological notion (*believe that).* Frege doesn't mind that because he very much isn't a naturalist. He takes principles like `descriptions that are coextensive need not be synonyms' to be, as it were, at metaphysical ground level. Descriptions *just are* the kinds of things that can be coextensive but not cointensive; there is nothing more to be said about that without resort to still further semantic vocabulary. But, of course, hard-headed cognitive scientist are ipso facto naturalists, and so are we. So *now* what ?

The story we like goes something like this: the form of words (1) is ambiguous. It has both the reading 2 and the reading 3. The point of present interest is that, although reading 2 makes 1 false, reading 3 makes 1 *true*. So, on the one hand, John *doesn't* believe the proposition *GW was FOC*; but, on the other hand, there's a reasonably robust sense in which he *does* believe the propositkion that GW was FOC, even though he sincerely denies that he does. After all, John believes the proposition that GW was GW; and, in point of historical fact, GW was the FOC.

1. J believes *GW* was the OFP.

2 J believes of GW (when described as GW) that he was OFP.

3. GW is such that, when GW is described as OFP, J believes of him that he was OFP.

4. *GW was OFP*.

5 It's GW that J believes was OFP

From the logico-syntactic point of view this is all unsurprising; what has happed is just that` GW` is moved from its `opaque` position *inside* the scope of `believes' in 2 to a `transparent' position *outside* the scope of `believes 'r in (3). It's being in the former position renders `GW' inferentially inert in 2 and its being in the latter position renders `GW' inferentially active in (3). So the substitution of coextensives is valid outside the scope of PA operators but not inside the scope of PA operators. (Frege thinks that's because terms in opaque contexts denote not their extensions but their *in*tensions. But that issue needn't concern us for present purposes.)

OK so far. Indeed, there's plenty of precedent. Compare *quoted* expressions: Suppose J *uttered* 5; and suppose that, as a matter of fact, but *unknown to J*, the man with the Martini was Abe Lincoln. Then it's unsurprising that J can perfectly sincerely, *though falsely*, deny that he said that Abe Lincoln was drinking a Martini. One way to put it is that, although J does believe that the Martini-man was Abe, he *doesn't know that he does.* That's not a paradox. `Abe is drinking a Martini' follows from `Abe *is the Martini-man*` only given a pertinent fact *to which J. isn't privy*; namely that the man with the Martini is Abe, not GW.

So , everything is fine: John 's denying that he believes (/said) that Abe is the drinker isn't, after all, a counter example to the very plausible claim that one's beliefs drive one's behavior. What the example does show, however, is the importance of distinguishing between *J believes Abe is the Martini man when Abe is so described* (which is true) and *J believes that Abe is the Martini man when Abe is described as, for example GW.* Or, to put it still differently, the example shows that, in one sense of `what you believe', some of what you believe depends not only on `what's going on in your head' but also on `what's going on in the world`. This theme will recur later in the discussion

But though we think that some such story is likely true, we are committed naturalists , so there isn't any story of that sort that we are allowed to tell. As previously remarked, naturalists aren't allowed to use semantic vocabulary in their construal of PA explanations; and, of course, `proposition`, `true' and `so described' are all semantic terms par excellence, So the circularity/non-naturalizability dilemma persists in persisting: We need a naturalistic and *uncircular* construal of what John believes/says and the like, but we haven't got one. So *now* what?

Our proposal traces back at least to Carnap (references)[3]. Here's the basic idea of Carnap's treatment: One should take very seriously the observation that names, descriptions and the like, when in the scope of expressions like `so described', `as such`, `under the guise', `under the mode of presentation….' and the like, work in much the way that quoted expressions do. To say that J believes that GW was OFP when GW is so described would be to say something like: J believes that tokens of the representation type `GW was OFP' are true; whereas, to say that J believes that GW was GW is to say something like: J believes that tokens of the representation type `GW was GW' are true. In effect, Carnap's idea is that quoted linguistic expressions might do much the same work that propositions are traditionally supposed do in explicating semantic notions like *conceptual content, belief etc*. From the point of view of a logician whom the inferential inertness of PA contexts would otherwise cause to lose sleep, that suggestion might well seem plausible.

Notice, in particular, that if Carnap is right, the `GW' in 4 refers *not* to the father of our country but to the *form of words* `OFP', which is the one that is used (in English) to refer to the Father of Our Country. So if what 1 mea ns, more or less what that 4 says, then it's unsurprising that 4 doesn't imply 5, *even though GW was in point of fact OFP*.

5. John believes that GW was OFP.

Expressions inside quotation marks don't refer to what the same expression do when the quotation marks are removed.

---

[3] Though we don't at all suppose that Carnap would approve of the way we use his sort of view. As we read Carnap he was a *Realist about PAs*, but he was some sort of *reductionist* about *PA explanations*, which he thought could be dispensed with in favor of explanations in brain science. We are *non-eliminative* Realists both about PAs *and* about PA explanations; in particular, we're morally certain that propositional attitude explanations don't reduce to anything that can be said in the vocabulary of brain science.

We're fond of Carnap's answer to Frege's puzzle, because it invokes the independently plausible thesis that talking about beliefs is interestingly like talking about linguistic expressions. Both thoughts and sentences are, after all, systematic, productive and compositional; and the similarity between, on the one hand, the relation of thoughts to their constituent concepts and, on the other hand, the relation of sentences to their constituent words, is surely too striking to ignore. So, just as the LOT version of RTM argues for the assimilation of mental content to linguistic content, so Carnap's account of the logic of PA contexts argues for their assimilation to quotations. Whichever way you approach it, thinking seems a lot like talking. That's perhaps unsurprising too since we use language to say what we think. This is the way things *ought* to work in empirical theory construction: everybody takes in everybody else's wash. No stick can hold itself up, but every stick helps hold up every other.

But, in principle, Carnap can do without the story about PA contexts working like quotation contexts. Technical details to one side, *any* cognitive theory that embraces RTM --- any cognitive theory according to which the relation between minds and propositions is mediated by mental representations--- can resolve Frege's problem in much the same way that Carnap does, *whether or not* it assimilates the inferential inertness of PA contexts to that of quotation contexts. The basic point is that Frege's sorts of examples don't show what Frege thought they do: that thoughts and concepts must have intensions. The most they show is that their extensions can't be the *only* parameter along which thoughts and concepts are individuated; there must be at least one other.

Frege just took for granted that, since coextensive thoughts(/concepts) can be distinct, it must be differences in their intensions that distinguish among them. But RTM, in whatever form, suggests another possibility: thoughts and concepts are individuated by their extensions *together with their vehicles.* The concepts THE MORNING STAR and THE EVENING STAR are distinct *because the corresponding mental representations are distinct.* That must be so since the mental representation that expresses the concept THE MORNING STAR has a constituent that expresses the concept MORNING, but the mental representation that expresses the concept THE EVENING STAR does not. That's why nobody can have the concept THE EVENING STAR who doesn't have the concept MORNING and nobody can have the concept THE EVENING STAR who doesn't have the concept EVENING

Likewise, the thought that *Cicero was fat* and the thought that *Tully was fat* are distinct because the corresponding mental representations contain different names of Tully (i.e of Cicero). Carnap's assimilation of PA contexts to quotation contexts is, to be sure, a special case of that sort of explanation; but he can do without it so long as he cleaves to some or other version of RTM. The serious question about inferential inertness isn't whether PA contexts are a species of quotation contexts; it's whether you need *differences among intensions* to individuate concepts (/thoughts), or whether differences between mental representations would do the job instead.

Our next point is, even less encouraging to Carnap's suggestion: Even if the quotation story is OK with a logician, a psychologist might well refuse to swallow it. The problem is that believing that a certain token of the sentence type `Cicero was fat' is true will have quite different psychological effects depending on whether or not it's an English speaker who believes it. [4]

That being so, however, the suggestion that that belief is a relation between persons and propositions that is mediated by a Language of Thought seems to imply not just that anybody who can think that Cicero was fat must have a language of thought, but that they must all have *the same* one; which is indeed a lot to swallow. This isn't, of course an argument against LOT per se. Suppose MRs are mental images; then the image you have when you think P and the image that I have when I think P

[4] You can, of course, perfectly well believe that the sentence `Cicero was fat' is true even if you have no idea what it means. Someone points to a token of that sentence and says `that's true'.

must be more similar to one another than either is any image that you form when either of us thinks Q. The consequence that the sameness of thoughts requires the sameness of the corresponding MRs holds not just for LOT but for *all* versions of RTM that could be of any use to psychology.

It seems to us that this is a parting of the ways; at a minimum, it's a nasty snarl of options. If you accept the suggestion that differences among the vehicles of thoughts (rather differences between their intensions) explain why referentially equivalent PAs can differ in truth value, then you will have to hold either:

1. that everyone who can think the thought that Cicero was fat uses the same mental representation to do so as (hence that if you think in German and I think in English, then at most one of us can think that Cicero was fat); Or
2. that, quite possibly, we can't all think that thought (or, indeed any other)thought.
   (Mix and match combinations of 1&2 two might also be considered.)

If you opt for 2. then if mental representations are expressions in the Language of Thought, it's wide open that each us thinks in a different LOT, and LOT is, in a reasonable sense, a private language. Now, it is notorious that Wittgenstein thought he had proved that there can't be such a thing as a private language; and lots of philosophers have agreed that indeed there can't. But the consensus is, perhaps, less striking than it seems at first since there isn't much agreement about what it takes for a language to be private. One plausible reading is that Wittgenstein denied the possibility of a language that isn't shared by a social group (Kripke apparently reads the `Private Language Argument' that way REFERENCE)) If that is the right interpretation of Wittgenstein's `Private Language Argument (PLA)' then to endorse it is, at a minimum, to hold that being a vehicle of interpersonal communication is an essential condition for being a vehicle of thought, and PLA would be an a priori disproof of what we take to be the best current hope for a cognitive psychology.

But, anyhow, we were stuck with private languages (so understood) as soon as we said that there is no such thing as meaning. The idea that we can all think that Cicero was fat requires, at a minimum, that some of your thoughts translate into some of mine; and, since it's *meaning* that translation is supposed to preserve, if there is no meaning, then there is no translation. And, while we're at it: if there is no meaning, then there is no communication, no paraphrase, no abridgment, no plagiarism and so forth for *any* of the cognitive process or relations of which meaning preservation is taken to be a defining property.) And if thinking in a laguage of which no communication, none of these is possible isn't thinking in a private language, it will surely do for the purposes Wittgenstein has in mind. Well, so be it.

We aren't, of course, denying that --- in a rough, ready, and commonsensical sort of way, translation, communication, paraphrase, thoughts, as well as abridgement, synopsis, plagiarism of thoughts happen all the time. Of course they do. Rather, our suggestion is that none of these is the name of a 'natural kind' of semantic relation;. In which case, if any version of RTM is true. there couldn't be a *theory* of the translation, communication, abridgement, synopsis etc, of thoughts in anything like the way in which there are theories of photosynthesis, or glacial drift, or (maybe) evolution. In fact, we rather think that, deep down, everybody agrees that even for languages like English, to say nothing of LOT. Nobody *really* thinks that there's a strict matter of fact about what is the `right' translation of a certain text; or whether two expressions are, strictly speaking, synonyms; or which of two metaphors is the better one; or which of two jokes is the funnier one; or whether a paraphrase misses something that's essential to a text; or what, precisely, is the message transmitted in a certain communication. Such decisions are endlessly sensitive to contextual detail; often enough they are matters of taste. (Hamlet told Ophelia to get herself to a nunnery. That puzzled Ophelia,and critics have been arguing about what he meant for five hundred years or so. The end still is not in sight.) Quite probably you *can't*

say what a good metaphor, or translation, or synopsis is in any vocabulary that's not already up to its ears in semantic notions (unlike, for example, the vocabulary of neural science, or of biochemistry.) Compare E.O. Wilson's remarkably naïve suggestion that science and criticism might eventually team up to answer such questions once and for all.) REFERENCE.

If we're right about all that, then it's not just theories of conceptual content to which notions like translation and meaning are unavailable; they aren't available for the purposes of *any* strictly scientific and naturalistic enterprise. A serious and naturalistic theory of content might tell us what anaphora is; or what, if anything, generic nouns denote, or whether `the' is a quantifier, or maybe even what the sufficient conditions for reference are; indeed, we think that the vindication of cognitive science requires that a serious and naturalistic theory of content do at least the third. We don't, however, expect it to tell us what a good translation of Pride and Prejudice is, or even which of two translation Pride and Prejudice the better. The moral is: once you've given up on intensions and the like, you might as well sign up for Private Languages; also, you as well cease to believe that many of the purposes that theories of language have been are serious goals for linguistics of cog sci. That might show that there really *must* be intentions. Or it might only show that quite a lot of linguistics and cog sci has been been barking up the wrong tree.

Where does all that leave us? There are think there are three options:

a) Give up on RTM (hence on claiming that the semantic content of thoughts and concepts are pairs of their vehicles and their extensions.)

b) Give up on the chance of a serious account of communication, paraphrase, translation and the like and suppose that, strictly speaking. The only kind of cognitive psychology there can be is *individual* psychology (much as many linguists think that the only grammars there can be are grammars of idiolects).

c) Take it seriously that we all think in much t he same Language of Thought. That we do is not, of course true a priori, but it might well be true de facto. After all, if we all think in a mental language, it can't, on pain of circularity, be one that we've learned, it must be a product of the innate neural architecture of our brains. And we do all have very similar sorts of brains (compared, anyhow, with the brains that any other kinds of creatures have).

Those are, as we can see, the available options. Given what we take to be the really hopeless failure of cognitive science to devise a remotely plausible account of conceptual content, we can't in good conscience recommend that you endorse the first. But, at least for present purposes, we'll settle for either of the other two

What we've just been saying is intended to reply to a non-Fregian argument for the claim that mental (/linguistic) content has to be more than just referential; namely, that all sorts of notions that are routinely (and indispensably) used in talking about content aren't easily translated into talk about reference; translation itself is among them. But, as we've just been seeing, that doesn't show that there are such things as intensions or that naturalism can't be sustained. So far as we can tell, there aren't any a serious arguments against the possibility of a naturalistic science of mind, so long as the semantics it assumes is referential. On the other hand, even meaning nihilists (like us) should grant that translation, communication, paraphrase and other like notions are essential for workaday purposes, and that there is something other than reference that is preserved in all of them. Identity of extension together with identity of vehicles is the best candidate that we've been able to think of so far.

Other arguments against referentialism

There is a scattering of other arguments that have, from time to time, been said to refute the suggestion that conceptual content is purely referential. None of these goes as deep as Frege's observation that expressions in PA contexts are characteristically inferentially inert; in particular, that substitution of otherwise co-referential expressions isn't reliably truth-preserving in such contexts. But some of them have nevertheless seemed persuasive. We're about to survey a couple of alternatives to Frege's sort of polemic against naturalistic theories of mind.[5]

`Empty` concepts

There is no Devil and there are no unicorns; so one might reasonably claim that the concepts THE DEVIL and THE ONLY UNCORN IN CAPTIVITY are coextensive; both denote the empty set. And so too, mutatis mutandis, do the corresponding English expressions. But (so we suppose) THE DEVIL and THE ONLY UNICORN are *different* concepts; nor are the expressions `the devil' and `the only unicorn' synonymous. So, once again the conclusion: there must be more to conceptual content (/lexical meaning) than extension.

Now, we might complain that, strictly speaking, all of that is tendentious in one way or another. The inference from '`The Devil' has no extension' to 'The Devil' refers to the empty set' is too quick. The right thing to say about `the Devil' isn't, perhaps, that it refers to the empty set, but that it doesn't refer at all. `Still,` you might reply, `the devil has no manners' isn't meaningless, and it isn't a synonym of `the only unicorn has no manners', both of which it ought to be if meaning is reference.' To which we might then reply: `We didn't say that meaning is reference; we said that there is no such thing as meaning.' If you think this rejoinder is frivolous, we repeat: there *is* more to content than reference; but that doesn't, in and of itself, argue that part (still less all) of content is meaning. ' And, as we've been seeoing,l , if there isn't any such thing as meaning, there can't be any such thing as synonymy since synonymy is, by definition, identity of meaning;. A fortiori. there is no such thing as the synonymy of DEVIL and UNICORN'. So who wins this sort of argument? Or is it a draw?

We don't think this question is `merely philosophical'. But we can imagine that a hard-headed cognitive scientist might not agree. What's true, in any case, is that we have yet to explain why, even if there is no synonymy, there are more or less reliable `intuitions of synonymy'. If those aren't really intuitions of identity of meaning, what are they intuitions of? [6] It's sometimes suggested, in aid of `Inferential Role' accounts of meaning (see ch 1) that, at a minimum, they can explain how there could be differences in content between referentially empty concepts.

Presumably anybody who knows what `is a unicorn' means should be willing to infer that if there aren't any unicorns, then there aren't any unicorns' horns. But, because it's possible to believe in unicorns but not in The Devil, it's possible for someone who knows what `devil' and `unicorn' mean and is prepared to infer from `no unicorns' to `no unicorns' horns' to nevertheless refuse to infer from `no

---

[5] The reader may wish to suggest that, on closer consideration, that some or all of these `alternatives' are, in fact, themselves instances of Frege-arguments. That may be so; for our present purposes, it doesn't matter.

[6] Actually, the `inter-judge reliability` of intuitions about identity/difference of content isn't self-evident. Linguists have been at daggers drawn for years about whether or not `John killed Bill' and `John caused Bill to die (/to become dead?) are synonyms; or even whether it is possible for one to be true when the other is false (which is not, by any means, equivalent to their having the same meaning.) See also the current scuffle over what, if anything, the results in `experimental philosophy' show about the stability and reliability of synonymy intuitions. (reference). We don't propose to insist on this, but we think somebody probably should.)

unicorns' to `no Devil's horns'. So. even assuming 'unicorn' and 'devil' are coextensive (because both denote the empty set), their inferential roles (hence their meanings, according to IRS) are different. Likewise, someone who hasn't heard about the Morning Star being The Evening Star might reasonably decline the inference from `The Morning Star is uninhabited' to `The Evening Star is uninhabited'. By contrast, if `bachelor' and `unmarried man' really are synonyms, then the inference from `is a bachelor' to `is an unmarried man' should be accepted by anybody rational who understands both expressions. Arguably, the identity of meaning with inferential role would explain all these facts, whereas, prima facie, the sort of meaning nihilism that we are preaching cannot.

So far so good: if there are such things as inferential roles and if coextensive concepts (/expressions) can have different inferential roles, that would explain the (presumptive) intuition that coextensive concepts needn't differ in content. So let's suppose, for the purposes of argument, that there are such things as inferential roles (and that the worry, raised in Ch1. that inferential role semantics is intrinsically holistic can be coped with somehow or other.) In that case, there is, after all, a respectable argument for IRS, hence for there being meanings as well as extensions No?

No. The suggestion that meaning is inferential role is subject to the same objection as all the other stories about what meaning is that we've discussed so far: namely that since inference is itself a *semantical* notion, such suggestions are circular. For example, inferring from one thought to another isn't just *any old* case of one thought causing another; it's a case of thought-to-thought causation that is generally *truth preserving* (associationists to the contrary notwithstanding). The hard part of figuring out what thinking is, is to explain *why* it preserves truth when it does, and why it doesn't preserve truth when it doesn't. And, of course, *truth* (like *reference*) is a paradigmatically mind/world relation.

For all that, perhaps you feel a certain nostalgia for the view that something like inferential role is preserved in translation, synonymy, paraphrase, communication and the like. So be it. We are prepared to split the difference: What is appealing about the idea that content and IRS are somehow the same thing can be retained *without* construing IRS as a theory of content.

Here's the idea: *Matching conceptual roles* are indeed what good translation… etc aims to achieve; but they aren't a supervenience base for intentional content. That's because (as previously remarked) there is no such thing as intentional content. Rather, what goes on is something like this: Every concept has a (possibly null) extension; but also, each mental representation is surrounded by a (rather hazy) belt of connections between its tokenings and tokenings of other mental representations. Extensions must, of course, be preserved in translation, communication and the like; if I ask you for a mouse, an elephant won't do; if you bring me an elephant, then you're being mean; or perhaps you don't know the difference between mice and elephants; most likely you just didn't understand the request. Likewise, a rough alignment of our conceptual roles is a desideratum. Perhaps I believe, and you know that I believe, that gray mice are nice on toast but brown mice are not; and suppose that I believe, and you know that I believe, that it's nearly time for lunch. None of that is plausibly part of the content of my (or of your) MOUSE concept; not, at least, if content is supposed to connect, in the usual ways, with analyticity, modality and the like. Still, communication between us has been less than a success if, given my preferences, and given what you believe about my preferences, and given what I believe that you believe about my preferences… you bring me a brown mouse for lunch.

We think this is the right way to think about how inferential roles connect to communication and, mutatis mutandis, to other cognitive processes that are commonly said to be meaning preserving). That would explain why such processes are so invariably vague and open-ended and context-sensitive; in fact, why they aren't, and quite possibly can't be, the domains of serious theories. And, we emphasize, you can have all that without supposing there are *any relations at all* between inferential roles and conceptual content. Even a meaning nihilist can have all of it; for free.

We've known students to be outraged by such proposals as that there aren't any such things as translations, metaphors, communication, definitions and the like. But, truly, =we only mean the kind of thing that a botanist means by saying that there is no such thing as grass. *Of course* there is grass for workaday purposes; but not for the purposes of scientific taxonomy or scientific explanation, which are, as Beatrice would put it "too costly to wear for working days".

`One over many'

Plato worried about what it is that chairs have in common *as such.* (That's why he cared so much about defining `justice': His idea was that 'justice' and 'chair' are both grist for the philosophical mill because each applies to, and only to, things that satisfy its definition.) If so, there is after all a reasonable response to `why do we need meanings?' Namely, it's because some truths about which properties that the things in a concept's extension have *essentially* are grounded in semantic truths about the concept; in effect, it's part of the meaning of `chair' that chairs are to sit on. Metaphysics is just one damned analyticity after another.

Our reply is uncharacteristically brief: We don't believe a word of that. It is (as Kripke has remarked) typically empirical inquiry, not lexicography, that reveals essences. It was chemists who figured out the essence of water, not philosophers.

People who think that there is a semantic solution to Plato's problem about the one over the many, do so because they think that `What do chairs have in common?' can be answered *in a vocabulary that doesn't include `chair'* (i.e in any vocabulary in which the definition of `chair' can be framed) . But we doubt that there *is* any such vocabulary. What chairs have in common as such is that they, and only they, are chairs. (Semanticists who think otherwise are cordially invited to define `chair'. We doubt they will succeed.)

Which Link?

Suppose there is a causal chain from a thing-in-the-world to a representational mental (/neural) state; and suppose (what can't but be true) that this chain has more than one link. The question arises, *which link in the chain is the referent* of the state? If you think that there are intensions, and that intensions determine extensions, there is no problem: it's the link that satisfies the intention. George Washington is the extension of the concept GEORGE WASHIGTON because the concept GEORGE WASHINGTON is actually a description, and George Washington is the one and only thing-in the-world-that satisfies the description. Perhaps he's even the one and only thing in any *possible* world that does. But what if you don't think that there are such things as intensions?

That question is discussed (and, we think, answered) in Fodor (REFERENCE), so won't repeat the proposal here. The general idea is that the referent is the most proximal link in a world-to-representation chain is the one at which all the actual *and counterfactual* chains from the world to that link intersect. (If we aren't allowed counterfactuals, we give up). `That chair' refers to that chair because all the causal chains that end with my saying `that chair' did (or would have) intersected at that that chair.

`But suppose there is only one such chain, actual or counterfactual? What is the referent then?' We think we're committed to the following (more or less Berkelian) prediction: If the actual causal chain is the *only* one that's possible (perhaps because a certain creature has only one sensory mechanism, or only one connection between its sensory mechanisms and its brain) then that creature can't refer to things-in-the-world (in effect, it can't see things as being-in-the-world.) Roughly, your token of`that chair' refers to *that chair* only if. all else equal, all the causal chains that end (/have ended/would have ended ) in that token intersect at that chair. To be sure, It's not very obvious how such a claim might be tested; but perhaps the following datum is relevant: If the retina is `stabilized` (so that movements of

the eye or of the percept always produce the proximal image of the percept at the same place on the retina and the same brain state) *the percept fades.* In effect, you can't refer to a thing-in-the- world that has only one way to getting at your brain.

Those are all the non-Fregian arguments we can think of against the Contrarion thesis that there is nothing to conceptual content except reference. We end the chapter with what we take to be a very powerful argument that favors it: If conceptual content is referential, the notorious `pet fish' problem (see Ch2) disappears; since extensions compose, the extension of `pet fish' is just what it should be: all and only things that are both pets and fish are in theextension of 'pet fish'.

And, anyhow, reference *must* be all that there is to content: since reference is the only semantic relation of which there is a prayer of a naturalistic account. The chapters that follow will provide a preliminary sketch of a kind of causal account of the reference of concepts that we think might actually work.

This page deliberately left blank

Chapter 4 : Reference Within The Perceptual Circle:
Experimental Evidence for Mechanisms of Perceptual Reference

Introduction

If, as we suppose, reference is a causal relation between referents-in-the world and tokens of the symbols that refer to them --- and is hence *not* to be explained in terms of intensions or their satisfaction conditions--- then a theory of reference is required to provide necessary and sufficient conditions for the cause of a symbol's being tokened to be its referent. And, if the theory is to meet reasonable naturalistic constraints, its characterizations of such conditions mustn't presuppose unnaturalized semantic or intensional concepts. But there are plausible reasons to doubt that any such project can actually be carried through. It is, for example, perfectly possible for someone who lives in New Jersey in 2013 AD to refer to someone who lived in Peking in 200 BC; e.g., by uttering that person's name. And it seems, to put it mildly, unobvious that a causal relation between the bearer of a name and its utterance would be sufficient, or necessary, or in some cases, even possible, for the one to refer to the other.[1]

Perhaps, however, a strategy of divide and conquer might help here: First provide a theory that accounts for cases where the relevant causal relation between a symbol and its referent is relatively direct; then work outward from the core cases to ones where they are less so. That is, in fact, the path we have been following; it has motivated several aspects of the discussion so far. For one: if you take reference to be a causal relation between referents and symbols, you are well-advised not to take *utterances* as paradigms of the latter. There is patently nothing that it is necessary to *say* about a thing in order to refer to it. Just *thinking* about it will do; and one doesn't say (or even publish) everything that one thinks. Likewise the other way around: To merely utter `Sally' is not thereby to refer to everyone ---or even to anyone--- who is so-called. This is one reason why a Mentalistic version a theory of reference is better suited for naturalization than a behavioristic one, all else equal. *Saying* is an action; whether one says "chair" depends on more than whether one sees a chair and knows that "chair" refers to chairs; one might, decide to keep one's chair-thought to oneself. But mentally tokening the concept CHAIR (e.g., seeing the chair *as* a chair) isn't typically a thing one *decides* to do. If, in such a case, one is attending to the chair that one sees, seeing it as a chair (hence tokening CHAIR) might well be a necessary consequence. Accordingly, one reason we've gone on so about *perceptual* reference ---reference to things in the perceptual circle (PC)--- is that if a causal theory of reference is ever to be true, the most likely candidate is the reference of a tokened mental representations to a thing that one perceives. For those cases, we think that this Chapter offers a plausible first approximation; namely that reference supervenes on a causal chain from percepts to the tokening of a Mentalese symbol by the perceiver. To that extent, we are in

[1] It might be suggested ---it might even be true--- that a necessary condition for the *utterance* of a name to refer to its bearer is that it be caused by the speaker's intention that it do so. But it's hard to see how a parallel condition could apply to reference in *thoughts* ; one doesn't usually have intentions with respect to the semantics of one's thoughts. Compare: 'From now on, when I *speak* about Marx I intend to be referring to Groucho, not Karl' with `From now on, when I think about Marx, I intend to be referring to Groucho, not Karl.' (Fodor, 2009) [`*Enough with the Norms*, Already' in Hieke and Leite (eds), Reduction Between The Mind And The Brain,' Ontgos Verlag.]

agreement with the Empiricist tradition. From our perspective, what was wrong with Empiricism was: first that it took the objects of perception to be typically mental (`bundles of sensations' or something of the sort); and second that it took the objects of thoughts, insofar as they aren't about things that are in the PC, to be constructions out of sensory elements. (Skinner made things worse by substituting his Behaviorism for the Empiricist's Mentalism, and his conditioning theory for their Association of Ideas.)

So, according to the Empiricists, and also according to us, early stages of perceptual processing provide canonical representations of sensory properties of things-in-the-world; and a process of conceptualization then pairs such canonical sensory representations with perceptual beliefs (i.e., with beliefs of the *that's a chair* variety.) The perceiver's background knowledge then mediates the inferential elaboration of his perceptual beliefs (`there's a chair, so there's something I might sit on') in ways that militate for behavioral success. But about this latter process ---the interaction of perceptual beliefs with background knowledge--- nothing is known that's worth reporting (for more on this, however, see Fodor, 2000).

This kind of theory imposes (relatively) strict limitations on the availability of previous cognitive commitments to the fixation of perceptual beliefs; the operations that perceptual mechanisms perform are more or less mandatory once a canonical sensory description of the referent is given. So the Empiricists were right that there is a robust sense in which theories of perception are at the heart of theories of mind-world semantic relations; perceptual processes are by and large `data driven'. Causal interactions with things in the world give rise to sensory representations, and sensory representations gives rise to perceptual beliefs. We emphasize, however, that this is intended as an empirical claim about the etiology of perceptual beliefs; in particular, it is intended to be empirical psychology rather than a priori epistemology

We think this sort of causal sequence is sufficient to establish a reference relation between (tokenings of) mental representations and the things-in-the-world that they are mental representations of. We think that is how semantic content enters the world; it's how, in the first instance, mental representations get `grounded' in experience. Since we're assuming that referential content is the only kind of conceptual content there is, this amounts to (a very schematic, to be sure) metaphysical theory of the semantics of mental representations; and, since the mind-world relations that this kind of theory invokes are, by assumption, causal, the proposal is thus far compatible with the demands that Naturalism imposes on the cognitive sciences.

In short, we think the causal chains that support the reference of mental representations to things-in-the-world are of two distinguishably different kinds: One connects distal object *within* the PC to perceptual beliefs; the other connects distal objects *outside* the PC to mental representations that refer to them. This Chapter is about the former; the next Chapter is about the latter.

Perception, Attention and objects

Arguably the area of Cognitive Science that has made the most progress in recent years has been Vision Science (experimental and computational), which devoted a considerable part of its energy to what is called Visual Focal Attention.[2] In so doing, it has found itself rubbing up

[2] The term "focal attention" is used to distinguish the sort of attention that is important to us in this chapter – the relatively narrowly focused region or set of properties where visual resources are brought to bear, as opposed to more informal uses of the term as in "pay attention" or "attend to your work not to what's on TV".

against the sorts of issues that we have been discussing. In particular, the goal of naturalizing reference relies on findings from the study of focal attention. Many philosophers have recognized the importance of focal attention to reference, particularly to demonstrative reference, and have suggested that to make a demonstrative reference to something in the perceived world involves attending to it. We think that something like this is on the right track and indeed, on the track that leads to a possible way to naturalize reference. But much more needs to be said about that nature of different mechanisms involved in attention since the intuitive sense of attention does not itself provide the essential mechanism needed for reference. Before we get to this point we will sketch some background on the role that focal attention plays in various perceptual functions and suggest that this does not by itself give us a basis for the world-mind link that we need. The present goal is to bridge the very significant gap between what the *transducers* (or sensors, as they are referred to in psychology) provide to the *early vision system* and what, in turn, the early vision system provides to the cognitive mind. Thus, as we will see shortly, this story is committed to the view that most visual processing is carried out without input from the cognitive system, so that vision is by-and-large cognitively impenetrable and encapsulated in a modular architecture (Fodor, 1983; Pylyshyn, 1999).

Among the ways of understanding the relation between focal attention and visual reference are those that derive from different approaches to the foundations of psychology, and particularly of the study of vision: these are *behaviorism*, *information-processing psychology* and the `*direct perception*' ideas of J.J. Gibson. .These are certainly not the only possibilities. For example there is a fair amount of current interest in what have been called `*embedded vision*` or situated vision' and in motor-based vision theories (O'Regan & Noë, 2002).But these can be viewed as deriving from the three foundational views just mentioned. We begin by the account of *focal attention* that each has provided.

1. Attention has often been viewed as the brain's way of matching the high speed and high capacity of visual inputs from sensors with the relatively slow speed and limited capacity of subsequent visual processes and a short-term memory that receives information from the early stages. It has been likened to a spotlight that focuses limited perceptual resources at places in the visual world. There is a huge experimental literature on focal attention and we can only touch on a very small portion of this work that is relevant as background for this chapter. We will take for granted some of the general conclusions (listed below) concerning visual focal attention, that have been drawn from many experiments.
2. Attention, like a spotlight, can be switched or moved along a (usually linear) path between visually salient objects in the field of view. This can even happen without eye movements.
3. Attention can be shifted by exogenous causes (as when it is attracted by something like a flash of light or the sudden appearance of a new object in the field of view); or it can be controlled endogenously, as when people voluntarily move their attention in searching for some visual feature.
4. Although it seems that, under certain conditions, attention may be moved continuously between different objects of interest,[3] the more usual pattern is for attention to switch to, and

[3] There is even controversy as to whether data suggesting that attention can move continuously through space can be better explained by a more discrete switching of attention, together with a certain inertia inherent in the build-up and decay of attention as it spreads and recedes in space (Sperling & Weichselgartner, 1995).

adhere to, an object. If the object is moving, then attention will stick to the moving object: that is, it will `track` that object. Attention is thus said to be object-based. This automatic tracking of objects is one of the most consequential properties of focal attention that we will discuss in this chapter.

5. The default situation is that attention tends to be unitary (hence the spotlight metaphor), although it can sometimes be broadened or narrowed (so-called "zooming" of attention) and under some conditions it can be split into two.[4]
6. Attention is required for encoding some properties of objects in the visual field, especially for encoding conjunctions of properties.

These features of focal attention are mentioned here in order to contrast them with a mechanism that is more fundamental and more directly relevant to the theme of this book: a mechanism referred to as a *visual index* or a *FINST.* [5] FINSTs are a species of mental representations sometimes likened to such Natural Language demonstratives as the terms *this* or *that*, although there are significant differences between FINSTs and demonstratives. FINSTs bear a resemblance not only to demonstratives, but also to proper names, computational pointers and deictic terms. But since , all these analogies are misleading in one way or another so we will persevere in the FINST neologism.

We begin by illustrating how FINSTs arose as an explanatory notion in experimental psychology. We will discuss some empirical phenomena beginning with the detection of such properties of sets of objects as the geometrical shape formed by the objects or their cardinality.

Picking out and binding objects to predicate arguments

Determining the cardinality and spatial pattern of a set of objects

When the numerosity of a set of individual visual objects is no more than about 4, observers can report the number very rapidly and without error and. Performance is not influenced by shape (except when objects form some simple familiar shape, such as a square or equilateral triangle, when enumeration is especially rapid) or nor color, nor by whether observers are pre-cued as to the location where the objects will appear (Trick & Pylyshyn, 1994). Although the time to enumerate a small set still increases with the number of items, the increase is very small (i.e. slope of the RT *vs* number graph is about 50-70 milliseconds per additional object). Enumeration of more than 4 items shows a different pastern; here shape, color, and location are relevant to enumeration performance. Also it takes much longer for each additional item enumerated (the RT *vs* number slope is greater) and pre-cueing their location facilitates this enumeration process. The explanation we propose is that the appearance of a visual object can cause a FINST indexes to be grabbed. Several such indexes, up to a maximum of about 4 or 5 may be grabbed simultaneously. Since indexes can be used to rapidly switch attention to the indexed objects, the cardinality of the set of indexed objects can then be determined by

---

[4] These are simplifications based on interpretations of experimental findings. As often in experimental psychology there is room for disagreement (Cavanagh & Alvarez, 2005; Scholl, 2009); (but see Pylyshyn, 2007).

[5] The term is an acronym for ' Fingers of Instantiation': FINSTS are mental representations that serve to ***instantiate***, or assign a value to, a variable; particularly a variable that serves as an argument to a function or predicate. The term first appeared in(Pylyshyn, Elcock, Marmor, & Sander, 1978).

sequentially attending and counting them (or perhaps even just counting the number of active indexes which, by assumption, does not exceed 4). If the number of objects exceeds 4, or if the objects cannot be individuated because they are too close together then this method of enumeration is not available. In that case observers must use other means to mark already-counted items and to search out the yet-uncounted ones. One can imagine a variety of ways of doing this, including moving attention serially, searching for each item or even to subitize subsets of items and then adding the results to compute the answer. These and other options on how this might be done have been proposed and tested (Mandler & Shebo, 1982; Trick & Pylyshyn, 1994; Watson & Humphreys, 1999; Wender& Rothkegel, 2000). But the relevant point here is that the quick and accurate way of doing it, by using indexes to access items in order to enumerate them, is not available when the items cannot be individuated and indexed. The remaining obvious ways of counting require a serial scan which searches for and visits each item while incrementing a counter. Thus counting more than 4 items is expected to be slower (as items are located and marked in the course of being enumerated) and, unlike in subitizing, to be sensitive to the spatial distribution of items or to visually precueing their location (e.g., providing information as to which quadrant of the screen they will appear in). This is what we found (Trick & Pylyshyn, 1994).

The same idea may be applied to the recognition of many simple patterns of objects. For example (Ullman, 1984) examined how one might detect patterns such as **Inside(x,L)** or **SameContour(x, y, L**), where x, y, … are individual objects and L is a contour object (illustrated in Figure 1). However the visual system detects the presence of these patterns, it must first identify which individual objects the predicate will apply to.[6] This is where the FINST indexes come in. Once the arguments of the predicates are bound to particulars using FINST indexes, the recognition process can be executed. In the predicates examined by (Ullman, 1984), it turns out that the evaluation must rely on a serial processes such as "area filling" or "contour tracing" or "subitizing".

[6] This claim is too strong in general since one cannot rule out the possibility that a predicate like **Inside**(x,L) might be evaluated even if the arguments are not bound to unique individual objects. The variables could be linked to a 'filter' function that yields an object under appropriate circumstances, in which case the value of the predicate may be known even if the particular individual is not bound to one of the arguments of the predicate. In some programming systems this may be represented as **Inside**($x,L), where the term $x indicates a filter function that may determine if an object satisfying the predicate exists even if it is not indexed at the time. An even stronger version of this filter option might be one in which the filter indicates that there is an x that satisfies the predicate even though it has not been indexed at the time (we might think of this as an existentially quantified variable **∃**(x) which asserts that there is an x that makes **Inside**(x,L) true even though x is not bound to it at the time). Whether these alternatives occur in vision is a larger question that is outside the scope of this theory.

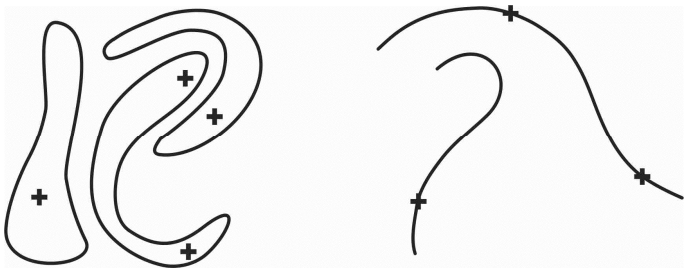

Figure 1. In order to evaluate the predicates Inside(x,L) or SameContour(x,y,L), which specify how the + marks are spatially related to the contours, the arguments of the predicates must be bound to particular objects in the figure, namely the + marks and the contours. FINST indexes serve this external binding function.

The picture we are proposing is sketched in Figure 2 which shows the indexes providing a referential link from conceptual representations to distal objects. This figure does not show what the conceptual representations are like or how they use FINSTs to refer to the objects in the world. It does, however, show the conceptual representations as being organized into "Object Files" which play a role in allowing object properties to be represented as conjoined or as belonging to the same object (as discussed in the next section).

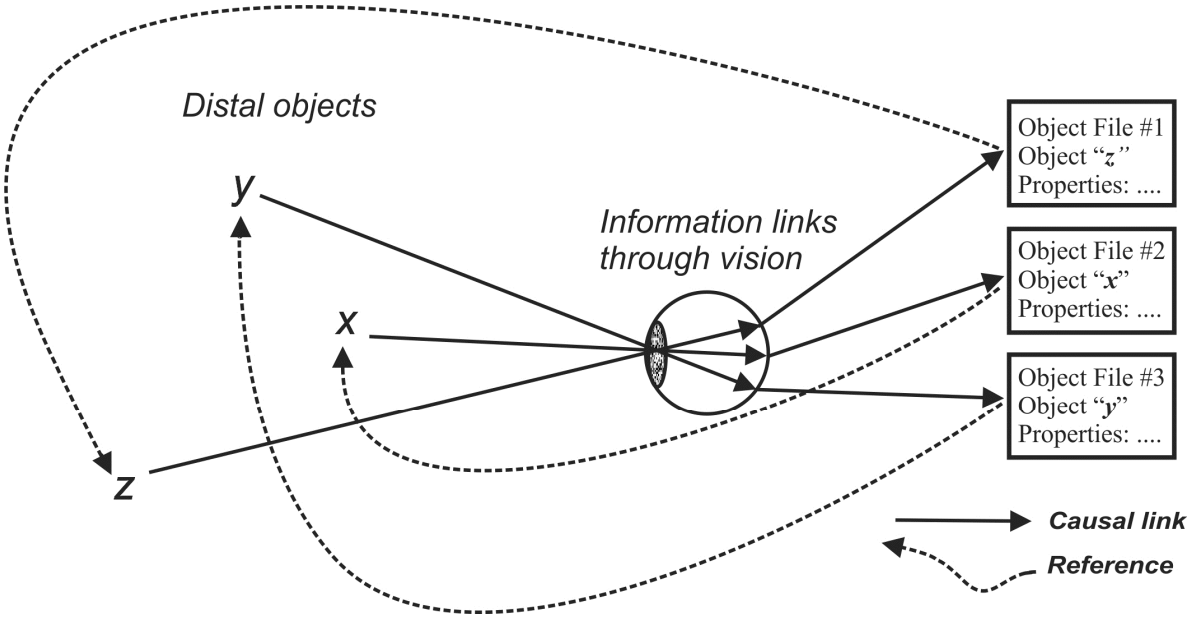

Figure 2. Illustration of how FINST work in bridging distal objects with representations of their sensory properties. It also shows how the conceptual representations may be stored or indexed in terms of the individual objects to which they refer. These "Object Files" are discussed below.

Solving the Property Conjunction Problem (aka "the binding problem")

Figure 2 shows that representations of (at least some) properties of indexed objects are stored in the *Object File* corresponding to that object. This is not intended to suggest that objects are identified by first encoding some of their properties. On the contrary, our view is that properties of an object are associated with objects *after* the assignment of indices, if at all. Moreover, the early stages of visual processing must not conflate property tokens so that later stages cannot tell which individuals properties belong to. In standard treatments, in both psychology (Treisman,

1988) and philosophy (Clark, 2000; Strawson, 1959), this problem is solved by encoding the *location* at which the property tokens are detected; so that the earliest stages of vision provide information in the form: "*Property-P-at-Location-L*". But this account can't be the general case since we can distinguish between figures even if the location of the relevant property tokens is the same in both, namely at their common center (shown in Figure 3). There is a still further problem with location-based property since most properties do not have punctate locations; they cover a region. Which region? Clearly it's the one that corresponds to the boundaries of the object which bears those properties. So the visual system cannot apply property-at-location encoding without first identifying the object to which the properties are ascribed; so it cannot escape individuating objects *before* it decides which properties belong to which object. Berkeley was wrong to think that objects are bundles of tokens of properties; rather, as Locke thought, they are the individuals to which properties (locations included) belong.

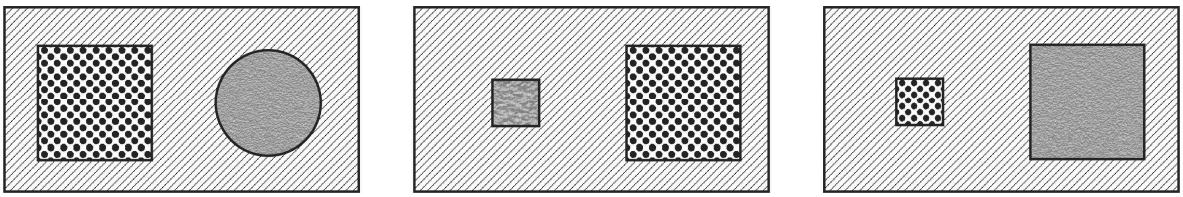

Figure 3. The early vision system must report information in a form that makes it possible to discriminate these three figures that are composed of identical sets of shape, texture and size although in different combinations. To specify how it does so is to solve what has been called the Binding Problem. From (from Pylyshyn, 2007).

The perception of a property token as a *property of an object* is extremely general; *In a visual scene properties are perceived as* belonging *to object tokens,* In fact when we perceive ambiguous figures, such as Figure 4, a change in the percept is accompanied by a change in the object to which the contour belongs.

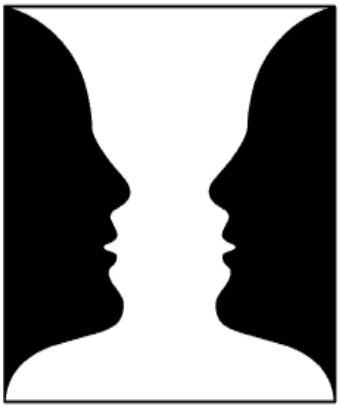

Figure 4. The "Rubin's Vase" formed by two profiles is an ambiguous figure which appears to change from being seen as a vase to being seen as two profiles. As the percept changes, the contour changes from belonging to the vases to belonging to the profiles. This sort of "belonging" is computed early in the visual process.

The correspondence Problem: Keeping track of objects even though their properties change

Visual representations must be constructed dynamically. Each second of our waking life our eyes are moving very rapidly, in movements called *saccades,* during which they can reach up to 900° per second, and they are blind to any pattern they move over. Between successive saccades they rest only briefly in what are called *fixations*. Visual representations must be constructed over a series of such eye fixations. There is also evidence that visual representations are constructed over time even *within* individual eye fixations.[7] In either case there is the problem of establishing a correspondence between objects identified at different times.

One way this could be done is by encoding each object at time t_1 in terms of a rich enough set of properties that they would uniquely identify the same object at time t_2. But this is unlikely, even if such properties could be found, for several reasons. One is that the solution to this correspondence problem must work even when all the objects have identical features and differ only in their histories – they count as different individuals because they came from different locations or had different properties a short time earlier. Another is that the objects' properties may be changing over time (e.g., the objects might be morphing in shape as well as moving into regions with different light) so keeping track of their properties does not guarantee that they will be correctly identified as the same individual object at a later time. And even more importantly, empirical evidence shows that the visual system does not solve correspondence problems in that way (see, for example, Figure 8). Consider, for example, the 'kinetic depth effect' in which dots, painted on the surface of a cylinder, are projected onto a two-dimensional plane. When the cylinder is rotated the dots are seen as moving in depth around the axis of the cylinder, even though of course they are moving on the two-dimensional plane. In order to create this illusion, individual dots must be tracked as the same moving dots over time. In other words a correspondence must be established between dots in each instant of time so they can be seen as the same moving dot rather than a sequence of different dots. What is important about this correspondence is that it depends *only* on the spatiotemporal pattern traced out by individual dots and not on any properties that these dots might have. Experiments show that such properties of a dot as its color, size, and shape do not determine the correspondence between dots over time. This finding is quite general and has been shown in a variety of apparent motion studies (Kolers & Von Grunau, 1976). Shimon Ullman (Ullman, 1979) showed that computing 3D structure from the motion of a texture field depends on the latter containing discrete objects among which certain spatiotemporal relations hold, not on their properties being unchanging.[8] It appears that relations of identity/difference of moving objects depends on *first* deciding which parts of a display belong to the same object and *then* determining which objects have moved, *not vice versa.* The matching process for individual objects follow what are sometimes referred to as Korte's Laws, which specify the relations about time and distance required to establish the perception of smooth motions from a series of discrete object positions.

One of the main characteristics of visual perception that led Pylyshyn (Pylyshyn, 1989, 2001) to postulate FINSTs is that vision appears not only to pick out several individual objects automatically, but also to keep track of them as they move about unpredictably.

---

[7] Direct evidence for this claim is provided by (Kimchi, 2000; Parks, 1994; Sekuler & Palmer, 1992).

[8] There is more to determining which objects correspond than just using proximity and timing. In fact a wide range of configural properties also bear on how correspondence is established (see Pylyshyn, 2003b, for more detailed discussion), but they all involve spatiotemporal factors.

An early demonstration that observers keep track of information associated with moving objects is described in (Kahneman, Treisman, & Gibbs, 1992). As illustrated in Figure 5, they showed observers a pair of squares each containing a different letter. Then the letters disappeared and both squares moved along a roughly circular trajectory, each ending up at 90° from its initial position, and in every case ending equidistance from where the letters started. Then a letter appeared in one of the squares and the observer had to read it as quickly as possible. Kahneman et al found that the time to read the letter was significantly faster if it appeared *in the same square* that it had initially been in. (The experiment controlled for such factors as the locations of end-points, speeds, and so on) other words being in the same box that it started in, enhances speed of recognition of the letters.

Kahneman et al explained these results by appealing to something they called an *Object File,* a term that they introduced. When an object first appears in the visual field, an Object File is created for it. Each file is initially empty but may be used to store information concerning the object for which it was created. Thus there is a connection between a file and the object associated object with its initial. The nature of connections between objects and files is, in effect, the main subject of this book.

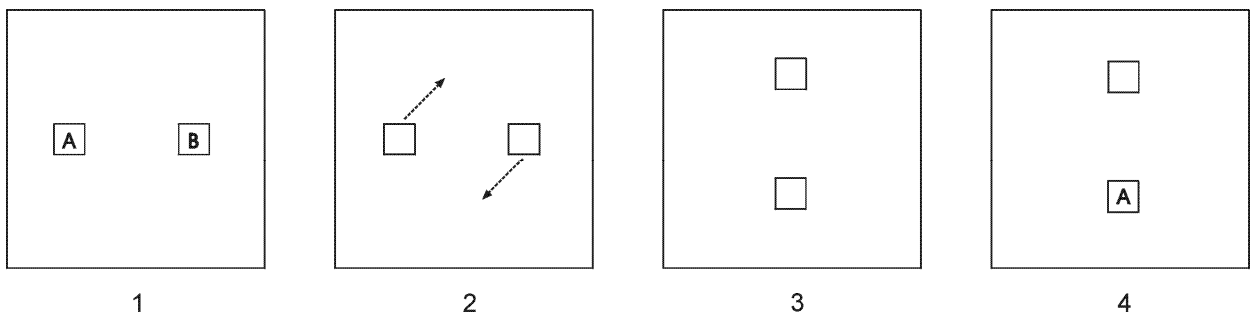

Figure 5. Illustration of the Kahneman et al., (1992). Demonstration that letter priming travels with the object it started in . Here observers see letters in boxes that then are cleared and moved to a new location. Subjects must then name the letter that appears in one of the two boxes; either the same box it had been in or a different box. When the letter reappeared in the same box it is named more quickly than if it appeared in the other box, even though nothing else favored the same-box case (not the distance moved nor other properties that are controlled).

At about the time that the Kahneman experiment was being carried out, Pylyshyn and his students demonstrated, in hundreds of experiments (described in Pylyshyn, 2001, 2003b; 2007 and elsewhere), that observers could keep track of up to 4 or 5 moving objects without encoding any of their distinguishing properties (including their location and the speed or direction of their movement). Because these studies are germane to our current thesis they are described in more detail below.

Multiple Object Tracking experiments

An experimental paradigm (**'Multiple Object Tracking'**), was inspired by the hypothesis that we can keep track of a small number of individual objects without keeping track of any of their token visual properties. Let's begin by describing this extremely simple experiment, illustrated in Figure 6.

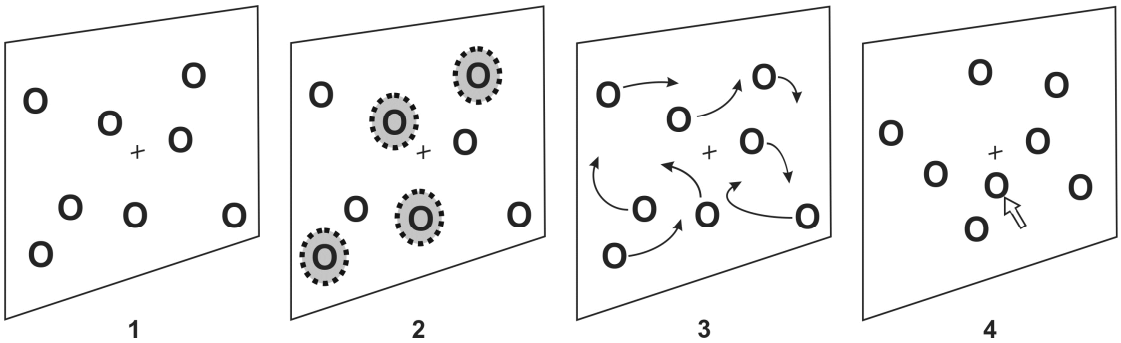

Figure 6. Schematic time-slice view of a trial of the Multiple Object Tracking (MOT) task. In (1) a set of objects (in this case 8 circular objects) is shown, (2) Some (generally half) of the objects, referred to as "targets," are made distinct (often by flashing them on and off a few times, indicated in panel 2 for expository purposes by the halo round the targets), (3) Then all the objects move in unpredictable ways (in some cases bouncing off one another and in other cases passing over one another); (4) After a certain interval, all objects stop moving and the observer uses the computer mouse to select the subset of targets that had been identified earlier.

A few of the findings, that have been replicated hundreds of times[9] are as follows:

- Almost everyone (even children as young as 5 years of age; (see Trick, Jaspers-Fayer, & Sethi, 2005) is able to do this task with a display of 8 objects (four targets and 4 nontargets; 3 for the five-year olds) and an accuracy of more than 80%. (There are some individual differences and some effects of prolonged practice – even practice on somewhat different visual tasks, such as video games (Green & Bavelier, 2006).
- There are a number of factors that affect performance, the most reliable of which is the distance between objects and the amount of time that pairs of objects remain close to one another. Other factors are in dispute. For example it has been widely observed that speeding up the motion produces a decrement in performance (Alvarez & Franconeri, 2007), but it has also been shown that this is most likely due to the confound of speed with average distance between objects (Franconeri, Jonathan, & Scimeca, 2010; Franconeri, Lin, Pylyshyn, Fisher, & Enns, 2008). Similarly, many people feel intuitively that the task uses attentional resources, so that having to perform an auxiliary attention-demanding task would lower tracking performance. Yet many studies have failed to support this intuitive conclusion (Franconeri, et al., 2010; Franconeri, et al., 2008; Leonard & Pylyshyn, 2003).[10] But notwithstanding such disagreements on details, it appears that tracking is both robust and limited by intrinsic properties of the cognitive architecture; in particular, keeping of track of a (small number) of objects does *not* appear to exploit prior knowledge of which objects have which properties.
- In addition, it appears that tracking is a primitive mechanism in the sense that it does *not* appear use processes, such as extrapolation of objects' trajectories, to predict the future

[9] A bibliography of a selection of publications using MOT is maintained by Brian Scholl at: http://www.yale.edu/perception/Brian/refGuides/MOT.html (as of January 2013 there were 125 papers from 37 different peer-reviewed journals).

[10] The very concept of attentional resource is problematic (Franconeri, In Press). Among other problems it is a free parameter which makes it too easy to fit a pattern of performance.

location of objects when they disappear from sight. When objects disappear briefly, either by being extinguished on the video display or by passing behind a (virtual) opaque screen, tracking performance is typically not impaired (Scholl & Pylyshyn, 1999). With slightly longer gaps, performance is best when objects reappear exactly where they had disappeared, as opposed to where they *would have been* had they kept moving in a straight line while invisible (Keane & Pylyshyn, 2006), This suggests that there is local spatial memory but no prediction of future location. Direct study of the role of encoding of direction in tracking when interrupted by gaps shows that when objects reappear having suddenly changed direction by up to 60 degrees, tracking is not impaired (Franconeri, Pylyshyn, & Scholl, 2012).

- Distinctive properties of objects do not appear to help in tracking them. Indeed, observers do not even notice when objects disappear and then reappear a short time later having changed their color or shape (Scholl, Pylyshyn, & Franconeri, 1999). The effect on tracking when the color or shape of objects changed continually during a trial were also examined. For example, tracking when no two objects have the same property at any given time (their properties changed asynchronously) was compared with tracking when all objects have the same property at a given time (their properties changed synchronously). When the motion algorithm kept objects from hitting or intersecting one another (e.g., they bounced off an invisible "fence" around the objects, or the objects were kept apart using an "inverse square repulsion" method), there was no difference in tracking between the all-objects-same and no-objects-same conditions. This shows that relations among object properties is not used to enhance tracking except perhaps in distinguishing objects that come very close together.

In all the studies described so far, the "targets" are indicated by flashing them. Flashing is a property that can be rapidly and accurate detected in search tasks and is minimally sensitive to the number of nontargets in the task. This sort of property is often said to cause "popout" in search tasks. In our terms, this kind of property is likely to capture or "grab" visual indexes. But what about properties or property-pairs that can differentiate targets from nontargets but do not popout (such as certain color differences)? These sorts of properties do not, for example, allow the segregation of different regions. In Figure 7, the panel on the left shows that one can easily visually distinguish the pattern formed by the small squares; but in the panel on the right it takes effortful search *to see the boundaries of the pattern.* The circle-square distinction yields what is called a *popout* discrimination while the circle with diagonals at different angles does not, even though both can easily be distinguished if we focus attention on them.

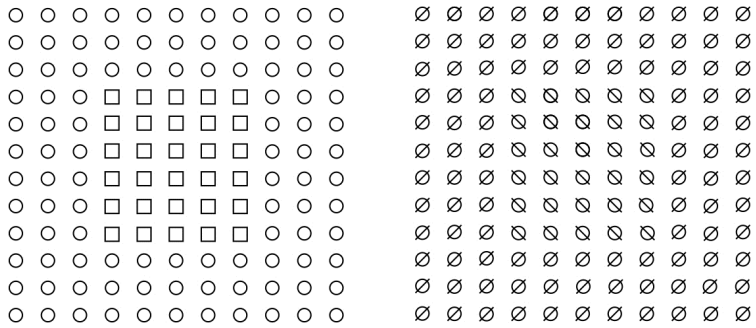

Figure 7. Objects with clearly distinguishable properties may nonetheless differ in how readily and automatically they cluster and how potent they are in grabbing FINST indexes.

Yet it seems that we can pick out a subset of objects to be tracked, even if the difference between targets and nontargets is *not* a popout feature: We can, for example, decide to track the red objects and ignore the blue ones or to track the square objects and ignore the round ones (as in Figure 7). In several experiments (described in Pylyshyn & Annan, 2006) observers can indeed select targets in an MOT task on the basis of non-popout features, even under the extreme conditions when the targets are the ones that *did not* blink. But in order to do so observers require more time to examine the initial cue display containing both types of objects. Moreover, the more targets that must be distinguished in this way, the longer observers need to examine the cue displays. This was interpreted as suggesting that in cases where features are indexed (FINSTed) under voluntary control, the process requires using focal attention to search the visual scene serially searching for particular task-relevant features. As each of these objects are found by focal attention, it enables the automatic index-grabbing and tracking processes to run their course. So the suggestion is that, in the case of task relevant properties, focal attention acts as a preparatory *enabling* mechanism for a basically reflexive "grabbing" operation. Or, to put the same claim in other terms, the empirical results so far suggest that the early processes of visual object tracking are `modular': bottom up, encapsulated, and object-based, rather than dependent on the recognition of specific properties. Which properties cause indexes to be grabbed is determined by the architecture of the early vision system, which also determines whether or not any of these properties are encoded (including, in particular, the property of being located at a particular place in some larger frame of reference[11]). The earliest perceptual operation closely resembles the function carried out by natural language *demonstratives*, such as the words *this* or *that*: They pick out but do not describe individual perceptual objects. Computer pointers do the same in data structures, notwithstanding the fact that they are called "pointers" which misleadingly suggests that they pick out objects by their location. The "locations" in a computer system are only abstractly related to physical places.[12]

[11] In any case it appears that in MOT objects are tracked in scene, rather than retinal, coordinates (Liu et al., 2005).

[12] This is frequently misunderstood since in computers one speaks of locations (and location-based retrieval). But even though data are at **some** location(s) or other at each instant in time (frequently distributed throughout the computer memory and changing rapidly as control shifts from program to program in the course of task-sharing random access memory) it is not by computing this sort of "location" that the data are retrieved from memory. The pointers correspond to addresses which, in turn, are more perspicuously thought of as names rather than locations (for more on this point see Pylyshyn, 2003b).

The existence of an encapsulated stage in visual processing is important to the story we have been telling because if there is to be a naturalized account of reference it will have to begin where patterns of light first come in contact with the visual system. Moreover at this stage the process will have to be characterized without reference to conceptualized information, such as propositional knowledge. On pain of circularity, conceptual knowledge has to make contact at some stage with causal but non-semantic and non-intensional properties of referents. According to the view we have been developing, what happens at this stage begins with objects `grabbing' indexes and ends with a description of the input in terms of conceptual descriptions stored in long-term memory. This process does not use information in cognitive memory, although it very probably uses a nonconceptual memory internal to the module.

It seems counterintuitive that early vision might not use information from memory about the likelihood that the token object currently indexed is a member of a particular kind. For example. the visual system often seems to use information about the world at large in computing the three-dimensional form of particular objects from the 2D proximal pattern on its retina(Pylyshyn, 1999). To understand this apparent inconsistency we need to distinguish between general, more-or-less permanent information – which may plausibly be built-in to the architecture of the visual system – and information about particular objects or types of objects. David Marr was arguably the person most responsible for making that distinction, although in somewhat different terms than we use here. His claim was that evolutionary pressures resulted in certain constraints on interpretation being "wired in" to the visual system. He called these "natural constraints" and researchers after him were able to uncover many such constraints(see, for example, Hoffman, 1998; Pylyshyn, 2003b).

For example, the apparent motion produced by putting objects in one temporal frame in correspondence with objects in a second temporal frame could be explained in several ways, including the application of simple heuristics.[13] Take the example illustrated in Figure 8.When the top pattern is repeatedly followed by the bottom pattern, what is seen is "apparent motion" Which particular motion is perceived depends on which objects in the first (top) figure are seen to correspond to which objects in the second (bottom) figure. This example of the visual solution of a "correspondence problem" reveals a property of the visual architecture and also the operation of a natural constraint. Intuition might suggest that the simplest correspondence would be based on the closest neighbor or on a linear translation which would yield the motion shown in panel A. But now suppose the second (bottom) pattern is rotated as in Panel B. There are several possible sets of correspondences, including one linear translation and two different rotations. What is seen, however, is a square pattern of objects moving and also rotating counterclockwise from the first pattern into the second. The principle that we proposed to explain this derives from the fact that in *our world* punctate features tend to arise mostly from discontinuities on the surface of rigid objects. If the visual system had internalized this natural constraint, it would solve the correspondence problem by pairing not just the nearest items, but the pattern consistent with the four items being on a rigid surface and then choosing the minimum rotation. This leads to a solution shown in Figure 8B where the nearest neighbor

[13] Such heuristics are almost never available to introspection and are frequently recondite and surprising as we will see in the examples below. This should not be read as suggesting that a heuristic is learned or discovered and represented as such. It's just a natural constraint whose origins are unknown (although it is tempting to assign them a teleological explanation).

condition is modulated by the rigidity constraint (Dawson & Pylyshyn, 1988). Many such solutions of perceptual problems can be traced to general properties of the world that the organism lives in. But what would happen if there was other information concerning which object in the initial display corresponds to which object in the second display. For example, suppose the items differed in shape, color or texture, as in panel C – would the similarity of objects override the simple rigidity constraint. The answer is it does not. The spatiotemporal constraints override the color or shape continuity, thus demonstrating the primacy of spatiotemporal constraints.

Later we will exhibit cases where spatiotemporal constraints override not only such properties as color and texture, but even very strong principles of physical necessity.

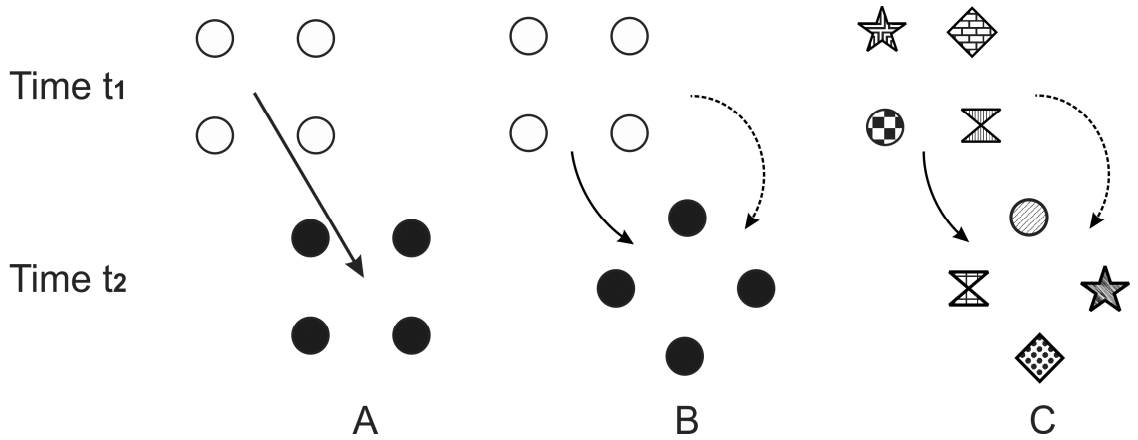

Figure 8.This shows a regularity to how the correspondence problem is solved. The two figures on the left illustrate that there is more to this principal than the "nearest neighbor" and the figure on the right shows that the correspondence is solved without regard to properties of the objects themselves.

Since *everything* that happens in a creature's environment conforms to laws of nature, one might well wonder *which* environmental regularities come to be mirrored by properties of perceptual architecture and which do not. Given the empiricist tradition that has colored the past half-century of psychology and biology, one might expect the most frequent or perhaps even the most important regularities to be integrated into the architecture of the organism. But it turns out that the constraints embodied in the architecture of vision tend *not* to be the ones that are most frequently encountered. The visual constraints that have been discovered so far are based almost entirely on principles that derive from laws of optics and/or projective geometry. Properties such as the occlusion of relatively distant surfaces by surfaces closer to the viewer are among the most prominent of these principles, as are principles attributable to the reflectance, opacity and rigidity of bodies. However, other properties of our world – about which our intuitions are equally strong – do not appear to have a special status in the early vision system. In particular the resolution of perceptual conflicts is rarely resolved so as to respect such physical principles as that solid objects do not pass through one another. Consequently some percepts constructed by the visual system fail a simple test of rationality or of coherence with certain basic facts about the world known to every observer. Apparently, which features of the world are architecturally represented can't be explained by appeal to merely statistical regularities in the environment.

Take the example of the Pulfrich double pendulum illusion (as described by Leslie, 1988). Two solid pendulums, constructed from sand-filled detergent bottles, are suspended by rigid metal rods and swinging opposite phase. When viewed with a neutral density filter over one eye

(which results in slower visual processing in that eye) both pendulums are seen as swinging in an elliptical path, but with other one seen as following behind the other (this manipulations results in the phase between the two pendulums being between 0 and 180 degrees). As a result of these differences in their perceived trajectories, the rigid rods are seen as passing *through* one another even though they are also seen to be solid and rigid. In such cases, `impossible' interpenetrations of solid objects do not seem to be blocked by the visual system, even though they are clearly at variance with what we know about how things happen in the physical world. Other examples of this same violation of solidity constraints include are well known to perceptual. Psychology. the demonstration involving: When a trapezoidal window is rotated about a vertical axis (see Figure 9) it is seen not as rotating but as oscillating back and forth about the axis. When a rod is attached to the axis so it rotates with the window, it is seen as rotating while the rigidly attached window is seen as only oscillating. Consequently, at certain points in the rotation the rod is seen as penetrating the solid window frame (this and other such examples are discussed by Rock, 1983).

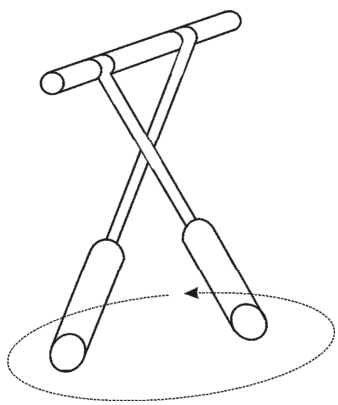

Figure 9. The "Pulfrich double pendulum". The two pendulums swing out of phase side by side, with one swinging left while the other is swinging right. When viewed binocularly, with one eye looking through a neutral density (gray) filter, they both appear to swing in ellipses. The result is that the pendulums are seen as passing though one another. This and many other examples show that vision does not incorporate even physical constraints as obvious as the impenetrability of solid objects.

Another example is the famous Ames Room; a distorted room constructed in such a way that the projection of every line and polygon in the room onto the retina of an observer looking through a peephole is identical to the projection that would arise arising from an actual3D room.(The reason it is possible to build such a room is that there are many ways to build a 3D object that projects a certain2D pattern – the 3D to 2D mapping is many-to-one). But if someone walks from one side of this room to another, the observer at the peephole sees a person enlarge and shrink as she appears in different parts of the room. This happens because the different parts of the room are higher or lower, but in such a way that all the edges and two dimensional figures (including walls and windows) appear to be rectilinear when view from the fixed point (several full-sized Ames rooms have been built. One is viewable at the San Francisco Exploratorium and shown in Pylyshyn, 2003b).

Summarizing the moral of this section: Many cases in which our visual system provides unambiguous (and usually veridical) percepts despite the inherent ambiguity of the 2D image, can be explained without having to assume that the visual system draws inferences from specific knowledge regarding what a 3D scene is likely to contain. Although the *beliefs* we come to have about what is in the scene invariably take into account what we know, the output of the early

vision system, or to put it loosely (though believe correctly) *the way particular things look,* does not take such knowledge into account. The visual system is so constructed, presumably because it has become tuned over eons of evolutionary history, that the range of interpretations it is able to make are severely restricted. This range of alternatives is specified by principles (or "rules") such as those discussed by (Hoffman, 1998; Marr, 1982; Richards, 1980) which can be understood as limitations of the possible interpretations that our visual system is capable of making, just as universal grammar can be seen as expressing the limitation in the structural inferences (grammatical analyses) possible in our language acquisition. In the case of vision, the structural limitations (or the natural constraints) are invariably associated with spatio-temporal and/or optical properties. Consequently, they do not appear to just reflect high-frequency properties of the world so much as those critical properties that animals need to keep track of in order to survive as mobile agents. The question remains why some such reliable environmental regularities come to be instantiated in the perceptual architecture while others do not. The probability that one object in a creature's environment will pass through another is zero; but the architecture of the visual system appears *not* to be constructed to reject perceptual interpretations in which they do. No merely evolutionary explanation (e.g., an increased probability of contribution to the gene pool) can account for such facts; a creature that persists in trying to walk through solid objects is unlikely to contribute much to the local gene pool.[14]

Objects and Encapsulation

The empirical evidence we have reviewed suggests that initial contact between the world and its perceptual representation begins with only a highly restricted fraction of the objects and states of affairs being represented. A plausible first approximation suggests that percepts are initially represented as "indexed objects" or as "this" or "that". Evidence suggests that the first stage in building perceptual representations is the individuation and indexing of objects. The very process of indexing objects requires that they can generally be tracked as they move or change their properties: they are picked out and tracked as individual objects. Indeed, we strongly suspect (contrary to much of the recent psychological literature such as described in Spelke, 1990) that `objects' *just are* things that can be tracked. Accordingly, since tracking is a reflex, which is to say that it doesn't involve the application of any concept, the concept OBJECT need not come into process of visual perception *at all.* That's just as well since, so far at least, no one has been able to provide a plausible account of what concept OBJECT *is.*[15]

The perceptual system indexes and tracks objects not by picking out things that fit certain descriptions (`physical`, or `follows a smooth trajectory`, or some such), but rather via a world-to-mind causal link. Certain things in the perceptual world happen to grab indexes and can then be tracked as they change properties and locations. It is this primitive tracking process that

[14]This is just a special case of the general principle that explanations that appeal to `natural selection' are invariably post hoc. For discussion, see (Fodor & Piattelli-Palmarini, 2010).

[15]The concept of an object is, of course, from time to time a constituent of *thoughts*; (mostly, we suppose, of relatively abstract and sophisticated thoughts). That is, however, entirely neutral whether the `Object Concept' is primitive. In particular, we are *not* suggesting that OBJECT can be *defined* in terms of the concept TRACKABLE; or, indeed that it can be defined at all. This is one of the reasons why, according to (Pylyshyn, 2007) FINSTs are said to be `grabbed by` (rather than applied to) `*things*' rather than to objects. Nobody thinks ***things*** is a technical term of art.

determines that they are the same objects when they change their locations in a quasi-continuous motion[16]. Thus our treatment of the early stages in perception has turned on the notions of *object* and of *tracking*, both of which are, of course, causal world-to-mind process; and neither of which need involve conceptualization. as is reference. according to us. It goes without saying that the picture is incomplete in many ways, including the question how information *other than* what is contained in Object Files is assimilated into cognition, and how one can refer to (such as relational properties among several percepts) or whose referents are not currently visible. More on such matters in Chapter 5. There is, in any event. increasing evidence over the last decade that the amount of visual information that is made available to cognitive processes by visual perception per se is much less than has often been supposed. Intuition suggest that that is made available to is into the cognition at large. The phenomenology of vision suggests that visual perception computes a finely detailed and panoramic display of dynamic information that the environment provides' whereas there is now considerable evidence that nothing like this quantity and quality of information is passed on from early vision to the cognitive mind. For example:

None of the preceding speaks to the question whether large amounts of information may be stored very briefly in some sort of `iconic` form prior to FINSTing and other perceptual processes that early vision performs: But we do believe that the information channel from the eye to visual cognition (in particular, from the eye to the fixation of perceptual beliefs – e.g., *this is a chair; that is a pineapple,* and so forth) – is much more limited than has widely been believed. The evidence that this is so comes from experiments and clinical observations of that show various sorts of "blindness," including:

- *Change blindness,*in which despite a rich phenomenology, people are unable to report major changes in a scene that occur during a brief disruptive gap caused by a saccade or by presentation of a series of images varying in some major content;
- *Inattentional blindness,* in which people fail to notice a task-irrelevant stimulus feature that occurs when they are attending to something else – even when they are looking directly at the object that they fail to notice;
- *Repetition Blindness*. When patterns are presented in rapid serial streams many of the items are not recalled, even when they are present twice(Kanwisher, 1991).
- Different sorts of brain damage symptoms including:
  - *Blindsight,* in which patients with lesions in their primary visual cortex are unable to report things in a "blind" part of their visual field and yet can react to them non-verbally (e.g., by pointing);
  - *Visual agnosia*, especially in cases where patients who are unable to recognize familiar patterns (e.g., faces) are still able to carry out accurate motor actions such as reaching and grasping and moving around: Examples include the celebrated case of DF discussed by (Milner & Goodale, 1995), or a case described by (Humphreys & Riddoch, 1987);

[14] Notice that the motion of the *proximal stimulus* – the motion of objects projected close to the viewer – such as on computer screen or the retina – is never actually continuous, even when they are seen as such. This shows that whatever detects motion must do so *before* the information is passed on to the index-tracking system. This is compatible with what is known about how motion is detected by receptor cells very early in the visual system, see, for example (Koch & Ullman, 1985) and also (Pylyshyn, 2003a) for possible models of motion encoding.

There are many other differences between how the visual world appears in one's conscious experience and the information (or lack of information) that can be demonstrated by other means (e.g., in controlled experiments). For example as far back as 1960 it was shown that information available from brief displays is neither the small amount demonstrated by experiments in reading or searching, yet not as encompassing as one's phenomenology suggests (i.e., that we have a panoramic picture-like display of finely detailed information).What we do have is partially analyzed information that is highly incomplete and that must be scanned serially in certain restricted ways [the original "visual icon" (Sperling, 1967) posited a form of representation between vision and response he called a "motor program" that prepares the response]. A strong case for the poverty of functional information in relation to visual phenomenology was made by (O'Regan, 1992).

It's important to bear in mind that encapsulation of a mental mechanism can work differently for different topics; it might be encapsulated with respect to some kinds of information but not others; or in respect of some of its functions but not others. A great deal of research on perceptual-motor coordination has shown that some visual information is unavailable for recognition and for conscious description but plays an important role in motor control. Some patients who had severe visual agnosia (so much so that they could not visually recognize their partners) nevertheless could navigate their way and could reach around obstacles to grasp some tem of interest (Milner & Goodale, 1995). In the present case, we need to ask whether there are considerations other than phenomenology for believing that more information is available to cognition than is encoded in Object Files (which is, we're assuming, largely about the `sensory' properties of the object that grabbed a certain FINST). The assumption of `object files' was, after all, initially motivated largely by considerations about the `Binding Problem' (or the 'conjunction problem'), by the desire to explain how we distinguish between two stimuli that have the same collection of properties but in different combinations (how a red square next to a green triangle is distinguished from a green square next to a red triangle). The suggestion that perceived properties are bound to FINSTed objects (rather than the other way around) has much illuminated such matters.

Our current guess is that information available in the early vision module is largely concerned with the shapes of objects. But the idea that there are knowledge-independent shape representations suitable for use in category recognition --say by looking in a form-to-category dictionary-- is well known in computer vision. A great deal of work has gone into the development of computer methods for deriving object categories from shape categories, including the use of generalized cylinders or cones (see Figure 10) deformable surfaces, structural decompositions (Pentland, 1987), Fourier methods and many other mathematical techniques. In short, a variety of approaches to shaped based categorization of visual objects is available from research on computer-vision. Some of them, mentioned above, attempt to fit what might be thought of as a rubberized surface, (a net or soap bubble) to the external parts of the scene, yielding a three-dimensional profile of the scene. The aim is to create general deformable models to both efficiently encode a scene and also to capture similarities between scenes. The techniques provide a mathematical surface deformed to encapsulate the shape of the scene so that salient aspects (e.g., location of types of maxima, minima and inflections) are represented by mathematical parameters that carry rich information allowing subsequent identification of the scene.

Another approach, proposed by many researchers in both computational vision and human vision, relies on part-decomposition as a step towards a canonical descriptions of objects

(Biederman, 1987; Marr & Nishihara, 1987). The standard re-useable parts (referred to as *Geons)* are identified by a technique that is not unlike parsing a sentence into constituents and syntactical relations among them. An example of the decomposition of a human shape, with each part represented by generalized cylinder (or cones) called Geons, is shown in Figure 10. The use of some standard form of object-part representation, such as generalized cylinders, allows an object to be represented in an economical manner: cylinders have as few as 7 degrees of freedom (diameter, length, orientation in 3D, and location of the origin of one end of the cylinder). These individual components may be hinged together so their relative orientation is important.

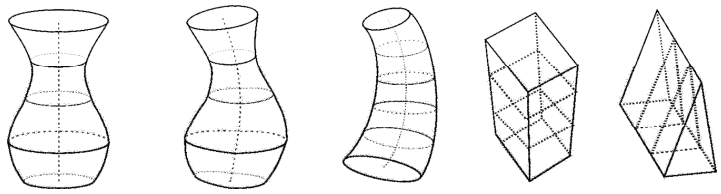

Figure 10. Illustration of generalized cones and polygons used as the basis for representing arbitrary shapes. Primitives used in constructing canonical shapes have been called "geons" (Biederman, 1987).

A number of experiments showing that people's ability to recognize objects under conditions of rapid presentation, as well as the errors they commit in generalizing them to similar objects, can be explained by assuming a representation that decomposes the objects into components (Biederman, 1987; Biederman, 1995). Pairs of objects with similar components and relations are perceptually similar. In addition a variety of other findings fall nicely into this picture. For example, pairs of figures were presented in a memory or identification experiment where parts of the figures were deleted. When the deleted part corresponded to a Geon it was more often misrecognized even though the amount of line segment erased was the same as in the control pairs. Recognition of objects under degraded perception conditions was improved when the objects were primed immediately before presentation by a brief display of the parts and their relations. These and many other experiments suggest that objects are analyzed into component parts and relations in the course of recognizing them as tokens of familiar objects.

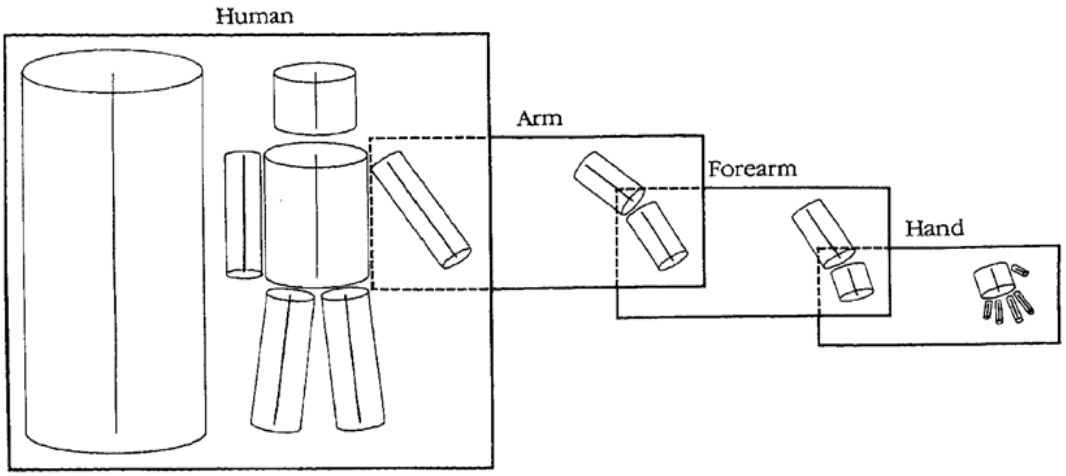

Figure 11: A sketch of a representation within a modular vision system that shows the decomposition of a shape into parts each represented by generalized cylinders. Such a system may use FINST indexes to keep track of the parts.(from Marr & Nishihara, 1978).

One might wonder why we have placed so much emphasis on *shape*, as opposed to other properties of indexed objects. It's because empirical data suggest that shape is perhaps the most salient perceptual property that distinguishes objects visually. It is also one of the central perceptual properties in virtue of which objects are seen as belonging to the same ***basic category*** (to use the term introduced by Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976) and tend to have other properties in common as well (see also the review in Mervis & Rosch, 1981). It is also a property that can be represented in a constructive manner since it lends itself to a hierarchical structure (as shown in Figure 11). Shape also tends to be invariant under many transformations such as translation, scaling, lighting conditions, and direction of view [17]. Because of its importance for the type-classification of token objects, shape has been extensively studied in computational vision where methods of encoding have been developed.

There are some conditions that must be met by the modular system we envision.[18] We have already introduced the idea of a canonical representation of shape, denoted $\mathfrak{L}(\boldsymbol{x})$. The visual system might map the shape of the object before it onto a canonical form and then may look up this shape in a table and 'recognize' it –i.e., propose a conceptually described (perhaps familiar) object that might fit the form. In order to compute $\mathfrak{L}(\boldsymbol{x})$, the early vision module must possess enough machinery to map a token object onto an equivalence class designated by $\mathfrak{L}(\boldsymbol{x})$ using only sensory information and module-internal processes and representations, without appealing to general knowledge. The module must also have some 4-5 Object Files, because it needs those to solve the binding problem as well as to bind predicate arguments to objects (and also to use the succesful Recognition-By-Parts process for recognizing complex objects, as described by Biederman, 1987). Processes inside the visual module would allow it to look up particular shape-

[17] It remains controversial whether the human early visual system encodes shapes in a viewpoint-independent manner or whether it encodes a series of viewpoint-dependent two-dimensional shapes called ***aspects,*** or both.

[18] It is no accident that researchers working on the development of canonical shapes count themselves as modularists (Biederman, 1995; Bowyer & Dyer, 1990; Koenderink, 1990b; Marr, 1982).

types $\mathfrak{L}(\boldsymbol{x})$ in a catalog of shape types. Yet this machinery is also barred from accessing cognitive memories, thus it could not carry out general inference processes.

The idea of a canonical form $\mathfrak{L}(\boldsymbol{x})$ is one that has seen considerable work in computational vision. It is a rather more complex idea than we have suggested so far. First of all, it is not a fully-specified three-dimensional shape. Because it represents an equivalence class of object shapes, many detailed properties are omitted in the many-one mapping from token object to canonical form. The canonical representation may also embody a set of related viewpoints, so $\mathfrak{L}(\boldsymbol{x})$ may consist of a set of what are called ***aspects*** of the shape. More generally, a shape representation may consist of a set of aspects that form a topological structure (called an ***aspect graph***) of the edges, junctions, and other discontinuities seen from a series of viewpoints. As the viewpoint changes gradually such a representation will remain fixed until, at some critical angle, additional discontinuities come into view and existing ones become occluded by parts of the object. Thus an aspect graph is a graph of potential aspects that capture the canonical form of an object. These have been studied mathematically in detail by (Koenderink, 1990a, 1990b). David Marr used the term "2½-D sketch" to denote a representation that embodies both a orthographic (pictorial) and partial depth information. We may view our canonical form as an aspect graph or a 2½-D sketch. As noted above, it is also closely related to the ***basic category*** described by (Mervis & Rosch, 1981; Rosch, et al., 1976) which plays an important role in recognition and memory.

So we come to a highly provisional picture of the causal relations between world and mental representations as well as causal relations *among mental representations* might play a role in early vision (illustrated in Figure 12). Like many other semantic referentialists, we hold that the content of a creature's mental states supervenes on causal chains between them and things in-the-world. But, unlike referentialists who are also behaviorists, we assume that such causal chains have quite often themselves got mental states among their links; and unlike the British Empiricists, we don't think that causal relations between mental states are typically Associative.

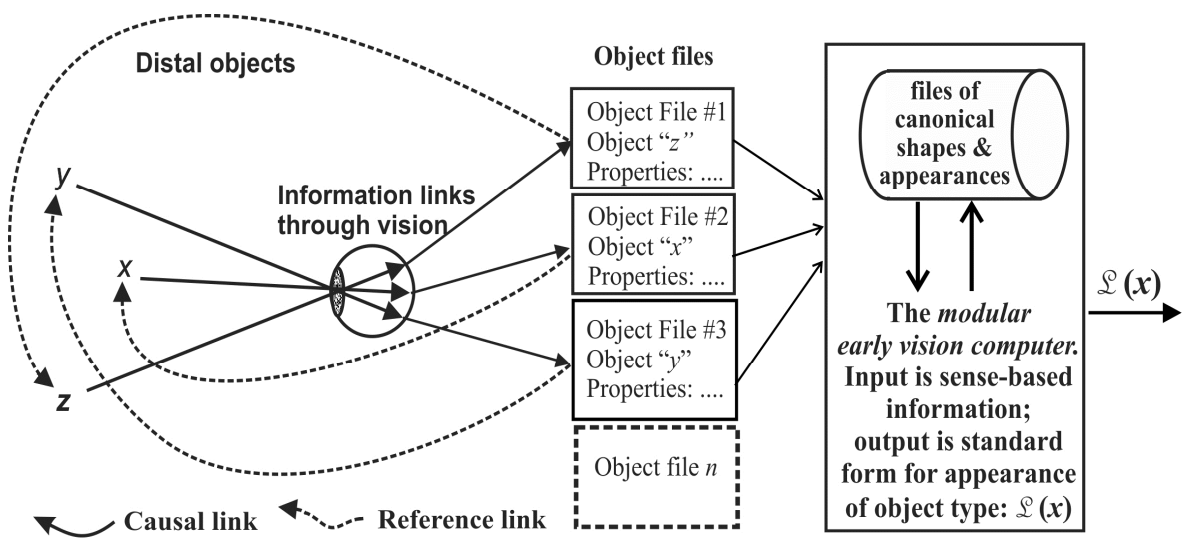

Figure 12: The FINST model of the modular vision system augmented to compute canonical equivalence classes of objects' sensory-based appearance. Other links (not shown) might allow for the possibility that focal attention could scan distal objects for the presence of other indexable objects meeting some general criterion (e.g., being to the left of a particular indexed object).

We want to emphasize the distinction between two sorts of claims that Figure 12 is intended to illustrate: The first is that there are reasonably plausible ways in which computational perceptual processes might support causal relations between things-in-the-world and the fixation of perceptual beliefs. We think this provides an essential first step towards the naturalization of reference; Chapter 5 hopes to carry the naturalization process further. The second claim is that this sort of model is compatible with data suggesting that many such processes are more or less encapsulated from cognition at large; perception is one thing, thought is another, New Look theorists to the contrary notwithstanding (Fodor, 1983; Fodor & Pylyshyn, 1981; Pylyshyn, 1999). We're aware that these two claims are independent in principle, but, as things now stand, we're prepared to bet on both.

Take away

If reference is a relation between a (mental) representation and something in-the-world, then semantic theory is faced with the question what this relation could consist in. And, if naturalism is a ground rule, the answer it gives must be some or other variant on the notion of a causal hook-up. But what kind of causal hook-up could that be? Notice that, though they are certainly related, this `Mind-World' problem is *not* the traditional Mind-Body problem. Suppose, for example, it turns out that mental states really are neural states (whatever, exactly, that means.) So we can all sleep safe in our beds; the traditional Mind-Body problem is solved; dualism is false; beliefs, desires and the like are just brain states. Still, the question persists: how do these brain states manage to be *about things in the world?* How could they have *semantic contents?* How could they refer to things? Short of answering such questions, neurological monism doesn't give us propositional attitude realism or representational realism more generally. And we need propositional attitude or representational realism because all the serious cognitive science we've got takes for granted that an organism's behavior is (very often) caused by interactions among its representations and by interactions between its representational contents and things-in-the-world.

Our working assumption in this book is that semantic relations between the mind and the world are `grounded in' causal relations between the mind and things in the mind's Perceptual Circle. This isn't, or course, to say that all you can think about is the things that you can see. But it does claim that *the mental representation of things that aren't in the PC depends on the mental representation of things* that *are.* Very roughly, it depends on reference-making causal chains that run from things you can see (or hear, or touch, or smell, or otherwise sense) to mental representations of things outside the PC. In effect, reference in thought depends on perceptual reference, which in turn depends on sensory transduction; and at each step, it's causation that provides the glue.

Some very sketchy suggestions about how minds can think about (hence refer to) things that aren't in the PC will be the topic of Chapter 5. Patently (and empiricists to the contrary notwithstanding) minds do think about things outside their PCs quite a lot of the time: for example, they remember things, imagine things, infer things, predict things, hope for things, dread things; and so forth, on and on. And the things that minds remember, infer, hope for, dread, etc aren't, in any serious sense 'constructions' (logical or associative) out of sensations. This chapter, however, has been occupied solely with *perceptual* reference; i.e. with reference-making causal chains that run from things in the PC to mental representations of such things in the head the perceiver, typically via the detection of sensory properties. It's just a truism that you can only perceive things that are in your the Perceptual Circle.

We've been making suggestions about kinds of psychological mechanisms that might underwrite this truism. Some of our suggestions have appealed to methodological or metaphysical assumptions (naturalism requires mental processes to be causal); some of them appeal to what seems, so far, to be the drift of empirical findings (perceptual processes are, to a surprising extent, encapsulated from the perceiver's background of prior cognitive commitments; the basic mind-world semantic relations in visual perception is the demonstrative reference of FINSTs); and some of them are plausible simply for want of serious alternatives (perceptual and cognitive processes are both species of computations) . This is not the sort of a priori demonstration that philosophers sometimes yearn for. But, as far as we know, it's the sort of patching-things-together by which science almost always proceeds. The next step is to think about how causal (and other) relations between things that aren't in the PC and things that are might ground the mental representation of the former.

APPENDIX: Gavagai Again

We've been campaigning for a semantics according to which:

1. The content of a concept is its extension; the content of CAT is the things belonging to the set of (actual or possible) cats; the content of LUNCH is the things belonging to the set of actual or possible lunches: the content of TUESDAY is the things belonging to the set of actual or possible Tuesdays; and so forth.
2. For cases that fall within the PC, the paradigm of reference is where the tokening of a symbol is caused by something that is in its extension. (What happens in cases of reference to things that *aren't* the PC is the topic of Chapter 5.)

But the sophisticated reader might well wonder if reference is a sufficiently strong condition on a semantics for mental representations (or, equivalently, of linguistic representations.) `Frege Arguments' are usually offered as reasons for wondering that; but so too are arguments for the `indeterminacy' of reference (IR) of the sort that Quine introduced in *Word and Object* (Quine, 1960). And whether, if we thought that there were sound arguments for the indeterminacy of reference, we would have to abandon the project to which this book is primarily devoted. But, for kinds of reasons that this Appendix will set out, we don't.

As with so many considerations that have importantly influenced discussions of semantics, it is less than fully clear how IR arguments should best be formulated. It is notorious that, prima facie, Quine's way of doing so depends on epistemological assumptions that maybe ought not to be taken for granted (roughly, that the data for the theorist's assignment of referents to a speaker's tokenings of referring expressions consist solely of correlations between tokenings of the expression and tokenings of things-in-the-world). This kind of epistemology has had some rough going over the years since *Word and Object* appeared. But, we don't think that the case for IR really depends on it. For purposes of this discussion, we will consider a formulation that we think captures the spirit of Quine's indeterminacy arguments while leaving out the behaviorism.

The basic point is very straight-forward: `Cause' determines transparent contexts: If it's true that ***a*** caused ***b***, that remains true on any substitution of a coextensive expressions for `***a***' and/ or `***b***'; if it's true that yeast causes bread to rise, and that bread is what Granny eats with her butter, then it is likewise true that that yeast causes what Granny eats with her butter to rise; if it's true that Paul Revere awakened Our First President, then if George Washington was Our First President, it is likewise true that Paul Revere awakened Our First President; and so on. Though it's quite true that there is a way of reading `refers to' as transparent[19] that is presumably *not* the reading that referential semantics has in mind when it says that causation determines reference. What's wanted, for the purposes of a referential semantics, is something compositional, and while (the word) `BUTTER' refers to (substance made from milk) `butter' meets this condition, `BUTTER refers to the yellow stuff that's on the table' does not. What shows this is that you can have the concept BUTTER even if you don't have the concept TABLE (and vice versa). The long and short is: you can't individuate conceptual contents by their causes (though it's left open that maybe you can individuate concepts by their intensions, meanings, senses, or the like). To put it more in the way that Quine does: If the presentation of a rabbit in the PC causes the informant to say `gavagai', then, presumably , the presentation of an assemblage of undetached rabbit parts does too, and versa. So if mental(/linguistic) representation requires compositionality, then the semantics of mental (/linguistic) representation isn't referential.

From our point of view, what's wrong with that line of argument, is that we don't hold that the content of a concept is its referent; we hold that the content of a concept is its referent *together with the compositional structure of the mental representation that expresses the concept.* PART is part of the mental representation UNDETACHED RABBIT PART, but not of the mental representation RABBIT; so the former may well be the content of our RABBIT concept ---the concept that we express with the word `rabbit' and that is a constituent of our rabbit -thoughts --- *UNDETACHED RABBIT PAR*T can't be. The moral: what's wrong with Frege arguments is also wrong with indeterminacy of reference arguments. Concepts can't be individuated just by their referents; but we never said that they could be.

In fact, there is some pretty persuasive empirical evidence that the content of the mental representation that expresses the concept RABBIT doesn't include the mental representation that expresses the concept PART (for a less abbreviated presentation of these results, see (RFEFERFENCE) Here as elsewhere S-R behaviorism underestimated the empirical constraints on psychological theories that experimental ingenuity may devise.

Consider a tracking experiment just like the ones described earlier in this chapter, except that the stimulus is not an array of dots but an array of `dumbells' (a dumbbell is a straight rod with weights on both ends). It turns out that subjects can track dumbbells but can't track their weights (unless we remove the rod that connects them.) Connecting the parts of a dumbbell creates a single new object, not just an arrangement of the parts of one, (see Scholl, Pylyshyn & Feldman, 2001). Since its weights are undetached parts of a dumbbell if anything is an

[19] For present current purposes, a context is `transparent) iff the substitutions of coreferrening expressions preserves truth in that context. Contexts that aren't transparent are said to be `opaque'.) See chapter 2.

undetached part of anything, the conclusion appears to be that what subjects see when they see dumbbells isn't *arrangements of undetached dumbbell parts* but, unsurprisingly, dumbbells. The world is the totality of *things* not undetached parts of things.

This result doesn't, of course, show that the reference of `dumbbell' is determinate; still less that it determinately refers to dumbbells. There are, after all, indefinitely many ways of specifying the extension of `dumbbell' and it remains open that some or other of these is an ineliminable alternative to `the members of the set of dumbbells'. (Quine mentions, for example, `rabbit stages' and `rabbit fusions`). The real point is that the constraint we want to impose on a semantics for mental representations, though not intensional, is much stronger than what Quine or Frege had in mind. Identifying the content of a representation with its referent together with its vehicle is a way of saying that mental representations that express concepts must be compositional. We doubt very much that Quine's sort of examples show that there're is residual indeterminacy when this condition is satisfied.

The reader will notice that we have carefully refrained from trying to define OBJECT or THING. We don't doubt that there are such things as things and objects; but we do doubt that there are such things as definitions (See Chapter 2). But, that said, we wouldn't be much surprised if trackability turns out to be the essence of thingness.

References cited in Chapter 4

Alvarez, G. A., & Franconeri, S. L. (2007). How many objects can you attentively track?: Evidence for a resource-limited tracking mechanism. *Journal of Vision, 7*(13), 1-10.

Biederman, I. (1987). Recognition-by-components: A theory of human image interpretation. *Psychological Review, 94*, 115-148.

Biederman, I. (1995). Visual object recognition. In S. M. Kosslyn & D. N. Osherson (Eds.), *Visual Cognition* (second ed.). Cambridge, MA: MIT Press.

Bowyer, K. W., & Dyer, C. R. (1990). Aspect graphs: An introduction and survey of recent results. *International Journal of Imaging Systems and Technology, 2*(4), 315-328.

Cavanagh, P., & Alvarez, G. A. (2005). Tracking multiple targets with multifocal attention. *Trends in Cognitive Sciences, 9*(7), 349-354.

Clark, A. (2000). *A Theory of Sentience*. New York: Oxford University Press.

Dawson, M., & Pylyshyn, Z. W. (1988). Natural constraints in apparent motion. In Z. W. Pylyshyn (Ed.), *Computational Processes in Human Vision: An interdisciplinary perspective* (pp. 99-120). Stamford, CT: Ablex Publishing.

Fodor, J. A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, Mass.: MIT Press, a Bradford Book.

Fodor, J. A. (2000). *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. Cambridge, MA: MIT Press.

Fodor, J. A. (2009). Enough with the norms, already. In A. Hieke & H. Leitgeb (Eds.), *Reduction: Between the Mind and the Brain*. Frankfurt, DE / New Brunswick, NJ: Ontos Verlag.

Fodor, J. A., & Piattelli-Palmarini, M. (2010). *What Darwin Got Wrong*: Farrar, Straus and Giroux.

Fodor, J. A., & Pylyshyn, Z. W. (1981). How Direct is Visual Perception? Some Reflections on Gibson's 'Ecological Approach'. *Cognition, 9*, 139-196.

Franconeri, S., Jonathan, S. J., & Scimeca, J. M. (2010). Tracking Multiple Objects Is Limited Only by Object Spacing, Not by Speed, Time, or Capacity. *Psychological Science, 21*(920-925).

Franconeri, S., Lin, J., Pylyshyn, Z., Fisher, B., & Enns, J. (2008). Evidence against a speed limit in multiple-object tracking. *Psychonomic Bulletin & Review, 15*(4), 802-808.

Franconeri, S. L. (In Press). The nature and status of visual resources. In D. Reisberg (Ed.), *Oxford Handbook of Cognitive Psychology*. Oxford, UK: Oxford University Press.

Franconeri, S. L., Pylyshyn, Z. W., & Scholl, B. J. (2012). A simple proximity heuristic allows tracking of multiple objects through occlusion. *Attention, Perception and Psychophysics, 72*(4).

Green, C., & Bavelier, D. (2006). Enumeration versus multiple object tracking: The case of action video game players. *Cognition, 101*(1), 217-245.

Hoffman, D. D. (1998). *Visual Intelligence: How We Create What We See*. New York: W.W. Norton.

Humphreys, G. W., & Riddoch, M. J. (1987). *To see but not to see: a case study of visual agnosia*. Hillsdale, NJ: Lawrence Erlbaum.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology, 24*(2), 175-219.

Kanwisher, N. (1991). Repetition Blindness and Illusory Conjunctions: Errors in Binding Visual Types With Visual Tokens. *Journal of Experimental Psychology: Human Perception and Performance, 17*(2), 404-421.

Keane, B. P., & Pylyshyn, Z. W. (2006). Is motion extrapolation employed in multiple object tracking? Tracking as a low-level, non-predictive function. *Cognitive Psychology, 52*(4), 346-368.

Kimchi, R. (2000). The perceptual organization of visual objects: A microgenetic analysis. *Vision Research, 40*(10-12), 1333-1347.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology, 4*, 219-227.

Koenderink, J. J. (1990a). The brain a geometry engine. *Psychological Research, 52*(2-3), 122-127.

Koenderink, J. J. (1990b). *Solid Shape*. Cambridge, MA: MIT Press.

Kolers, P. A., & Von Grunau, M. (1976). Shape and colour in apparent motion. *Vision Research, 16*, 329-335.

Leonard, C. J., & Pylyshyn, Z. W. (2003). Measuring the attention demand of Multiple Object Tracking (MOT). [Abstract]. *Journal of Vision, 3*.

Leslie, A. M. (1988). The necessity of illusion: Perception and thought in infancy. In L. Weiskrantz (Ed.), *Thought Without Language*. Oxford: Oxford Science Publications.

Liu, G., Austen, E. L., Booth, K. S., Fisher, B. D., Argue, R., Rempel, M. I., et al. (2005). Multiple-Object Tracking Is Based on Scene, Not Retinal, Coordinates. *Journal of Experimental Psychology: Human Perception and Performance, 31*(2), 235-247.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W.H. Freeman.

Marr, D., & Nishihara, H. K. (1978). Representation and Recognition of Spatial Organization of Three-Dimensional Shapes. *Proceedings of the Royal Society of London, B, 200*, 269-294.

Marr, D., & Nishihara, K. (1987). Representation and Recognition of the Spatial Organization of Three Dimensional Shapes. *Proc. Roy. Soc. London, 200*, 612-620.

Mervis, C. B., & Rosch, E. (1981). Categorization of Natural Objects. *Annual Review of Psychology, 32*, 89-115.

Milner, A. D., & Goodale, M. A. (1995). *The Visual Brain in Action*. New York: Oxford University Press.

O'Regan, J. K. (1992). Solving the "real" mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology, 46*, 461-488.

O'Regan, J. K., & Noë, A. (2002). A sensorymotor account of vision and visual consciousness. *Behavoral and Brain Sciences, 24*(5), 939-1031.

Parks, T. E. (1994). On the microgenesis of illusory figures: a failure to replicate. *Perception, 23*, 857-862.

Pentland, A. (1987). *Recognition by Parts.* Paper presented at the ICCV London.
Pylyshyn, Z. W. (1989). The role of location indexes in spatial perception: A sketch of the FINST spatial-index model. *Cognition, 32*, 65-97.
Pylyshyn, Z. W. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences, 22*(3), 341-423.
Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition, 80*(1/2), 127-158.
Pylyshyn, Z. W. (2003a). Return of the Mental Image: Are there really pictures in the brain? *Trends in Cognitive Sciences, 7*(3), 113-118.
Pylyshyn, Z. W. (2003b). *Seeing and visualizing: It's not what you think*. Cambridge, MA: MIT Press/Bradford Books.
Pylyshyn, Z. W. (2007). *Things and Places: How the mind connects with the world (Jean Nicod Lecture Series).* Cambridge, MA: MIT Press.
Pylyshyn, Z. W., & Annan, V., Jr. (2006). Dynamics of target selection in Multiple Object Tracking (MOT). *Spatial Vision, 19*(6), 485-504.
Pylyshyn, Z. W., Elcock, E. W., Marmor, M., & Sander, P. (1978). *Explorations in visual-motor spaces.* Paper presented at the Proceedings of the Second International Conference of the Canadian Society for Computational Studies of Intelligence, University of Toronto.
Quine, W. V. O. (1960). Word and Object. Cambridge, MA: MIT Press.
Richards, W. (1980). Natural computation: Filing a perceptual void. Paper presented at the The 10th Annual Conference on Modelling and Simulation, University of Pittsburgh.
Rock, I. (1983). *The Logic of Perception*. Cambridge, Mass.: MIT Press, a Bradford Book.
Rosch, E. H., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*, 382-439.
Scholl, B. J. (2009). What have we learned about attention from multiple-object tracking (and vice versa)? In D. Dedrick & L. Trick (Eds.), *Computation, Cognition and Pylyshyn* (pp. 49-78). Cambridge, MA: MIT Press.
Scholl, B. J., & Pylyshyn, Z. W. (1999). Tracking multiple items through occlusion: Clues to visual objecthood. *Cognitive Psychology, 38*(2), 259-290.
Scholl, B. J., Pylyshyn, Z. W., & Franconeri, S. L. (1999). When are featural and spatiotemporal properties encoded as a result of attentional allocation? *Investigative Ophthalmology & Visual Science, 40*(4), 4195.
Sekuler, A. B., & Palmer, S. E. (1992). Visual completion of partly occluded objects: A microgenetic analysis. *Journal of Experimental Psychology: General, 121*, 95-111.
Spelke, E. S. (1990). Principles of object perception. *Cognitive Science, 14*, 29-56.
Sperling, G. (1967). Successive Approximations to a Model for Short Term Memory. *Acta Psychologica, 27*, 285-292.
Sperling, G., & Weichselgartner, E. (1995). Episodic theory of the dynamics of spatial attention. *Psychological Review, 102*(3), 503-532.
Strawson, P. F. (1959). *Individuals: An essay in descriptive metaphysics*. London: Methuen.
Treisman, A. (1988). Features and objects: The fourteenth Bartlett memorial lecture. *The Quarterly Journal of Experimental Psychology, 40A*(2), 201-237.
Trick, L. M., Jaspers-Fayer, F., & Sethi, N. (2005). Multiple-object tracking in children: The "Catch the Spies" task. *Cognitive Development, 20*(3), 373-387.
Ullman, S. (1979). *The interpretation of visual motion*. Cambridge, MA: MIT Press.
Ullman, S. (1984). Visual routines. *Cognition, 18*, 97-159.

This page deliberately left blank

Chapter 5 Reference Beyond The Perceptual Circle

Why do we go on so about what does and doesn't happen in the in the Perceptual Circle (PC)? Not, certainly, because we think what paradigm Empiricists did: that all one's beliefs are, or reduce to, beliefs about one's sensations and perceptions. Rather, it's because we think that, if a theory of reference is to be of use to cognitive science, it must be naturalistic; and we think that if a theory of reference is to be naturalistic, it must posit a causal chain that runs from the things that thoughts are about to tokens of mental representations that refer to them; and we think that that, for the case where the referent is a current object of perception, there is at least a first approximation to a story to tell about the character of such causal chains: It's some or other variant of the one we told in ch4: At the early stages of perceptual processing, transductive mechanisms provide preconceptual representations of perceptible features of the light reflected from things and events in the PC. At some later stage, the perceptual system computes from such transducer outputs representations in a specialized (perhaps geometrical) vocabulary. `Basic level` representations of distal percepts are derived from these, perhaps by the application of a form-to-category dictionary. (Since it looks like this, it's probably a swan) We aren't at all inclined to insist on the details; but, given the current drift of research in Vision Science, it strike us as plausible that some story along those lines might eventually be forthcoming.[1]

Complaint: " But still, everything that you've discussed so far has been about *perceptual* reference; and, surely, it is perfectly possible to refer to things that aren't currently objects of perception, hence to things that aren't 'in the PC'). So, even if, *per impossible*, every word of Chapter 4 is true, theories of referential content and theories of the fixation of perceptual beliefs are both in need of a vocabulary of mental representation that apply to things that *aren't* currently in view: For example, remote thing, past things, future things, characters in novels, perfect triangles, `theoretical entities' etc can all be referred to and thought about, though none of them can be currently perceived, and many of them can't be perceived at all. And, perhaps worst of all, referents of concepts that weren't in available to perception at one time have been known to enter the PC later on; PARAMECIUM is a paradigm case.[2] So then: What about reference *beyond* the perceptual circle? If the kind of naturalistic theory you've been pushing can't make sense of that, what good it is it to psychologists (or to anyone else?)

That is, we think, a perfectly reasonable objection; but we still aren't ashamed of endorsing a semantics that stresses the primacy of reference to things in the PC. On anybody's story, things that are in the PC have semantic properties that things that aren't in it don't; thus, only things in the PC can be

[1] It bears emphasis that, although mental representations may be structured (eg. geometrically) at the stages of visual perceptual processing that we've been discussing, and although they refer to distal objects, the mental processes that take place are nonetheless assumed to be encapsulated. In particular, the vocabulary available to refer to basic-level distal objects is exiguous compared to what is routinely available to express the contents of perceptual beliefs at large. Maybe the former includes: ROUND, TRIANGULAR, ABOVE, BELOW, , BLUE, FRONT, BACK, NEARER TO, FURTHER FROM and other concepts that occur in typical basic-level representations; but it's assumed not to include: JOHN, JOHN'S TROUSERS, THE HOUSE NEXT DOOR, etc, even though all of those express properties that *non*basic objects can be perceived to have. Trousers are, perhaps basic-level objects, in which case, the concept TROUSERS may appear in the form-to-basic-category dictionary. But `John's trousers' surely isn't a basic level description (it's a mark of basic-level predicates that they are typically morphologically simple.) Accordingly, John's trousers are seen as such only if they are first seen as trousers.

[2] It used to be widely held (by pragmatists, verificationists, operationalists, procedural semanticists and the like) that the distinction between `observable' and `theoretical' entities is central not only in epistemology but also in ontology and theories of conceptual content. For example, that (`strictly speaking') to say that the invention of the microscope made it possible to see paramecia is to equivocate either on `paramecia' or on `see' (or both). This kind of thesis strikes us as a *reductio* of the ontology and semantics that lead to it and has now gone largely out of fashion (however, see the appendix to Chapter 2).

referred to by *demonstratives*, which, according to us, play a vital role in the earliest stages of perceptual representation. (Since only things that are in the PC can be FINSTed;, the FINST idea accords comfortably with the idea that demonstratives are a primitive kind of mental representation.) Or, for another example, and with only one kind of exception we can think of; you can't *attend to* things that aren't in the perceptual circle, though you can, of course, think about them, remember them, pine for them, and so on.[3]

We do admit that, at this point, that it would be convenient to have intensions to lean on since, according to the Fregeian tradition, a concept's *ex*tension contains all and only what satisfies the concepts *in*tension, so (demonstratives aside) it is not required that concepts that apply *beyond* the PC, also apply *within* the PC. But if one thinks, as we certainly do, that that the demands of naturalism require mental states to have causal powers, thereby excluding a semantics of senses for mental representations) what except reference is there left for conceptual content to be? It's worth noting that it's not only Fregians who have irons in this fire. If there are no intentensions, that is, to be sure, an embarassment for Frege. But likewise, what kind of content other than reference would be left to supervene on the `language-games', `rules of informal logic', `conceptual roles' etc. of which `analytic' philosophers are so fond? What with one thing and another, a naturalistic theory of content –one that eschews intensions--- had better have something to say about reference to things beyond the perceptual circle. Some hard cases follow; we're not at all sure that we've covered all that such cases; but maybe these will suffice to suggest how others might be managed.

Too far away

Perhaps the least perplexing kind of reference to things outside the perceptual circle is to things that are currently too far away to be perceived. Surely we can refer to such things both in language and in thought? What might a naturalistic account of reference make of that?

We could try leaning on causal counterfactuals (a token of such and such a mental representation *would have been* caused by such-and-such if such-and-such *had been* in the PC); and we will indeed, opt for that sort of treatment at various points in our story. But counterfactuals won't do the job here. We hold that tokenings of mental representations have the referents they do because they come at the ends of causal chains of which the referents are among the links. But it's hard to see how the mere truth of a counterfactual could cause anything . First blush, the fact that seeing Bossie *would* cause me to believe that Bossie is a cow *if she had been here* could explain my actually believing that Bossie is a cow only if, in fact, Bossie *is* here. By assumption, Bossies that are very far away can't be current percepts; so it seems that, even if a causal theory might work for *perceptual reference,* it couldn't work for the reference of thoughts.

Would it help to give up and accept the more or less Empiricist view that other-than-perceptual references to Bossie are constructs out of sensory experiences? (That wouldn't, of course, entail that Bossie herself is a `bundle of sensations'; it's presumably common ground that Bossie is protoplasm through and through). But this only delays the problem, which reappears in respect to the reference of the sensory constituents that *non*-perceptual beliefs are supposed to be constructed from. Consider, for example, my current belief that Bossie is brown; and suppose that I am not now sensing either Bossie or brownness. According to the Empiricist picture, the content of my current perceptual beliefs must be

[3] Sensations: pains, afterimages, ringings in your ears and the like, are counterexamples, assuming that feeling an itch, hearing a ringing in your ears and the like counts as perceiving it.) Sensations are thus like percepts in some respects but not in others, and we don't understand why that's so. Many philosophers and psychololgists have tried to explain it; but, none of them has succeeded the best of our knowledge. Dire epistemological issues turn on this since it's often said that sensations are the sorts of mental things that their owners can't be mistaken when they say/think they that they are having. We're neutral however; we don't do epistemology.

constructed out of current sensations of brown. So, what about the content of my belief that Bessie is brown when she *isn't* in the PC? You can, after all, believe that Bossie is brown with your eyes closed.[4]

Why couldn't the `brown` constituent of my current belief that Bossie is brown be a *memory* of one of my *past* sensations of brown?" Because that too leaves us back where we started. We were worried about how, if not by satisfying an intension, anything that isn't currently being experienced could be the referent of a current token of a mental representation. Well, sensations that I only *remember* having had aren't currently being experienced; so the question now arises how *they* could be I the extension of tokens of concepts like SENSATION I REMEMBER HAVING HAD. Empiricits tended to gloss over this problem because they often held that remembering a token of a past sensation involves behaving a token of that sensation (Remembering a sensation as `replaying' the sensation). But this is no good. Remembering an agonizing pain doesn't require that you have one now.

We've been imagining a program for constructing a causal theory of reference that starts with reference to things in the PC and works out from there. To which you might object that perceptual reference must be different from reference to things outside the PC because, tautologically, things outside the PC aren't percepts. But that ignores the *mobility* of one's Perceptual Circlle*;* your PC follows you around as you go from place to place. That being so, you can have perceptions of brown (or of Bossie) practically whenever you wish. All you need to do is: find something brown to have a sensation of (mutatis mutandis, find Bossie and have a sensation of her).

Which link?

Many people who deny that reference (even *perceptual* reference) can be naturalistically explained think that's because there's a puzzle about which link in a causal chain is the referent of a mental representation. The worry is that, unless there is there`s an 'interpreter' on the scene to assign a referent to the representation, there would be `no fact of the matter` about which link in the chain is the one being referred to. And, of course INTERPRETATION is itself a semantic/intensional concept and so is itself in want of naturalization. (fn)?? (references See, eg. Dennett ( ), Davidson ( ), Quine ( ). But we think this line of thought overlooks a plausible option: Namely that the right link can be located by the (actual or counterfactual) applications of a procedure of `triangulation', (which may be found in any book on coastal navigation qv). For elaboration, see Fodor, LOT 2. (REFERENCE).

Too long ago

Still, what we've said so far couldn't be the whole story about reference to things that aren'in the PC. For, even vIf perceptual reference is supposed to be the basic kind, we still have to worry about things that aren't in anybody' PC see *now,* but once were. And, of course, my perceptual circle isn't mobile in *time* in the way that it is in space. At best, I can *remember* what I saw; *pace* Proust, I can't go back and have another look. Suppose, for example, that Mentalese contains proper names; and assume, for the sake of argument, that proper names are primitives (rather than abbreviations of descriptions.) Still, though I have never seen Pericles, I can refer to him; I can think about him and talk about him. What could an intension-free semantics say about that?

Kripke tells a story about how it is possible for him, here and now, to refer to Moses: Roughly, someone once baptized Moses `Moses' (using, presumably, the Hebrew equivalent of `Moses' to do so). In consequence, people who were present at the baptism came to use `Moses' as Moses' name for Moses. Then people who *weren't* present at the baptism heard about it from people who were, and consequently they too came to use `Moses' to refer to Moses. And so on, link by link, along an arbitrarily

[4] It's not clear to us why merely counterfactual interpreters mightn't do for our purposes since, it is presumably not the case that interpreters serve as links in the causal chains connect referents to perceivers that they interpret. But never mind.

long chain that started with Moses being baptized and eventually got to Kripke's using `Moses' to refer to Moses.

That seems entirely plausible. But Kripke apparently *doesn't* think that it would meet the demand for naturalization (which is, anyhow, not a project in which he seems to be much interested.) His reason is that the chain of transmission that connects him to Moses starts with a baptismal *intention* (the intention that the person baptized should henceforth be called `Moses') and, moreover, continues via the mental states (beliefs, memories, communicative intentions and so forth) of people who came to use the name in consequence of the baptism; and, on pain of the usual problems about circularity, naturalists aren't allowed to take beliefs, memories, intentions and the like for granted in explaining how metal or linguistic representations come to have the referents that they do.

We don't find that objection convincing. Let's assume that a causal-transmission-chain story would be a more or less correct account of how it cones about that our current utterances of `Moses' (/thoughts about Moses) can refer to Moses. It seems to us to mistaken to argue, as we take Kripke to do, that naturalists are prohibited from telling that story about how `Moses' got from Moses to us. It's true that, if the transmission of baptismal intentions were proposed as a metaphysical account of what *reference is* (what it is for `Moses' to refer to Moses), then it would indeed be circular. But the story about the transmission of baptismal intentions *doesn't* purport to be a Theory of Reference in that sense; rather, it's a theory of reference *transmission.* According to our kind of naturalist, reference consists of some sort of causal relation between a representation and the thing it refers to. According to our kind of naturalist, such chains are grounded in perceptual reference. The story about the transmission of reference along a causal chain is supposed to explain how, *assuming that a reference-making mind/world connection is in place,* it can be inherited from minds to minds over time. De facto, the causal chains that connect our mental (/linguistic) representations of things in the future, like mental representations of things in the past, include, in all sorts of ways, tokenings of beliefs, memories, intentions etc. among the links of causal chains. But why should that perplex a naturalist? *Transmission* of reference is constituted by causal relations between people over time. But reference itself is a causal relation between mental representations and the things-in-the-world that they represent. A theory about the transmission of content can perfectly legitimately take contentful mental states for granted, even though a theory about what content *is* mustn't do so on pain of circularity.

Still, it might objected, even if a Naturalistic account of the *transmission* of reference needn't make unexplicated appeals to intensional states and processes, still Kripke's story about how Moses came to be so-called does so. That's because Kripke holds that *the initial* causal link In the transmission chain --- the one that initially fixes the reference of `Moses'--- is a baptism; and a baptism is ipso facto an intentional act, an act that is *intended* to be a naming. So an attempt to naturalize one of Kripke' s chains of reference-transmission would, of necessity, fail at the first link. But that argument too is unconvincing because, though transmission-chains *can* be initiated by baptisms, they certainly don't need to be. Names can be just `picked up`. Suppose that, in a fit of pique, I say to my cat `you idiot, don't do that'. And suppose that, having happened to have overheard my saying it. It is then perfectly possible that you should come to believe (albeit quite wrongly) that my cat's name is `You Idiot', and hence come to so refer to him. The next link in the chain might then pick up this putative name of my cat from you…. and so forth. So the practice of referring to my cat as `You Idiot' might be both initiated without anybody ever having the intention that my cat should be so called. Why, in principle, shouldn't a reference chain that started in that way lead to Kripke's using `Moses' to refer to Moses? The long and short is that a theory about how reference is transmitted over time requires a causal chain that runs from an instance of reference fixation to an instance of reference inheritance. But, in principle at least, it doesn't matter whether any of the links in the chain are (including even the first one) is intentional, so long as there actually are things in the world that have the power to cause tokens of mental representations when they turn up in the PC.

Empty concepts

We remarked in Chapter 2 that the existence of `empty' but nonsynonymous concepts (concepts that are intuitively different in their semantic contents but both have null extensions) is a prima facie argument against the view that reference is all that there is to the semantics of mental representations. If reference is content, why aren't `griffin` and `unicorns' synonyms? Worse still, from our point of view, empty concepts seem to favor some sort of Inferential Role semantics over a purely referential one. `is a Griffin' and `is a unicorn' license different patterns of inference; `unicorn' implies has a horn, `Griffin' does not; maybe the intuitive difference between their contents supervenes on that. We think that there are, in fact, a number of different kinds of empty concepts; fictional characters work a little differently from frictionless planes, for example (see below). We will spare you any extensive enumeration; too much ado about nothing. But perhaps discussions of a few kinds of cases will persuade you that empty extensions don't make a conclusive case against referential semantics; still less a conclusive case for IRS.

A natural suggestion is that counterfactuals might take up some of the slack. There aren't any frictionless planes; but, because we know a lot about the laws of mechanics, we know a lot about how frictionless planes *would* behave if there *were* any: If a frictionless perfect sphere rolled down a frictionless inclined plane, its velocity would increase uniformly as a function of the inclination. Even though there aren't any frictionless planes or perfect spheres, *there are laws that would apply to them if there were. Why* shouldn't such laws make corresponding counterfactuals about them true. And why shouldn't such counterfactuals about unicorns be quite different from the counterfactual causal relations that griffins are in ?It's true, of course, that rhetorical questions aren't arguments; but we're not, at the moment, arguing for referential semantics; we're just trying to block some prima facie embarrassing objections to it.

Reply: "it's a mistake to assimilate unicorns to frictionless planes because, though there are, as you say, arguably laws about the latter, there are none about the former. Consider, for example, `if there were unicorns, they'd be herbivores' or even `if there were unicorns, they'd have hearts'. Which, if either, of these is true? The only counterfactual about unicorns that we know for sure to be true is `if there were unicorns, they would have only one horn'; which is true by definition.'

Fair enough; we warned you that saving referential semantics from empty extensions might require distinguishing among empty extensions of different kinds. Accordingly, our story about the semantics of `frictionless plane' is different from our story about the semantics of `unicorn'. Unicorns are *fictions*; frictionless planes, by contrast, are extrapolations from theories of mechanics that are independently certified. There are true counterfactuals about frictionless planes because there are laws of mechanics. There aren't any true counterfactuals about unicorns because there aren't any laws about unicorns (not even physical laws; we think `If there were unicorns, they could run faster than light' is neither true nor false.)

Further reply "What about square circles? It's not just that there *aren't* any; it's that there *couldn't* be any. So the extension of `square circle' is *necessarily* empty, as is the extension of `round triangle. Nevertheless the semantic content of `square circle' is intuitively different from the semantic content of `square triangle`, which is again different from that of `round triangle."

We don't think this kind of consideration refutes referential semantics, but we agree that it has interesting consequences. As far as we can see, a Purely Referential Semantics would require *that no concept whose extension is necessarily empty can be primitive.* There couldn't, for example, be a syntactically simple mental representation that denotes all and only square circles. The reason is straight forward: as we've been describing it, PRS implies that the extension of a representation is determined *by actual or counterfactual causal interactions* between things in its extension and its tokens. But there are no possible (a fortiori, no actual) causal interactions between square circles and *anything* because `there are no square circles' is *necessarily* true. So if there is a concept SQUARE CIRCLE (we're prepared

to assume, if only for the sake of argument, that there is), and if PRS is true, then the concept SQUARE CIRCLE must be compositional, hence structurally complex, hence *not* primitive.[5]

Fictions

Maybe. But think about, say, `Tonto` and `The Lone Ranger` (taking both expressions to be `true' names; i.e. names rather than shorthands for descriptions). It's plausible that certain truths about the extensions of true names are analytic: It's analytic that the name `Tonto' refers to Tonto or to no one; `It's analytic that the name `George Washington` refers George Washington or to no one; and so on. So even if a purely referential semantics can be made to cope with all the kinds of empty-extension cases discussed so far, it can't cope with intuitions about the content of the names of fictions: If The Lone Ranger and Tonto are fictions, there aren't even any *possible* worlds in which `The Lone Ranger and Tonto are other than coextensive, (i.e. both empty). So, according to referential semantics, they are *necessarily* synonymous. IRS, by contrast, entails no such embarrassment; it can hold that your intuition that `Tonto' and `TLR' aren't synonyms arises from your knowing that Tonto and TLR *play different roles* in the familiar story. If, in short, conceptual contents supervene (not on extensions but) on conceptual roles., it's maybe[6] ok for The `Lone Ranger` not to refer to Tonto after all.

The idea that content supervenes on inferential role has many enthusiasts, in both philosophy and cogsci. But the more you think about it, the more it seems to beg precisely the questions that semantic theories are supposed to answer. For, what's the difference between saying, that Tonto and TLR have different roles in the story and saying, that *the story represents* Tonto and TLR as different people? And if, as we rather suspect, these are just two ways of saying the same thing, then the putative example of how terms with the same (empty) extensions can nevertheless differ in content turns on a notion that is itself intensional: namely the notion of *representation in a story. Since the notion of a `role' is intensional, it's unsurprising, and not deeply very moving, that an Inferential Role Semantics could legitimize distinguishing between coextensive names.* Shifting the polemical turf from intentions about, meaning to intuitions about inferential roles isn't much use to naturalist if the explication of the latter begs the case against him.[7]

Still, it should be admitted that names fictions are prima facie counterexamples to purely referential Semantics. Quibbles aside, what PRS really wants to say is that *all empty concepts have the same semantic content* in any sense of semantic content that it is prepared to take seriously. And according to intuition that's just plain false. Intuition says that `Tonto` and `TLR` aren't synonyms. But we think that intuition is misled in saying this. Chapter 2 suggested a way of explaining why that is: Practically every term/concept comes with an aura of associations, attitudes feelings, beliefs, quasi-

[5] It could be, however, that the complexity of SQUARE CIRCLE is, as one says, ' merely implicit'; that it appears only when the compositional structure of the concept is unpacked; i.e. at what linguists often (misleadingly) call `the semantic level'; i.e. at the level of logical syntax.

[6] The note of caution is because this sort of argument turns on some very uncertain assumptions about the modal behavior of proper names (in particular that they are `rigid designators'; and also because there isn't, in fact, any more of a story about the individuation of `conceptual roles` than there is about the individuation of meanings. (see Ch. 2) Indeed, the familiar dilemma about holism vs analyticity) arises in the former just as it is in the latter.) How different must storiest be for `TLS' to refer one but not the other?

[7] Naturalists who like IRS often for granted that the *inferential* role of a concept can be explicated as something like *its causal role in mental processes.'* (See eg Block (reference)) But it's hard to see how this could work unless the distinction between `mental` processes and others assumes that the former are themselves ipso facto intensional; in which case the familiar circularities reappear. This is a special case of a point we will presently emphasize: *prima facie*, `dog' and `cat' are different in inferential role; `that's a dog' implies `that's a canine` but `that's a cat' implies `that's a feline'. But this constitutes a difference in inferential role only on the assumption that `canine' and `feline' *aren't synonyms.* We strongly suspect that IRSs that claim to be naturalistic are doomed to running in this sort of circle.

beliefs, recollections, expectations and the like. (Bertrand Russell (of all people) once held that `or' expresses an attitude of uncertainty!) We think it's quite plausible that auras can, and often do, influence intuitions that purport to (but don't) reflect bona fide semantic truths

Methodologically speaking, that's a good reason why cognitive scientists (linguists and philosophers very much included) should take intuitions with a grain of salt. This applies with special force to the case of characters in fictions because so much of the aura that one's concept of a character has is supplied *by the fiction.* Hamlet was chronically indecisive. How do we know? Because 'Hamlet' shows us that he was. Othello was noble but a dope. How do we know this? Likewise, Iago was a plain stinker. How do we know? And so forth. One of the things that authors of fictions do for a living is construct auras for an audience's concepts of their characters; and, roughly speaking, the stronger the aura, the more `life-like' the fiction. Shakespeare was *very* good at that. It's a middle-brow cliché that his characters seem to 'step off the page'. One could almost believe in Rosalind or Beatrice, if they weren't too good to be true. But (according to us) *none of that has anything to do with semantics.* What shows is that, like inferential roles, auras, don't support paradigm semantic properties like synonymy and modality.[8] It doesn't follow that auras can't *explain* such intuitions when they are mistaken. Intuitions are sometimes *evidence* for (/against) semantic theories; but they are never *constitutive* of the truth (/falsity) of semantic theories. They work much the way that observational reports do in the hard sciences; one has to take them seriously, but they're fallible.

So much for what issues non-synonymous concepts with empty extensions do and don't raise for referential semantics. We do agree, of course, that they are *prima facie* counterexamples. But we don't agree that they are remotely decisive *prima facie* counter-examples, and we won't give up on PRS unless we're really forced to; because (we may have mentioned this before) we think that PRS is the only available candidate for a naturalistic science of cognition, and we do think that, in the Cognitive Sciences as in the others, naturalism is *sine qua non.*

Too small

It is, we think, undeniable that some things are too small to see. And there are things that we can see now that used to be too small to see were once too small to see that we can see some things We can now see paramecia and the like by using microscopes; and saying that we can doesn't equivocate on `paramecium' or on ` see'. But, even back when paramecia were too small to see, they could perfectly will be referred to. Accordingly, a viable metaphysics of reference must be able to explain how they could have been referred to even when they were in anyone's PC. The answer seems clear enough from the point of view of the kind of causal theory of reference that we endorse: Reference requires a causal chain connecting tokens of mental representation to their referents. In paradigm cases of visual perception, one of the links in the chain is the reflection of light from a distal object onto a sensory transducers that respond to them. It used to be that that there were no such chains, but then Lewonhouk invented the microscope and there are some now. So it's perfectly ok to say that, even when they couldn't see paramecia, people could refer to them.[9]

[8] For example, it can't be *semantically necessary* that Iago was a stinker; he might have been quite a good fellow if his childhood had been happier.

[9] "What about things that are too small for anyone *ever* to perceive them; protons, for example?" Well, adding appropriate links to causal chains has made it possible to *detect* protons. Does seeing a proton's trace on a photographic plate count as seeing a proton? Or does it only count as seeing a trace of a proton? If you think that deep issues in epistemology (or in ontology, or in semantics) turn on this (as philosophers did who cared a lot about the distinction between `observations' and `theoretical inferences '), you are free to adjust the boundaries of the PC to suit your philosophical commitments. All that matters, according to our understanding of what content is, that though such ajustments may affect what one says about seeing, they needn't affect what one says about referring. what one says about *referring.* We think that is itself an argument for understanding content in the way we've been urging you.

Too big

II there any such thing as The Universe, or is `the universe' just a hypostatic way of saying `the totality of things` (Cf Ryle: `the university' as a hypostatic way of saying `all the colleges taken together'). And, if there is such a thing as the universe, is it possible, to refer to it according to a causal account of reference? Presumably that would depend on whether the universe can cause anything; on whether the universe can cause tokenings of mental representations; well, can it?, we suspect that such questions are merely frivolous. But, If they aren't, whose job is it to answer them? The semanticist's? The ontologist's? The physicist's? The metaphysician's? In any case, better them than us. We think a theory of reference is only obliged to cope with reasonably clear cases where a symbol refers. We'd be more than content with that.

Properties

`Abstract objects don't cause things. But we can refer to (not just instances of red but also) the color red; and to the *property of bein*g *red*; and both of these are abstract objects. So causal theories of reference must be false'. What, if anything, is wrong with that line of argument? It would be nice if there were some way around it since, as we've several times remarked, taking conceptual content to be referential and reference to be causal promises to avoid embarrassments that intension-based theories seem not to be able to dodge. These include not just worries about naturalism but, as Chapter 2 pointed out, by dispensing with notions like SENSE and INTENSION, referential semantics deflates `Fodor's Parardox', which purports to show that if concepts are individuated by senses, there can be no such thing as concept learning. This bears a little looking into.

Suppose what Fodor took for granted: if concepts are learned at all, it must by some process of hypothesis projection and confirmation (how else, after all, *could* they be learned?) It then looks as though all concepts will have to be innate. The Churchlands have suggested (REFERENCE) that this argument is a sort of *reductio* of the whole idea of a computational cognitive psychology; and even Fodor, in his darkest moments, has found that uninviting. By contrast, a referential theory of conceptual content can just shrug its shoulders. If the content of a concept is its reference, all you need to learn BACHELOR is *some way or other* of representing its extension; any mental representation that is coextensive with BACHELOR (any description that is true of all and only bachelors) will serve for that. In particular, it's *not* required that the mental representation of bachelors that mediates the learning of BACHELOR should be a synonym of `bachelor` since, according to a Purely Referential Semantics, there are no such things as synonyms. It's an agreeable consequence of referentialism that it takes concept acquisition to be at least *possible;* which, arguably, sense-semantics can't do.[10] We think that what makes the main trouble for a referential-causal theory of conceptual content aren't the issues about how concept learning is possible; it's how there can be reference to abstracta. In particular, since senses and the like are abstracta, a causal-referential theory of content is sorely in need of an alternative to `senses determine extensions'. But if senses don't, what does? We continue to hold the most promising answer to be that, in the central cases, conceptual content *is* referential, and reference *is*, (or supervenes on) a causal relation between mental representations and their referents. But, since it's common ground that abstracta don't have causal powers, how, could there be concepts that refer to abstracta? How, for example, could there be concepts of redness, or of being red?

How serious is this? We think it's no worse than not hopeless. To begin with, we think that reference to a property must somehow be grounded in reference to (actual and possible) individuals

[10] It doesn't, of course, follow from PRS that some, all, or even any, of one's concepts actually *are* learned; it doesn't even follow that there *is* such a thing as learning. All those are left as empirical issues (as they should be; ethology is not an a priori science, legions of philosophers to the contrary notwithstanding.)

that have that property. Merely possible individuals don't, of course, have causal powers any more than abstracta do. But we suppose that part of a naturalistic account of property-reference will appeal to counterfactuals of roughly the form `*if there were* an *x* that has property P, it *would* cause …' where the *x*'s can be possible-but-non-actual. (If the naturalist we're trying to placate won't allow us counterfactuals, so be it. But we doubt that intensional psychology is the only science that his very exiguous sort of naturalism would preclude, since it is plausible that the notion of an empirical law is very much bound up with counterfactuals, and we suppose that there are (or anyhow may be) laws of cognitive psychology.[11]

Consider a kind of example that Joe Levine is fond of: Suppose you say (or think): 'I want my house to be the color of that house'. This doesn't, of course, mean that I want the color to be stripped off that house and glued onto mine in the way that frescos sometimes are. What I want is that my house should be *a token of the same* color *type* as that house. But to what does the `color' refer in `token of the same color type as…?' Plausibly, it refers to a property; viz the property of which the color of that house is an instance, and of which I want the color of my house to be likewise an instance. But, to repeat, the question, how could an utterance (/thought) refer to a property if properties are abstracta and reference is a causal relation between representations and their referents? Properties don't have effects (though a state of affairs consisting of a property's being instantiated by an individual perfectly well can.) This distinction may strike a *very* hard-headed psychologist as frivolous, but it isn't. All sorts of nasty equivocations result if one doesn't attend to it. Consider `the color of Sam's house infuriates John'. Does that mean redness infuriates John (John is red-phobic), or does it mean that it infuriates John that Sam's house instantiates redness (he doesn't care if some other houses are red, but it bothers him that Sam's is)? If you think that distinction is frivolous too, imagine that you are a developmental psychologist who is speculating on the etiology of John's pathology; or a clinical psychologist who is trying to cure him of it.

One could understand the problem `how should a purely causal theory of reference cope with reference to properties?' as raising an issue in ontology; that way of reading it would take the question to be `*what kind of thing* could a properties be such that a Purely Referential Semantic iis true of terms that refer to them. But we don't recommend that you read it that way; this ontological, is, we think, one of the things that makes the problem about reference to properties seem not just hard for a causal theory but hopeless on the face of it. It would, for example, be a mistake a Purely Referential Semanticist to ask `what sort of thing-in-the-world (or even in Plato's heaven) bears the same reference-making relation to the expression `red' in `red is a color' that *this apple* bears to the `this apple' in a token of `this apple'. It's a mistake because, on the one hand, the semantic value of `apple' in `this apple' is its referent, and the content-making relation between referring expression and their referents has to do with tokens of latter causing tokens of the former; as we keep saying, it's a truism that properties don't have effects. So it might well seem that either causal/referential semantics isn't true or the truism is false. It appears that the cost of endorsing such a semantics is having to conclude that there isn't --- couldn't be--- *anything* that property-terms refer to. Many a nominalist has indeed argued in more or less that way.

[11] The notion of a law of nature is, of course, intensional, and there are laws of physics (chemistry/ mechanics etc.) that apply to chairs, but, on our view, none of those apply to chairs *as such.* But it would be a mistake to argue, by analogy, that there are no *psychological* laws about chairs. To the contrary, it seems entirely plausible that if a chair is in the perceptual circle (and if the appropriate background conditions are satisified) then it is *nomologically necessary* that it would cause certain things to happen; (in particular that it would cause certain mental representations to be tokened) in minds like ours. It might, for example, be nomologically necessary, that if a chair were now in your perceptual circle, it would now cause a token of a Mentalese representation that co-referential with the word `chair' to now be tokened in our minds . That sounds OK to us.

But it's a false dilemma: Causal/referential semantics, doesn't say that there are no properties (or even that there aren't *really* any properties). Come to think of it, we're not at all clear what it would it would mean to say either of those things. Rather, PRS should say that what the concept RED contributes to the semantics of THE COLOR RED isn't its referent; what it contributes is *a possession condition.* In particular, to have the concept THE COLOR RED one must have the concept that `red' expresses in (for example) `that red house'; roughly, it's to have a Mentalese Representation type that is caused to be tokened by red things as such. Since we are, for better or worse, already committed to a sufficient condition for having the concept that `red' expresses in `that red house' the semantics of `red' in `the color red' doesn't show that the causal-referential semantics game is over. It does. however, suggest that our `Purely Referential' semantics isn't won't, after all, be *purely* referential.[12] Having the concept that `red' refers to in `the color red' is having a mental representation type that is caused to be tokened by (actual and possible) things that are red; so the `red' in `the color red' expresses a *second order* property; it's a property that representations have in virtue of the causal properties of *other* representations; in particular, in virtue of their power to be caused by instances of redness. That seems reasonably natural since, from the metaphysical point of view, properties do seem to be second-order sorts of things. For there to be the property *red* is for some (actual or possible) things to be instances of redness. No doubt, if that is taken to be a *definition* of `the property red', or as the *sense* of the corresponding concept, it would be hopelessly circular. But we don't intend any such thing; according to us, concepts don't have definitions or senses; all they have is extensions. Accordingly, all you need in order to specify the content of a concept is a description that has the same extension that the concept itself does. Having the concept of the color red' is having a mental representation that is caused to be tokened by actual and possible red things.

* * *

So where there has the discussion in this Chapter gotten us? There are a number of prima facie *reasons* for thinking that a naturalistic theory of conceptual content would have to be some sort of causal theory of reference; and there are also a number of prima facie reasons for thinking that the content of one's concepts is constituted, in the first instance, by a relation between one's mind and things in one's perceptual circle; in effect, all other content-making relations are grounded in these. That may strike you as very bad news for Sherlock Homes, Julius Caesar, merely prospective grandchildren, protons, numbers, true triangles, the color *red*, and so on. So it may appear, that there is a large and heterogeneous collection of prima facie counterexamples to the only kind of naturalistic semantics that we've been able to think of. We would find that depressing except that that there isn't (actually) anybody that `Sherlock Homes' refers to; and, though `Caesar' and `proton' both have referents, we think that's compatible with a theory that says that they having referents depends on their tokens being caused by Caesar and protons respectively. That's just as well. Cog Sci *needs* a naturalistic semantics because it needs a viable version of the Representational Theory of Mind. It needs a viable version of the Representational Theory of Mind because it needs a naturalistic account of the content of mental representations; and. far as we can tell, RTM is the only serious candidate currently in view.

So here's what we take to be the polemical situation: On the one hand, it isn't disputed (except, perhaps, by philosophers attracted to a `phenomenology of the present moment', of whom there no

[12] The idea that some (or even all) concepts are individuated by their possession conditions is, of course, by no means novel; see, for example, see, for example Sellars, etc; and something of the sort may well be implicit in Wittgenstein.) (REFERENCES). But that kind of semantics is generally suggested as a general *alternative* to referential semantics, thus assimilating the semantics of bona fide referential concepts ('tree' `Cicero`) to `logical' concepts like `and' (and, in our view, making the question how concepts can have extensions appear insoluble (Brandom REFERFENCE). Inferences from `some xs are F' to `all xs are F' should be avoided whenever possible.

longer are many) that, theories of reference, causal or otherwise must somehow make sense of reference to things outside the PC; and, pretty clearly, there are things outside the PC that we can refer but to which, at best, we aren't *directly* causally connected (Sherlock Holmes and Julius Caesar for example; and there are things outside the PC (properties and number, for example) that aren't causally connected to mental representations or to anything else. Taken together, however, these are an odd lot: The reason that Sherlock Holmes is outside our perceptual circle is very different from the reason that Julius Caesar is; which is again very different from the reasons that perfect triangles and the property RED are; which are again outside the PC reasons that some very small things are; which is again different from the reason that our far-flung grandchildren are (to say nothing of our merely prospective great grandchildren). Likewise, things outside our light-cones. We see no reason to doubt that (barring Sherlock Holmes) all these things are, or could be, real. But the fact that the kinds of putative counterexamples to causal theories of reference are so strikingly heterogeneous suggests to us that perhaps they can be dealt with piecemeal That's what this Chapter has it in mind to do.

We think that (contrary to claims that philosophers and others have made from time to time) there are no really conclusive arguments against the view that conceptual content boils down to reference; or against the view that the vehicles of conceptual content are, in the first instance, mental representations; or against the view that the reference of mental representations supervenes on their causal relations. If we thought otherwise, we would have written some other book. Maybe, however, convincing arguments against one or other of those views will turn up tomorrow. We hope not; but if one does, we'll worry about it then.