OXFORD COGNITIVE SCIENCE SERIES

CONCEPTS

OXFORD COGNITIVE SCIENCE SERIES

General Editors
MARTIN DAVIES, JAMES HIGGINBOTHAM, JOHN O'KEEFE,
CHRISTOPHER PEACOCKE, KIM PLUNKETT

Forthcoming in the series

Context and Content
Robert Stalnaker

Mindreading
Stephen Stich and Shaun Nichols

Face and Mind: The Science of Face Perception
Andy Young

CONCEPTS

Where Cognitive Science Went Wrong

JERRY A. FODOR

CLARENDON PRESS · OXFORD
1998

Oxford University Press, Great Clarendon Street, Oxford OX2 6DP
Oxford New York
Athens Auckland Bangkok Bogota Bombay
Buenos Aires Calcutta Cape Town Dar es Salaam
Delhi Florence Hong Kong Istanbul Karachi
Kuala Lumpur Madras Madrid Melbourne
Mexico City Nairobi Paris Singapore
Taipei Tokyo Toronto Warsaw
and associated companies in
Berlin Ibadan

Oxford is a trade mark of Oxford University Press

Published in the United States by
Oxford University Press Inc., New York

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Data available

ISBN 0–19–823637–9
ISBN 0–19-823636–0 (pbk.)

1 3 5 7 9 10 8 6 4 2

Typeset by Invisible Ink
Printed in Great Britain
on acid-free paper by
Biddles Ltd, Guildford and King's Lynn

for Janet, KP and Anthony; nuclear family

Chorus: Zuruck!
Tamino: . . . Zuruck?
Da seh ich noch ein Tur,
Vielleicht find ich den Eingang hier.

—*The Magic Flute*

CONTENTS

ABBREVIATIONS AND TYPOGRAPHICAL CONVENTIONS

THE following conventions are adopted throughout:

Concepts are construed as mental particulars. Names of concepts are written in capitals. Thus, 'RED' names the concept that expresses *redness* or *the property of being red*. Formulas in capitals are not, in general, structural descriptions of the concepts they denote. See Chapter 3, n. 1.

Names of English expressions appear in single quotes. Thus ''red'' is the name of the homophonic English word.

Names of semantic values of words and concepts are written in italics. Thus 'RED expresses the property of *being red*' and ''Red' expresses the property of *being red*' are both true.

The following abbreviations are used frequently (especially in Chapters 6 and 7).

RTM: The representational theory of the mind
IRS: Informational role semantics
MOP: Mode of presentation
MR: Mental representation
IA: Informational atomism
SA: The standard argument (for radical concept innateness)
SIA: Supplemented informational atomism (= IA plus a locking theory of concept possession)
d/D problem: The doorknob/DOORKNOB problem.

1

Philosophical Introduction: The Background Theory

> Needless to say, this rather baroque belief system gave rise to incredibly complicated explanations by the tribal elders . . .
>
> — Will Self

My topic is what concepts are. Since I'm interested in that question primarily as it arises in the context of 'representational' theories of mind (RTMs), a natural way to get started would be for me to tell you about RTMs and about how they raise the question what concepts are. I could then set out my answer, and you could tell me, by return, what you think is wrong with it. The ensuing discussion would be abstract and theory laden, no doubt; but, with any luck, philosophically innocent.

That is, in fact, pretty much the course that I propose to follow. But, for better or for worse, in the present climate of philosophical opinion it's perhaps not possible just to plunge in and do so. RTMs have all sorts of problems, both of substance and of form. Many of you may suppose the whole project of trying to construct one is hopelessly wrong-headed; if it is, then who cares what RTMs say about concepts? So I guess I owe you some sort of general argument that the project isn't hopelessly wrong-headed.

But I seem to have grown old writing books defending RTMs; it occurs to me that if I were to stop writing books defending RTMs, perhaps I would stop growing old. So I think I'll tell you a joke instead. It's an *old* joke, as befits my telling it.

Old joke: Once upon a time a disciple went to his guru and said: 'Guru, what is life?' To which the Guru replies, after much thinking, 'My Son, life is like a fountain.' The disciple is outraged. 'Is that the best that you can do? Is that what you call wisdom?' 'All right,' says the guru; 'don't get excited. So maybe it's not like a fountain.'

That's the end of the joke, but it's not the end of the story. The guru noticed that taking this line was losing him clients, and gurus have to eat.

So the next time a disciple asked him: 'Guru, what is life?' his answer was: 'My Son, I cannot tell you.' 'Why can't you?' the disciple wanted to know. 'Because,' the guru said, 'the question "What is *having* a life?" is logically prior.' 'Gee,' said the disciple, 'that's pretty interesting'; and he signed on for the whole term.

I'm not going to launch a full-dress defence of RTM; but I do want to start with a little methodological stuff about whether having a concept is logically prior to being a concept, and what difference, if any, that makes to theorizing about mental representation.

It's a general truth that if you know *what an X is*, then you also know *what it is to have an X*. And ditto the other way around. This applies to concepts in particular: the question what they are and the question what it is to have them are logically linked; if you commit yourself on one, you are *thereby* committed, willy nilly, on the other. Suppose, for example, that your theory is that concepts are pumpkins. Very well then, it will have to be a part of your theory that having a concept is having a pumpkin. And, conversely: if your theory is that having a concept is having a pumpkin, then it will have to be a part of your theory that pumpkins are what concepts are. I suppose this all to be truistic.

Now, until quite recently (until this century, anyhow) practically everybody took it practically for granted that the explanation of concept *possession* should be parasitic on the explanation of concept *individuation*. First you say what it is for something *to be* the concept *X*—you give the concept's 'identity conditions'—and then *having* the concept *X* is just *having whatever the concept X turns out to be*. But the philosophical fashions have changed. Almost without exception, current theories about concepts reverse the classical direction of analysis. Their substance lies in what they say about the conditions for *having* concept *X*, and it's the story about *being* concept *X* that they treat as derivative. Concept *X* is just: *whatever it is that having the concept X consists in having*. Moreover, the new consensus is that you really must take things in that order; the sanctions incurred if you go the other way round are said to be terrific. (Similarly, *mutatis mutandis* for *being the meaning of a word* vs. *knowing the meaning of a word*. Here and elsewhere, I propose to move back and forth pretty freely between concepts and word meanings; however it may turn out in the long run, for purposes of the present investigation word meanings just are concepts.)

You might reasonably wonder how there possibly could be this stark methodological asymmetry. We've just been seeing that the link between 'is an *X*' and 'has an *X*' is conceptual; fix one and you thereby fix the other. How, then, could there be an issue of principle about which you should start with? The answer is that when philosophers take a strong line

on a methodological issue there's almost sure to be a metaphysical subtext. The present case is not an exception.

On the one side, people who start in the traditional way by asking 'What are concepts?' generally hold to a traditional metaphysics according to which a concept is a kind of mental particular. I hope that this idea will get clearer and clearer as we go along. Suffice it, for now, that the thesis that concepts are mental particulars is intended to imply that *having* a concept is constituted by having a mental particular, and hence to exclude the thesis that having a concept is, in any interesting sense, constituted by having mental traits or capacities.[1] You may say, if you like, that having concept *X* is having the ability to think about *X*s (or better, that having the concept *X* is being able to think about *X*s 'as such'). But, though that's true enough, it doesn't alter the metaphysical situation as traditionally conceived. For thinking about *X*s consists in having thoughts about *X*s, and thoughts are supposed to be mental particulars too.

On the other side, people who start with 'What is concept *possession*?' generally have some sort of Pragmatism in mind as the answer. Having a concept is a matter of what you are able to *do*, it's some kind of epistemic 'know how'. Maybe having the concept *X* comes to something like *being reliably able to recognize Xs and/or being reliably able to draw sound inferences about Xness*.[2] In any case, an account that renders having concepts as having capacities is intended to preclude an account that renders concepts as species of mental particulars: capacities aren't kinds of *things*; a fortiori, they aren't kinds of *mental* things.

So, to repeat, the methodological doctrine that concept possession is logically prior to concept individuation frequently manifests a preference

[1] I want explicitly to note what I've come to think of as a cardinal source of confusion in this area. If concept tokens are mental particulars, then having a concept is being in a relation to a mental particular. This truism about the *possession conditions* for concepts continues to hold whatever doctrine you may embrace about how concepts *tokens* get assigned to concept *types*. Suppose Jones's TIGER-concept is a mental token that plays a certain (e.g. causal) role in his mental life. That is quite compatible with supposing that what makes it a token of the type TIGER-concept (rather than a token of the type MOUSE-concept; or not a token of a concept type at all) is something dispositional; viz. the dispositional properties *of the token* (as opposed, say, to its weight or colour or electric charge).

The discussion currently running in the text concerns the relation between theories about the ontological status of concepts and theories about what it is to have a concept. Later, and at length, we'll consider the quite different question how concept tokens are typed.

[2] Earlier, less sophisticated versions of the view that the metaphysics of concepts is parasitic on the metaphysics of concept possession were generally not merely pragmatist but also behaviourist: they contemplated reducing concept possession to a capacity for responding selectively. The cognitive revolutions in psychology and the philosophy of mind gagged on behaviourism, but never doubted that concepts are some sort of capacities or other. A classic case of getting off lightly by pleading to the lesser charge.

for an ontology of mental dispositions rather than an ontology of mental particulars. This sort of situation will be familiar to old hands; proposing dispositional analyses in aid of ontological reductions is the method of critical philosophy that Empiricism taught us. If you are down on cats, reduce them to permanent possibilities of sensation. If you are down on electrons and protons, reduce them to permanent possibilities of experimental outcomes. And so on. There is, however, a salient difference between reductionism about cats and reductionism about concepts: perhaps some people think that they *ought* to think that cats are constructs out of possible experiences, but surely nobody actually does think so; one tolerates a little *mauvaise foi* in metaphysics. Apparently, however, lots of people do think that concepts are constructs out of mental (specifically epistemic) capacities. In consequence, and this is a consideration that I take quite seriously, whereas nobody builds biological theories on the assumption that cats are sensations, much of our current cognitive science, and practically all of our current philosophy of mind, is built on the assumption that concepts are capacities. If that assumption is wrong, very radical revisions are going to be called for. So, at least, I'll argue.

To sum up so far: it's entirely plausible that a theory of what concepts are must likewise answer the question 'What is it to have a concept?' and, *mutatis mutandis*, that a theory of meaning must answer the question 'What is it to understand a language?' We've been seeing, however, that this untendentious methodological demand often comports with a substantive metaphysical agenda: viz. the reduction of concepts and meanings to epistemic capacities.

Thus Michael Dummett (1993*a*: 4), for one illustrious example, says that "any theory of meaning which was not, or did not immediately yield, a theory of understanding, would not satisfy the purpose for which, philosophically, we require a theory of meaning". There is, as previously remarked, a reading on which this is true but harmless since *whatever* ontological construal of *the meaning of an expression* we settle on will automatically provide a corresponding construal of *understanding the expression* as *grasping* its meaning. It is not, however, this truism that Dummett is commending. Rather, he has it in mind that an acceptable semantics must explicate linguistic content just by reference to the "practical" capacities that users of a language have qua users of that language. (Correspondingly, a theory that explicates the notion of conceptual content would do so just by reference to the practical capacities that having the concept bestows.) Moreover, if I read him right, Dummett intends to impose this condition in a very strong form: the capacities upon which linguistic meaning supervenes must be such as can be severally and determinately manifested in behaviour. "An axiom earns its place in the

theory [of meaning] . . . only to the extent that it is required for the derivation of theorems the ascription of an implicit knowledge of which to a speaker *is explained in terms of specific abilities which manifest that knowledge*" (1993*b*: 38; my emphasis).

I don't know for sure why Dummett believes that, but I darkly suspect that he's the victim of atavistic sceptical anxieties about communication. Passages like the following recur in his writings:

> What . . . constitutes a subject's understanding the sentences of a language . . .? [I]s it his having internalized a certain theory of meaning for that language? . . . then indeed his behaviour when he takes part in linguistic interchange can at best be strong but fallible evidence for the internalized theory. In that case, however, the hearer's presumption that he has understood the speaker can never be definitively refuted or confirmed. (1993*c*: 180; notice how much work the word 'definitively' is doing here.)

So, apparently, the idea is that theories about linguistic content should reduce to theories about language use; and theories about language use should reduce to theories about the speaker's linguistic capacities; and theories about the speaker's linguistic capacities are constrained by the requirement that any capacity that is constitutive of the knowledge of a language is one that the speaker's use of the language can overtly and specifically manifest. All this must be in aid of devising a bullet-proof anti-scepticism about communication, since it would seem that for purposes *other* than refuting sceptics, all the theory of communication requires is that a speaker's utterances reliably cause certain 'inner processes' in the hearer; specifically, mental processes which eventuate in the hearer having the thought that the speaker intended him to have.

If, however, scepticism really is the skeleton in Dummett's closet, the worry seems to me to be doubly misplaced: first because the questions with which theories of meaning are primarily concerned are metaphysical rather than epistemic. This is as it should be; understanding what a thing is, is invariably prior to understanding how we know what it is. And, secondly, because there is no obvious reason why behaviourally grounded inferences to attributions of concepts, meanings, mental processes, communicative intentions, and the like should be freer from normal inductive risk than, as it might be, perceptually grounded attributions of tails to cats. The best we get in either case is "strong but fallible evidence". Contingent truths are like that as, indeed, Hume taught us some while back. This is, no doubt, the very attitude that Dummett means to reject as inadequate to the purposes for which we "philosophically" require a theory of meaning. So much the worse, perhaps, for the likelihood that philosophers will get from a theory of meaning what Dummett says that

they require. I, for one, would not expect a good account of what concepts are to refute scepticism about other minds any more than I'd expect a good account of what cats are to refute scepticism about other bodies. In both cases, I am quite prepared to settle for theories that are merely *true*.

Methodological inhibitions flung to the wind, then, here is how I propose to organize our trip. Very roughly, concepts are constituents of mental states. Thus, for example, believing that *cats are animals* is a paradigmatic mental state, and the concept ANIMAL is a constituent of the belief that *cats are animals* (and of the belief that *animals sometimes bite*; etc. I'm leaving it open whether the concept ANIMAL is likewise a constituent of the belief that *some cats bite*; we'll raise that question presently). So the natural home of a theory of concepts is as part of a theory of mental states. I shall suppose throughout this book that RTM is the right theory of (cognitive) mental states. So, I'm going to start with an exposition of RTM: which is to say, with an exposition of a theory about what mental states and processes are. It will turn out that mental states and processes are typically species of relations to mental representations, of which latter concepts are typically the parts.

To follow this course is, in effect, to assume that it's OK for theorizing about the nature of concepts to precede theorizing about concept possession. As we've been seeing, barring a metaphysical subtext, that assumption should be harmless; individuation theories and possession theories are trivially intertranslatable. Once we've got RTM in place, however, I'm going to argue for a very strong version of psychological atomism; one according to which what concepts you have is conceptually and metaphysically independent of what epistemic capacities you have. If this is so, then patently concepts couldn't *be* epistemic capacities.

I hope not to beg any questions by proceeding in this way; or at least not to get caught begging any. But I do agree that if there is a knock-down, a priori argument that concepts are logical constructs out of capacities, then my view about their ontology can't be right and I shall have to give up my kind of cognitive science. Oh, well. If there's a knock-down, a priori argument that cats are logical constructs out of sensations, then my views about *their* ontology can't be right either, and I shall have to give up my kind of biology. Neither possibility actually worries me a lot.

So, then, to begin at last:

RTM

RTM is really a loose confederation of theses; it lacks, to put it mildly, a canonical formulation. For present purposes, let it be the conjunction of the following:

> First Thesis: *Psychological explanation is typically nomic and is intentional through and through*. The laws that psychological explanations invoke typically express causal relations among *mental states that are specified under intentional description*; viz. among mental states that are picked out by reference to their contents. Laws about causal relations among beliefs, desires, and actions are the paradigms.

I'm aware there are those (mostly in Southern California, of course) who think that intentional explanation is all at best pro tem, and that theories of mind will (or anyhow should) eventually be couched in the putatively purely extensional idiom of neuroscience. But there isn't any reason in the world to take that idea seriously and, in what follows, I don't.

There are also those who, though they are enthusiasts for intentional explanation, deny the metaphysical possibility of laws about intentional states. I don't propose to take that seriously in what follows either. For one thing, I find the arguments that are said to show that there can't be intentional laws very hard to follow. For another thing, if there are no intentional laws, then you can't make science out of intentional explanations; in which case, I don't understand how intentional explanation *could* be better than merely pro tem. Over the years, a number of philosophers have kindly undertaken to explain to me what non-nomic intentional explanations would be good for. Apparently it has to do with the intentional realm (or perhaps it's the rational realm) being autonomous. But I'm afraid I find all that realm talk very hard to follow too. What *is* the matter with me, I wonder?[3]

> Second Thesis: *'Mental representations' are the primitive bearers of intentional content*.

Both ontologically and in order of explanation, the intentionality of the propositional attitudes is prior to the intentionality of natural languages; and, both ontologically and in order of explanation, the intentionality of mental representations is prior to the intentionality of propositional attitudes.

Just for purposes of building intuitions, think of mental representations on the model of what Empiricist philosophers sometimes called 'Ideas'. That is, think of them as mental particulars endowed with causal powers and susceptible of semantic evaluation. So, there's the Idea DOG. It's satisfied by all and only dogs, and it has associative-cum-causal relations to, for example, the Idea CAT. So DOG has conditions of semantic evaluation and it has causal powers, as Ideas are required to do.

[3] The trouble may well have to do with my being a Hairy Realist. See Fodor 1995*b*.

Since a lot of what I want to say about mental representations includes what Empiricists did say about Ideas, it might be practical and pious to speak of Ideas rather than mental representations throughout. But I don't propose to do so. The Idea idea is historically intertwined with the idea that Ideas are images, and I don't want to take on that commitment. To a first approximation, then, the idea that there are mental representations is the idea that there are Ideas *minus* the idea that Ideas are images.

RTM claims that mental representations are related to propositional attitudes as follows: for each event that consists of a creature's having a propositional attitude with the content *P* (each such event as Jones's believing at time *t* that *P*) there is a corresponding event that consists of the creature's being related, in a characteristic way, to a token mental representation that has the content *P*. Please note the meretricious scrupulousness with which metaphysical neutrality is maintained. I did *not* say (albeit I'm much inclined to believe) that having a propositional attitude *consists in* being related (in one or other of the aforementioned 'characteristic ways') to a mental representation.

I'm also neutral on what the 'characteristic ways' of being related to mental representations are. I'll adopt a useful dodge that Stephen Schiffer invented: I assume that everyone who has beliefs has a belief box in his head. Then:

> *For each episode of believing that P, there is a corresponding episode of having, 'in one's belief box', a mental representation which means that P.*

Likewise, *mutatis mutandis*, for the other attitudes. Like Schiffer, I don't really suppose that belief boxes are literally boxes, or even that they literally have insides. I assume that the essential conditions for belief-boxhood are functional. Notice, in passing, that this is *not* tantamount to assuming that "believe" has a 'functional definition'. I doubt that "believe" has *any* definition. That most—indeed, overwhelmingly most—words don't have will be a main theme in the third chapter. But denying, as a point of semantics, that "believe" has a functional definition is compatible with asserting, as a point of metaphysics, that belief has a functional essence. Which I think that it probably does. Ditto, *mutatis mutandis*, "capitalism", "carburettor", and the like. (Compare Devitt 1996; Carruthers 1996, both of whom run arguments that depend on not observing this distinction.)

RTM says that there is no believing-that-*P* episode without a corresponding tokening-of-a-mental-representation episode, and it contemplates no locus of original intentionality except the contents of mental representations. In consequence, so far as RTMs are concerned, to

explain what it is for a mental representation to mean what it does *is* to explain what it is for a propositional attitude to have the content that it does. I suppose that RTM leaves open the metaphysical possibility that there could be mental states whose content does not, in this sense, derive from the meaning of corresponding mental representations. But it takes such cases not to be *nomologically* possible, and it provides no hint of an alternative source of propositional objects for the attitudes.

Finally, English inherits its semantics from the contents of the beliefs, desires, intentions, and so forth that it's used to express, as per Grice and his followers. Or, if you prefer (as I think, on balance, I do), English *has no semantics*. Learning English isn't learning a theory about what its sentences mean, it's learning how to associate its sentences with the corresponding thoughts. To know English is to know, for example, that the form of words 'there are cats' is standardly used to express the thought that there are cats; and that the form of words 'it's raining' is standardly used to express the thought that it's raining; and that the form of words 'it's not raining' is standardly used to express the thought that it's not raining; and so on for in(de)finitely many other such cases.

Since, according to RTM, the content of linguistic expressions depends on the content of propositional attitudes, and the content of propositional attitudes depends on the content of mental representations, and since the intended sense of 'depends on' is asymmetric, RTM tolerates the metaphysical possibility of thought without language; for that matter, it tolerates the metaphysical possibility of mental representation without thought. I expect that many of you won't like that. I'm aware that there is rumoured to be an argument, vaguely Viennese in provenance, that proves that 'original', underived intentionality must inhere, *not* in mental representations *nor* in thoughts, but precisely in the formulas of public languages. I would be very pleased if such an argument actually turned up, since then pretty nearly everything I believe about language and mind would have been refuted, and I could stop worrying about RTM, and about what concepts are, and take off and go sailing, a pastime that I vastly prefer. Unfortunately, however, either nobody can remember how the argument goes, or it's a secret that they're unprepared to share with me. So I'll forge on.

Third Thesis: *Thinking is computation*.

A theory of mind needs a story about mental *processes*, not just a story about mental states. Here, as elsewhere, RTM is closer in spirit to Hume than it is to Wittgenstein or Ryle. Hume taught that *mental states are relations to mental representations*, and so too does RTM (the main difference being, as we've seen, that RTM admits, indeed demands, mental

representations that aren't images). Hume also taught that *mental processes* (including, paradigmatically, thinking) *are causal relations among mental representations*.[4] So too does RTM. In contrast to Hume, and to RTM, the logical behaviourism of Wittgenstein and Ryle had, as far as I can tell, no theory of thinking at all (except, maybe, the silly theory that thinking is talking to oneself). I do find that shocking. How *could* they have expected to get it right about belief and the like without getting it right about belief fixation and the like?

Alan Turing's idea that thinking is a kind of computation is now, I suppose, part of everybody's intellectual equipment; not that everybody likes it, of course, but at least everybody's heard of it. That being so, I shall pretty much take it as read for the purposes at hand. In a nutshell: token mental representations are symbols. Tokens of symbols are physical objects with semantic properties. To a first approximation, computations are those causal relations among symbols which reliably respect semantic properties of the relata. Association, for example, is a bona fide computational relation within the meaning of the act. Though whether Ideas get associated is supposed to depend on their frequency, contiguity, etc., and not on what they're Ideas *of*, association is none the less supposed reliably to preserve semantic domains: *Jack*-thoughts cause *Jill*-thoughts, *salt*-thoughts cause *pepper*-thoughts, *red*-thoughts cause *green*-thoughts, and so forth.[5] So, Hume's theory of mental processes is itself a species of RTM, an upshot that pleases me.

Notoriously, however, it's an inadequate species. The essential problem in this area is to explain how thinking manages reliably to preserve *truth*; and Associationism, as Kant rightly pointed out to Hume, hasn't the resources to do so. The problem isn't that association is a causal relation, or that it's a causal relation among symbols, or even that it's a causal relation among mental symbols; it's just that their satisfaction conditions aren't among the semantic properties that associates generally share. To the contrary, being Jack precludes being Jill, being salt precludes being pepper, being red precludes being green, and so forth. By contrast, Turing's account of thought-as-computation showed us how to specify causal relations among mental symbols that are reliably truth-preserving. It thereby saved RTM from drowning when the Associationists went under.

I propose to swallow the Turing story whole and proceed. First, however, there's an addendum I need and an aside I can't resist.

[4] And/or among states of entertaining them. I'll worry about this sort of ontological nicety only where it seems to matter.

[5] Why relations that depend on merely mechanical properties like frequency and contiguity *should* preserve intentional properties like semantic domain was what Associationists never could explain. That was one of the rocks they foundered on.

Addendum: if computation is just causation that preserves semantic values, then the thesis that thought is computation requires of mental representations only that they have semantic values and causal powers that preserve them. I now add a further constraint: many mental representations have *constituent* (*part/whole*) *structure*, and many mental processes are sensitive to the constituent structure of the mental representations they apply to. So, for example, the mental representation that typically gets tokened when you think . . . *brown cow* . . . has, among its constituent parts, the mental representation that typically gets tokened when you think . . . *brown* . . .; and the computations that RTM says get performed in processes like inferring from . . . *brown cow* . . . to . . . *brown* . . . exploit such part/whole relations. Notice that this *is* an addendum (though it's one that Turing's account of computation was designed to satisfy). It's untendentious that RTM tolerates the possibility of conceptual content *without* constituent structure since everybody who thinks that there are mental representations at all thinks that at least some of them are primitive.[6]

The aside I can't resist is this: following Turing, I've introduced the notion of computation by reference to such semantic notions as content and representation; a computation is some kind of content-respecting causal relation among symbols. However, this order of explication is OK *only if the notion of a symbol doesn't itself presuppose the notion of a computation*. In particular, it's OK only if you don't need the notion of a computation to explain what it is for something to have semantic properties. We'll see, almost immediately, that the account of the *semantics* of mental representations that my version of RTM endorses, unlike the account of *thinking* that it endorses, is indeed non-computational.

Suppose, however, it's your metaphysical view that the semantic properties of a mental representation depend, wholly or in part, upon the computational relations that it enters into; hence that the notion of a computation is *prior* to the notion of a symbol. You will then need some *other* way of saying what it is for a causal relation among mental representations to *be* a computation; *some way that does not presuppose such notions as symbol and content*.[7] It may be possible to find such a notion of computation, but I don't know where. (Certainly not in Turing,

[6] Connectionists are committed, willy-nilly, to *all* mental representations being primitive; hence their well-known problems with systematicity, productivity, and the like. More on this in Chapter 5.

[7] Not, of course, that there is anything wrong with just allowing 'symbol' and 'computation' to be interdefined. But that option is not available to anyone who takes the theory that thought is computation to be part of a *naturalistic* psychology; viz. part of a programme of metaphysical reduction. As Turing certainly did; and as do I.

who simply takes it for granted that the expressions that computing machines crunch are *symbols*; e.g. that they denote numbers, functions, and the like.) The attempts I've seen invariably end up suggesting (or proclaiming) that *every* causal process is a kind of computation, thereby trivializing Turing's nice idea that *thought* is.

So much for mental processes.

Fourth Thesis: *Meaning is information (more or less).*

There actually are, in the land I come from, philosophers who would agree with the gist of RTM as I've set it forth so far. Thesis Four, however, is viewed as divisive even in that company. I'm going to assume that what bestows content on mental representations is something about their causal-cum-nomological relations to the things that fall under them: for example, what bestows upon a mental representation the content *dog* is something about its tokenings being caused by dogs.

I don't want to pursue, beyond this zero-order approximation, the question just which causal-cum-nomological relations are content-making. Those of you who have followed the literature on the metaphysics of meaning that Fred Dretske's book *Knowledge and the Flow of Information* (1981) inspired will be aware that that question is (ahem!) mootish. But I do want to emphasize one aspect of the identification of meaning with information that is pretty widely agreed on and that impacts directly on any proposal to amalgamate an informational semantics with RTM: if meaning is information, then coreferential representations must be synonyms.

Just how this works depends, of course, on what sort of causal-cum-nomological covariation content is and what sort of things you think concepts represent (properties, actual objects, possible objects, or whatever). Suppose, for example, that you run the kind of informational semantics that says:

> *A representation* R *expresses the property* P *in virtue of its being a law that things that are* P *cause tokenings of* R *(in, say, some still-to-be-specified circumstances* C*)*.

And suppose, for the sake of the argument, that *being water* and *being* H_2O are (not merely coextensive but) the same property. It then follows that if it's a law that WATER tokens covary with water (in *C*) it's also a law that WATER tokens covary with H_2O (in *C*). So a theory that says that WATER means *water* in virtue of there being the first law is also required to say that WATER means H_2O in virtue of there being the second. Parallel reasoning shows that H_2O means *water*, hence that WATER and H_2O mean the same.

You may wonder why I want to burden my up to now relatively uncontroversial version of RTM by adding a theory of meaning that has this uninviting consequence; and how I could reasonably suppose that you'll be prepared to share the burden by granting me the addition. Both questions are fair.

As to the first, suppose that coextension is *not* sufficient for synonymy after all. Then there must be something else to having a concept with a certain content than having a mental representation with the kind of world-to-symbol causal connections that informational semantics talks about. The question arises: *what is this extra ingredient?* There is, as everybody knows, a standard answer; viz. that *what concepts one has is determined*, at least in part, *by what inferences one is prepared to draw* or to accept. If it is possible to have the concept WATER and not have the concept H_2O, that's because it's constitutive of having the latter, but not constitutive of having the former, that you accept such inferences as *contains* $H_2O \rightarrow$ *contains* H. It is, in short, received wisdom that content may be constituted in part by informational relations, but that unless coreference is sufficient for synonymy, it must also be constituted by inferential relations. I'll call any theory that says this sort of thing an Inferential Role Semantics (IRS).

I don't want content to be constituted, even in part, by inferential relations. For one thing, as we just saw, I like Turing's story that inference (qua mental process) reduces to computation; i.e. to *operations on symbols*. For fear of circularity, I can't *both* tell a computational story about what inference is *and* tell an inferential story about what content is. Prima facie, at least, if I buy into Inferential Role Semantics, I undermine my theory of thinking.

For a second thing, I am inclined to believe that an inferential role semantics has holistic implications that are both unavoidable and intolerable. A main reason I love RTM so much is that the computational story about mental *processes* fits so nicely with the story that psychological *explanation* is subsumption under intentional laws; viz. under laws that apply to a mental state in virtue of its content. Since computation is presumed to respect content, RTM can maybe provide the mechanism whereby satisfying the antecedent of an intentional law necessitates the satisfaction of its consequent (see Fodor 1994: ch. 1). But I think it's pretty clear that psychological explanation can't be subsumption under intentional laws if the metaphysics of intentionality is holistic. (See Fodor and Lepore 1992.)

For a third thing, as previously noted, the main point of this book will be to argue for an *atomistic* theory of concepts. I'm going to claim, to put it very roughly, that satisfying the metaphysically necessary conditions for

having one concept *never* requires satisfying the metaphysically necessary conditions for having any other concept. (Well, *hardly* ever. See below.) Now, the status of conceptual atomism depends, rather directly, on whether coreference implies synonymy. For, if it doesn't, and if it is inferential role that makes the difference between content and reference, then every concept must *have* an inferential role. But it's also common ground that you need more than one concept to draw an inference, so if IRS is true, conceptual atomism isn't. No doubt this line of thought could use a little polishing, but it's surely basically sound.

So, then, if I'm going to push for an atomistic theory of concepts, I *must not* hold that one's inferential dispositions determine, wholly or in part, the content of one's concepts. Pure informational semantics allows me not to hold that one's inferential dispositions determine the content of one's concepts because it says that content is constituted, exhaustively, by symbol–world relations.

It's worth keeping clear on how the relation between concept possession and concept individuation plays out on an informational view: the content of, for example, BACHELOR is constituted by certain (actual and/or counterfactual) causal-cum-nomic relations between BACHELOR-tokenings and tokenings of instantiated *bachelorhood*. Presumably *bachelorhood* is itself individuated, *inter alia*, by the necessity of its relation to *being unmarried*. So, 'bachelors are unmarried' is conceptually necessary in the sense that it's guaranteed by the content of BACHELOR together with the metaphysics of the relevant property relations. It follows, trivially, that *having* BACHELOR is having a concept which can apply only to unmarried things; this is the truism that the interdefinability of concept individuation and concept possession guarantees. But *nothing at all* about the epistemic condition of BACHELOR owners (e.g. about their inferential or perceptual dispositions or capacities) follows from the necessity of 'bachelors are unmarried'; *it doesn't even follow that you can't own BACHELOR unless you own UNMARRIED*. Informational semantics permits atomism about concept possession even if (even though) there are conceptually necessary truths.[8] This is a sort of point that will recur repeatedly as we go along.

So much for why I want an informational semantics as part of my RTM. Since it is, of course, moot whether I can have one, the best I can hope for is that this book will convince you that conceptual atomism is OK unless there is a decisive, independent argument against the reduction of meaning to information. I'm quite prepared to settle for this since I'm

[8] What it doesn't do is guarantee the connection between what's conceptually necessary and what's a priori. But perhaps that's a virtue.

pretty sure that there's no such argument. In fact, I think the dialectic is going to have to go the other way around: what settles the metaphysical issue between informational theories of meaning and inferential role theories of meaning is that the former, but not the latter, are compatible with an atomistic account of concepts. And, as I'll argue at length, there are persuasive independent grounds for thinking that atomism about concepts must be true.

In fact, I'm going to be more concessive still. Given my view that content is information, I can't, as we've just seen, afford to agree that the content of the concept H_2O is different from the content of the concept WATER. *But I am entirely prepared to agree that they are different concepts*. In effect, I'm assuming that coreferential representations are *ipso facto* synonyms and conceding that, since they are, *content* individuation can't be all that there is to *concept* individuation.

It may help make clear how I'm proposing to draw the boundaries to contrast the present view with what I take to be a typical Fregean position; one according to which concepts are distinguished along two (possibly orthogonal) parameters; viz. reference and *Mode of Presentation*. (So, for example, the concept WATER is distinct from the concept DOG along *both* parameters, but it's distinct from the concept H_2O only in respect of the second.) I've diverged from this sort of scheme only in that some Fregeans (e.g. Frege) identify modes of presentation with *senses*. By contrast, I've left it open what modes of presentation are, so long as they are what distinguish distinct but coreferential concepts. So far, then, I'm less extensively committed than a Fregean, but I don't think that I'm committed to anything that a Fregean is required to deny.

Alas, ecumenicism has to stop somewhere. The fifth (and final thesis) of my version of RTM does depart from the standard Frege architecture.

> Fifth Thesis: *Whatever distinguishes coextensive concepts is* ipso facto *'in the head'*. This means, something like that it's available to be a proximal cause (/effect) of mental processes.[9]

As I understand it, the Fregean story makes the following three claims about modes of presentation:

5.1 MOPs are senses; for an expression to mean what it does is for the expression to have the MOP that it does.

[9] I take it that one of the things that distinguishes Fregeans *sans phrase* from *neo*-Fregeans (like e.g. Peacocke 1992) is that the latter are *not* committed to Fege's antimentalism and are therefore free to agree with Thesis Five if they're so inclined. Accordingly, for the *neo*- sort of Fregean, the sermon that follows will seem to be preached to the converted.

5.2 Since MOPs can distinguish concepts, they explain how it is possible to entertain one, but not the other, of two coreferential concepts; e.g. how it is possible have the concept WATER but not the concept H_2O, hence how it is possible to have (de dicto) beliefs about water but no (de dicto) beliefs about H_2O.

5.3 MOPs are abstract objects; hence they are non-mental.

In effect, I've signed on for 5.2; it's the claim about MOPs that everybody must accept who has any sympathy at all for the Frege programme. But I think there are good reasons to believe that 5.2 excludes both 5.1 and 5.3. In which case, I take it that 5.1 and 5.3 will have to go.

—*What's wrong with 5.1*: 5.1 makes trouble for 5.2: it's unclear that you can hold onto 5.2 if you insist, as Frege does, that MOPs be identified with senses. One thing (maybe the only one) that we know for sure about senses is that synonyms share them. So if MOPs are senses and distinct but coextensive concepts are distinguished (solely) by their MOPs, then synonymous concepts must be identical, and it must not be possible to think either without thinking the other. (This is the so-called 'substitution test' for distinguishing modes of presentation.) But (here I follow Mates 1962), it is possible for Fred to wonder *whether John understands that bachelors are unmarried men* even though Fred does not wonder *whether John understands that unmarried men are unmarried men*. The moral seems to be that if 5.2 is right, so that MOPs *just are* whatever it is that the substitution test tests for, then it's unlikely that MOPs are senses.

Here's a similar argument to much the same conclusion. Suppose I tell you that Jackson was a painter and that Pollock was a painter, and I tell you nothing else about Jackson or Pollock. Suppose, also, that you believe what I tell you. It looks like that fixes the senses of the names 'Jackson' and 'Pollock' if anything could; and it looks like it fixes them as both having the *same* sense: viz. *a painter*. (*Mutatis mutandis*, it looks as though I have fixed the same inferential role for both.) Yet, in the circumstances imagined, it's perfectly OK—perfectly conceptually coherent—for you to wonder whether Jackson and Pollock were the *same* painter. (Contrast the peculiarity of your wondering, in such a case, whether Jackson was Jackson or whether Pollock was Pollock.) So, then, by Frege's own test, JACKSON and POLLOCK count as different MOPs. But if concepts with the same sense can be different MOPs then, patently, MOPs can't be senses. This isn't particularly about names, by the way. If I tell you that a flang is a sort of machine part and a glanf is a sort of machine part, it's perfectly OK for you to wonder whether a glanf is a flang.[10]

[10] You can't, of course, do this trick with definite descriptions since they presuppose

Oh well, maybe my telling you that Jackson was a painter and Pollock was a painter didn't fix the same senses for both names after all. I won't pursue that because, when it comes to senses, who can prove what fixes what? But it hardly matters since, on reflection, what's going on doesn't seem to have to do with *meaning*. Rather, the governing principle is a piece of logical syntax: If '*a*' and '*b*' are different names, then the inference from '*Fa*' to '*Fb*' is never conceptually necessary.[11] (It's even OK to wonder whether Jackson is Jackson, if the two 'Jacksons' are supposed to be tokens of different but homonymous name types.) It looks like the moral of this story about Jackson and Pollock is the same as the moral of Mates's story about bachelors and unmarried men. *Frege's substitution test doesn't identify senses*. Correspondingly, if it is stipulated that MOPs are whatever substitution *salve veritate* turns on, then MOPs have to be sliced a good bit *thinner* than senses. Individuating MOPs is more like individuating forms of words than it is like individuating meanings.

I take these sorts of considerations *very* seriously. They will return full strength at the end of Chapter 2.

—*What's wrong with 5.3*: This takes a little longer to say, but here is the short form. Your having *n* MOPs for water explains why you have *n* ways of thinking about water *only on the assumption that there is exactly one way to grasp each MOP*.[12] The question thus arises what, if anything, is supposed to legitimize this assumption. As far as I can tell, unless you're prepared to give up 5.3, the only answer a Fregean theory allows you is: sheer stipulation.

Terminological digression (I'm sorry to have to ask you to split these hairs, but this is a part of the wood where it is *very* easy to get lost): I use 'entertaining' and 'grasping' a MOP (/concept) interchangeably. Entertaining/grasping a MOP doesn't, of course, mean *thinking about* the MOP;

uniqueness of reference. If you mean by "Jackson" *the horse that bit John*, and you mean by "Pollock" *the horse that bit John*, you can't coherently wonder whether Jackson is the same horse as Pollock.

By the way, I have the damnedest sense of *déjà vu* about the argument in the text; I simply can't remember whether I read it somewhere or made it up. If it was you I snitched it from, Dear Reader, please do let me know.

[11] More precisely: it's never conceptually necessary unless either the inference from *Fa* to $a = b$ or the inference from *Fb* to $a = b$ is itself conceptually necessary. (For example, let *Fa* be: '*a* has the property of being identical to *b*'.)

[12] Or, if there is more than one way to grasp a MOP, then all of the different ways of doing so must correspond to the *same* way of thinking its referent. I won't pursue this option in the text; suffice it that doing so wouldn't help with the problem that I'm raising. Suppose that there is more than one way to grasp a MOP; and suppose that a certain MOP is a mode of presentation of Moe. Then if, as Frege requires, there is a MOP corresponding to each way of thinking a referent, all the ways of grasping the Moe-MOP must be the *same* way of thinking of Moe. I claim that, precisely because 5.3 is in force, Frege's theory has no way to ensure that this is so.

there are as many ways of thinking about a MOP as there are of thinking about a rock or a number. That is, innumerably many; one for each mode of presentation of the MOP. Rather, MOPs are supposed to be the *vehicles* of thought, and entertaining a MOP means using it to present to thought whatever the MOP is a mode of presentation of; it's thinking *with* the MOP, not thinking *about* it. End digression. My point is that if there is more than one way to grasp a MOP, then 'grasping a water-MOP is a way of thinking about water' and 'Smith has only one water-MOP' does *not* entail that Smith has only one way of thinking about water.

So, then, what ensures that there is only one way to grasp a MOP? Since Frege thinks that MOPs are senses and that sense determines reference (concepts with the same sense must be coextensive) he holds, in effect, that MOP identity and concept identity come to the same thing. So my question can be put just in terms of the latter: that one has as many ways of thinking of a referent as one has concepts of the referent depends on there being just one way to entertain each concept. What, beside stipulation, guarantees this?

Perhaps the following analogy (actually quite close, I think) will help to make the situation clear. There are lots of cases where things other, and less problematic, than Fregean senses might reasonably be described as 'modes of presentation'; viz. as being used to present the object of a thought to the thought that it's the object of. Consider, for example, using a diagram of a triangle in geometrical reasoning about triangles. It seems natural, harmless, maybe even illuminating, to say that one sometimes reasons about triangles *via* such a diagram; and that the course of the reasoning may well be affected (e.g. facilitated) by choosing to do so. In a pretty untendentious sense, the diagram functions to present triangles (or triangularity) to thought; OK so far.

But notice a crucial difference between a diagram that functions as a mode of presentation and a Fregean sense that does: in the former case, there's more—lots more—than one kind of object that the diagram can be used to present. The very same diagram can represent now triangles, now equilateral triangles, now closed figures at large, now three-sided figures at large . . . etc. depending on *what intentional relation the reasoner bears to it*; depending, if you like, on how the reasoner entertains it. In this sort of case, then, *lots* of concepts correspond to the same mode of presentation. Or, putting it the other way round, what corresponds to the reasoner's concept is not the mode of presentation per se, but the mode of presentation *together with how it is entertained*.

A diagram can be used in all sorts of ways to present things to thought, but a Fregean sense can't be *on pain of senses failing to individuate concepts*; which is, after all, what they were invoked for in the first place. So,

question: what stops senses from behaving like diagrams? What guarantees that each sense can serve in only one way to present an object to a thought? I think that, on the Frege architecture with 5.3 in force, nothing prevents this except brute stipulation.

As far as I know, the standard discussions have pretty generally failed to recognize that Frege's architecture has this problem, so let me try once more to make clear just what the problem is. It's because there is more than one way to think about a *referent* that Frege needs to invoke MOPs to individuate concepts; *referents* can't individuate concepts because lots of different concepts can have the same referent. Fine. But Frege holds that MOPs *can* individuate concepts; that's what MOPs are *for*. So he mustn't allow that different MOPs can correspond to the same concept, *nor may he allow that a MOP can correspond to a concept in more than one way*. If he did, then each way of entertaining the MOP would (presumably) correspond to a different way of thinking the referent, and hence (presumably) to a different concept of the referent. Whereas MOPs are supposed to correspond to concepts one-to-one.

So, the question that I'm wanting to commend to you is: what, if anything, supports the prohibition against proliferating ways of grasping MOPs? Frege's story can't be: 'There is only one way of thinking a referent corresponding to each mode of presentation of the referent because there is only one way of entertaining each mode of presentation of a referent; and there is only one way of entertaining each mode of presentation of a referent because I say that's all there is.' Frege needs something that can *both* present referents to thought *and individuate thoughts*; in effect, he needs a kind of MOP that is *guaranteed* to have only one handle. He can't, however, get one just by wanting it; he has to explain *how there could be such things*. And 5.3 is in his way.

I think that if MOPs can individuate concepts and referents can't, that must be because MOPs *are mental objects* and referents aren't. Mental objects are *ipso facto* available to be proximal causes of mental processes; and it's plausible that at least some mental objects are distinguished by the kinds of mental processes that they cause; i.e. they are functionally distinguished.[13] Suppose that MOPs are in fact so distinguished. Then it's hardly surprising that there is only one way a mind can entertain each MOP: since, on this ontological assumption, functionally equivalent MOPs are *ipso facto* identical, the question 'Which MOP are you

[13] This doesn't, please notice, commit me to holding that the individuation of thought *content* is functional. Roughly, that depends on whether Frege is right that whatever can distinguish coextensive concepts is *ipso facto* the *sense* of the concepts; i.e. it depends on assuming 5.1. Which, however, I don't; see above.

entertaining?' and the question 'Which functional state is your mind in when you entertain it?' are required to get the same answer.

Frege's structural problem is that, though he wants to be an *externalist* about MOPs, the architecture of his theory won't let him.[14] Frege's reason for wanting to be an externalist about MOPs is that he thinks, quite wrongly, that if MOPs are mental then concepts won't turn out to be public. But if MOPs *aren't* mental, what kind of thing *could* they be such that *necessarily* for each MOP there is only one way in which a mind can entertain it? (And/or: what kind of mental state could entertaining a MOP be such that *necessarily* there is only one way to entertain each MOP?) As far as I can tell, Frege's story offers nothing at all to scratch this itch with.

If, however, MOPS are in the head,[15] then they can be proximal mental causes and are, to that extent, apt for functional individuation. If MOPs are both in the head and functionally individuated, *then a MOP's identity can be constituted by what happens when you entertain it*.[16] And if the identity of a MOP is *constituted* by what happens when you entertain it, then *of course* there is only one way to entertain each MOP. In point of metaphysical necessity, the alleged 'different ways of entertaining a MOP' would really be ways of entertaining different MOPs.

The moral, to repeat, is that even Frege can't have 5.3 if he holds onto 5.1. Even Frege should have been a mentalist about MOPs if he wished to remain in other respects a Fregean. On the other hand (perhaps this goes without saying), to claim that MOPs must be *mental* objects is quite compatible with also claiming that they are *abstract* objects, and that abstract objects are *not* mental. The apparent tension is reconciled by taking MOPS-qua-things-in-the-head to be the tokens of which MOPS-qua-abstract-objects are the types. It seems that Frege thought that if meanings can be shared it somehow follows that they can't also be

[14] In this usage, an 'externalist' is somebody who says that 'entertaining' relates a creature to something mind-independent, so Frege's externalism is entailed by his Platonism. Contrast the prima facie quite different Putnam/Kripke notion, in which an externalist is somebody who says that what you are thinking depends on what world you're in. (Cf. Preti 1992, where the distinction between these notions of externalism is sorted out, and some of the relations between them are explored.)

[15] This way of talking is, of course, entirely compatible with the current fashions in Individualism, Twins, and the like. Twins are supposed to show that referents can distinguish concepts whose causal roles are the same. For the demonstration to work, however, you've got to assume that Twins *ipso facto* have the causal roles of their concepts in common; viz. that whatever *contents* may supervene on, what *causal roles* supervene on is *inside* the head. That's precisely what I'm supposing in the text.

[16] Notice that this is not to say that *concepts* are individuated by the mental processes they cause, since a concept is a MOP together with a content; and I've taken an informational view of the individuation of contents. It's thus open to my version of RTM that 'Twin-Earth' cases involve concepts with different contents but the same MOPs.

particulars. But it beats me why he thought so. You might as well argue from '*being a vertebrate* is a universal' to 'spines aren't things'.

We're almost through with this, but I do want to tell you about an illuminating remark that Ernie Sosa once made to me. I had mentioned to Ernie that I was worried about why, though there are lots of ways to grasp a referent, there's only one way to grasp a MOP. He proceeded to pooh-pooh my worry along the following lines. "Look," he said, "it's pretty clear that there is only one way to instantiate a property, viz. by having it. It couldn't be, for example, that the property *red* is instantiated sometimes by a thing's being red and sometimes by a thing's being green. I don't suppose that worries you much?" (I agreed that it hadn't been losing me sleep.) "Well," he continued, with a subtle smile, "*if you aren't worried about there being only one way to instantiate a property, why are you worried about there being only one way to grasp a mode of presentation*?"

I think that's very clever, but I don't think it will do. The difference is this: It is surely plausible on the face of it that 'instantiating property *P*' is just *being P*; being red is all that there is to instantiating *redness*. But MOP is a technical notion in want of a metaphysics. If, as seems likely, the identity of a mental state turns on its causal role, then if MOPs are to individuate mental states they will have to be the sorts of things that the causal role of a mental state can turn on. But it's a mystery how a MOP *could* be that sort of thing if MOPs aren't in the head. If (to put the point a little differently) their non-mental *objects* can't distinguish thoughts, how can MOPS distinguish thoughts if they are non-mental too? It's as though the *arithmetic* difference between 3 and 4 could somehow explain the *psychological* difference between thinking about 3 and thinking about 4.

That red things are what instantiate redness is a truism, so you can have it for free. But Frege can't have it for free that, although same denotation doesn't mean same mental state, *same MOP* does. That must depend on some pretty deep difference between the *object* of thought and its *vehicle*. Offhand, the only difference I can think of that would do the job is ontological; it requires MOPs to be individuated by their roles as causes and effects of mental states, and hence to themselves be mental. So I think we should worry about why there's only one way to grasp a MOP even though I quite agree that we shouldn't worry about why there's only one way to instantiate a property.

Well, then, that's pretty much it for the background theory. All that remains is to add that in for a penny, in for a pound; having gone as far as we have, we might as well explicitly assume that MOPs are mental representations. That, surely, is the natural thing to say if you're supposing, on the one hand, that MOPs are among the proximal determinants of mental processes (as per Thesis Five) and that mental processes are

computations on structured mental representations (as per Thesis Two). It's really the basic idea of RTM that Turing's story about the nature of mental processes provides the very candidates for MOP-hood that Frege's story about the individuation of mental states independently requires. If that's true, it's about the nicest thing that ever happened to cognitive science.

So I shall assume that it is true. From here on, I'll take for granted that wherever mental states with the same satisfaction conditions have different intentional objects (like, for example, wanting to swallow the Morning Star and wanting to swallow the Evening Star) there must be corresponding differences among the mental representations that get tokened in the course of having them.

Now, finally, we're ready to get down to work. I'm interested in such questions as: 'What is the structure of the concept DOG?' Given RTM as the background theory, this is equivalent to the question: 'What is the MOP in virtue of entertaining which thoughts have dogs as their intentional objects?' And this is in turn equivalent to the question: 'What is the structure of the mental representation DOG?'

And my answer will be that, on the evidence available, it's reasonable to suppose that such mental representations *have no structure*; it's reasonable to suppose that they are atoms.

2

Unphilosophical Introduction: What Concepts Have To Be

THIS is a book about concepts. Two of its main theses are:

—that if you are going to run a representational/computational theory of mind (that is, any version of RTM; see Chapter 1) you will need a theory of concepts.

And:

—that none of the theories of concepts that are currently taken at all seriously either in cognitive science or in philosophy can conceivably fill the bill.

To argue this, I shall first need to say what bill it is that needs to be filled. That's the burden of this chapter. I want to set out five conditions that an acceptable theory of concepts would have to meet. Several chapters following this one will be devoted to making clear by how much, and for what reasons, current theories of concepts fail to meet them.

A word about the epistemic status of the conditions I'm about to endorse: I regard them as fallible but not negotiable. Not negotiable, that is, short of giving up on RTM itself; and RTM remains the only game in town, even after all these years. In effect, I'm claiming that these constraints on concepts follow just from the architecture of RTMs together with some assumptions about cognitive processes and capacities which, though certainly contingent, are none the less hardly possible to doubt. (I mean, of course, hardly possible to doubt really, not hardly possible to doubt philosophically.) If this is indeed the status of these constraints, then I think we had better do what we can to construct a theory of concepts that satisfies them.

So, then, here are my five not-negotiable conditions on a theory of concepts.

1. Concepts are mental particulars; specifically, they satisfy whatever ontological conditions have to be met by things that function as mental causes and effects.

Since this is entailed by RTM (see Chapter 1), and hence is common to all the theories of concepts I'll consider, I won't go on about it here. If, however, you think that intentional causation explains behaviour only in the way that the solubility of sugar explains its dissolving (see Ryle 1949), or if you think that intentional explanations aren't causal at all (see e.g. Collins 1987), then nothing in the following discussion will be of much use to you, and I fear we've reached a parting of the ways. At least one of us is wasting his time; I do hope it's you.

2. Concepts are categories and are routinely employed as such.

To say that concepts are categories is to say that they apply to things in the world; things in the world 'fall under them'. So, for example, Greycat the cat, but not Dumbo the elephant, falls under the concept CAT. Which, for present purposes, is equivalent to saying that Greycat is in the extension of CAT, that 'Greycat is a cat' is true, and that 'is a cat' is true of Greycat. I shall sometimes refer to this galaxy of considerations by saying that applications of concepts are susceptible of '*semantic evaluation*': claims, or thoughts, that a certain concept applies to a certain thing are always susceptible of evaluation in such semantical terms as satisfied/unsatisfied, true/false, correct/incorrect, and the like. There are, to be sure, issues about these various aspects of semantic evaluability, and about the relations among them, that a scrupulous philosopher might well wish to attend to. But in this chapter, I propose to keep the philosophy to a bare minimum.[1]

Much of the life of the mind consists in applying concepts to things. If I think *Greycat is a cat* (de dicto, as it were), I thereby apply the concept CAT to Greycat (correctly, as it happens). If, looking at Greycat, I take him to be a cat, then too I apply the concept CAT to Greycat. (If looking at Greycat I take him to be a meatloaf, I thereby apply the concept MEATLOAF to Greycat; incorrectly, as it happens.) Or if, in reasoning about Greycat, I infer that since he's a cat he must be an animal, I thereby proceed from applying one concept to Greycat to the licensed application of another concept; the license consisting, I suppose, in things I know about how the extensions of the concepts CAT and ANIMAL are related.

In fact, RTM being once assumed, most of cognitive psychology, including the psychology of memory, perception, and reasoning, is about how we apply concepts. And most of the rest is about how we acquire the concepts that we thus apply. Correspondingly, the empirical data to which cognitive psychologists are responsible consist largely of measures of subject performance in concept application tasks. The long and short is: whatever else a theory of concepts says about them, it had better exhibit

[1] Or, at least, to confine it to footnotes.

concepts as the sorts of things that get applied in the course of mental processes. I take it that consensus about this is pretty general in the cognitive sciences, so I won't labour it further here.

Caveat: it's simply untendentious that concepts have their satisfaction conditions essentially. Nothing in any mental life could be the concept CAT unless it is satisfied by cats. It couldn't be that there are some mental lives in which the concept CAT applies to CATS and others in which it doesn't. If you haven't got a concept that applies to cats, that *entails* that you haven't got the CAT concept. But though the *satisfaction* conditions of a concept are patently among its essential properties, it does not follow that the *confirmation* conditions of a concept are among its essential properties. Confirmation is an epistemic relation, not a semantic relation, and it is generally theory mediated, hence holistic. On the one hand, given the right background theory, the merest ripple in cat infested waters might serve to confirm an ascription of cathood; and, on the other hand, no cat-containing layout is so well lit, or so utterly uncluttered, or so self-certifying that your failure to ascribe cathood therein would *entail* that you lack the concept. In short, it is OK to be an atomist about the metaphysical conditions for a concept's *having* satisfaction conditions (which I am and will try to convince you to be too), and yet be a holist about the confirmation of claims that a certain concept is satisfied in a certain situation. Shorter still: just as Quine and Duhem and those guys taught us, *there aren't any criteria*. So at least I shall assume throughout what follows.

3. Compositionality: concepts are the constituents of thoughts and, in indefinitely many cases, of one another. Mental representations inherit their contents from the contents of their constituents.

Some terminology: I'll use 'thoughts' as my cover term for the mental representations which, according to RTMs, express the propositions that are the objects of propositional attitudes. Thus, a belief that it will rain and a hope that it will rain share a thought as well as a proposition which that thought expresses. For present purposes, it will do to think of thoughts as mental representations analogous to closed sentences, and concepts as mental representations analogous to the corresponding open ones. It may strike you that mental representation is a lot like language, according to my version of RTM. Quite so; how could language express thought if that were not the case?

Qua constituents of thoughts, and of each other, concepts play a certain role in explaining why the propositional attitudes are productive and systematic. The outlines of this story are well known, though by no means untendentious:

Beliefs are *productive* in that there are infinitely many distinct ones that a person can entertain (given, of course, the usual abstraction from 'performance limitations'). Beliefs are *systematic* in that the ability to entertain any one of them implies the ability to entertain many others that are related to it in content. It appears, for example, to be conceptually possible that there should be a mind that is able to grasp the proposition that Mary loves John but not able to grasp the proposition that John loves Mary. But, in point of empirical fact, it appears that there are no such minds. This sort of symmetry of cognitive capacities is a ubiquitous feature of mental life.[2] It implies a corresponding symmetry of representational capacities since RTM says, 'no cognition without representation'. That is, RTM says that you can't grasp a proposition without entertaining a thought.

So, the question presents itself: what must mental representation be like if it is to explain the productivity and systematicity of beliefs? This question is loaded, to be sure: that the systematicity of the attitudes requires the systematicity of mental representation doesn't itself require that the systematicity of mental representation is what explains the systematicity of the attitudes. Perhaps both are the effects of a common cause. Maybe, for example, 'the world' somehow teaches the mind to be systematic, and the systematicity of mental representation is the by-product of its doing so.

The stumbling-block for this sort of suggestion is that the mind is much more systematic than the world: that John loves Mary doesn't make it true, or even very likely, that Mary reciprocates. Sad for John, of course, but where would The Western Canon be if things were otherwise? In fact, the only thing in the world that is as systematic as thought is language. Accordingly, some philosophers (Dan Dennett 1993 in particular) have suggested that it's *learning language* that makes a mind systematic.

But we aren't told how an initially unsystematic mind *could* learn a systematic language, given that the latter is *ipso facto* able to express propositions that the former is unable to entertain. How, for example, does a mind that can think that *John loves Mary* but not that *Mary loves John* learn a language that is able to say both? Nor is it clear what could make *language itself* systematic if not the systematicity of the thoughts that it is used to express; so the idea that the mind learns systematicity from language just sweeps the problem from under the hall rug to under the rug in the parlour. On balance, I think we had better take it for granted,

[2] It bears emphasis that systematicity concerns symmetries of cognitive *capacities*, not of actual mental states. It is, for example, patently not the case that whoever thinks that Mary loves John also thinks that John loves Mary. Compare van Gelder and Nicklasson 1994.

and as part of what is not negotiable, that systematicity and productivity are grounded in the 'architecture' of mental representation and not in the vagaries of experience. If a serious alternative proposal should surface, I guess I'm prepared to reconsider what's negotiable. But the prospect hasn't been losing me sleep.

So, to repeat the question, what is it about mental representation that explains the systematicity and productivity of belief? Classical versions of RTM offer a by now familiar answer: there are infinitely many beliefs because there are infinitely many thoughts to express their objects. There are infinitely many thoughts because, though each mental representation is constructed by the application of a finite number of operations to a finite basis of primitive concepts, there is no upper bound to how many times such operations may apply in the course of a construction. Correspondingly, thought is systematic because the same primitive concepts and operations that suffice to assemble thoughts like JOHN LOVES MARY also suffice to assemble thoughts like MARY LOVES JOHN; the representational capacity that is exploited to frame one thought implies the representational capacity to frame the other. Since a mental representation is individuated by its form and content (see Chapter 1), both of these are assumed to be determined by specifying the inventory of primitive concepts that the representation contains, together with the operations by which it is assembled from them. (In the case of the primitive concepts themselves, this assumption is trivially true.) As a shorthand for all this, I'll say that what explains the productivity and systematicity of the propositional attitudes is the *compositionality* of concepts and thoughts.

The requirement that the theory of mental representation should exhibit thoughts and concepts as compositional turns out, in fact, to be quite a powerful analytic engine. If the content of a mental representation is inherited from the contents of its conceptual constituents then, presumably, the content of a constituent concept is just whatever it can contribute to the content of its hosts. We'll see, especially in Chapter 5, that this condition is not at all easy for a theory of concepts to meet.

4. Quite a lot of concepts must turn out to be learned.

I want to put this very roughly since I'm going to return to it at length in Chapter 6. Suffice it for now that all versions of RTM hold that if a concept belongs to the primitive basis from which complex mental representations are constructed, it must *ipso facto* be *un*learned. (To be sure, some versions of RTM are rather less up front in holding this than others.) Prima facie, then, where a theory of concepts draws the distinction between what's primitive and what's not is also where it draws the

distinction between what's innate and what's not. Clearly, everybody is going to put this line somewhere. For example, nobody is likely to think that the concept BROWN COW is primitive since, on the face of it, BROWN COW has BROWN and COW as constituents. Correspondingly, nobody is likely to think that the concept BROWN COW is innate since, on the face of it, it could be learned by being assembled from the previously mastered concepts BROWN and COW.

A lot of people have Very Strong Feelings about what concepts are allowed to be innate,[3] hence about how big a primitive conceptual basis an acceptable version of RTM can recognize. Almost everybody is prepared to allow RED in, and many of the liberal-minded will also let in CAUSE or AGENT. (See, for example, Miller and Johnson-Laird 1978). But there is, at present, a strong consensus against, as it might be, DOORKNOB or CARBURETTOR. I have no desire to join in this game of pick and choose since, as far as I can tell, it hasn't any rules. Suffice it that it would be nice if a theory of concepts were to provide a principled account of what's in the primitive conceptual basis, and it would be nice if the principles it appealed to were to draw the distinction at some independently plausible place. (Whatever, if anything, that means.) Chapter 6 will constitute an extended reconsideration of this whole issue, including the question just how the relation between a concept's being primitive and its being innate plays out. I hope there to placate such scruples about DOORKNOB and CARBURETTOR as some of you may feel, and to do so within the framework of an atomistic RTM.

5. Concepts are *public*; they're the sorts of things that lots of people can, and do, *share*.

Since, according to RTM, concepts are symbols, they are presumed to satisfy a type/token relation; to say that two people share a concept (i.e. that they have literally the same concept) is thus to say that they have tokens of literally the same concept type. The present requirement is that the conditions for typing concept tokens must not be so stringent as to assign practically every concept token to a different type from practically any other.

[3] I put it this way advisedly. I was once told, in the course of a public discussion with an otherwise perfectly rational and civilized cognitive scientist, that he "could not permit" the concept HORSE to be innate in humans (though I guess it's OK for it to be innate in horses). I forgot to ask him whether he was likewise unprepared to permit neutrinos to lack mass.

Just why feelings run so strongly on these matters is unclear to me. Whereas the ethology of all other species is widely agreed to be thoroughly empirical and largely morally neutral, a priorizing and moralizing about the ethology of our species appears to be the order of the day. Very odd.

It seems pretty clear that all sorts of concepts (for example, DOG, FATHER, TRIANGLE, HOUSE, TREE, AND, RED, and, surely, lots of others) are ones that all sorts of people, under all sorts of circumstances, have had and continue to have. A theory of concepts should set the conditions for concept possession in such a way as not to violate this intuition. Barring very pressing considerations to the contrary, it should turn out that people who live in very different cultures and/or at very different times (me and Aristotle, for example) both have the concept FOOD; and that people who are possessed of very different amounts of mathematical sophistication (me and Einstein, for example) both have the concept TRIANGLE; and that people who have had very different kinds of learning experiences (me and Helen Keller, for example) both have the concept TREE; and that people with very different amounts of knowledge (me and a four-year-old, for example) both have the concept HOUSE. And so forth. Accordingly, if a theory or an experimental procedure distinguishes between my concept DOG and Aristotle's, or between my concept TRIANGLE and Einstein's, or between my concept TREE and Helen Keller's, etc. that is a very strong prima facie reason to doubt that the theory has got it right about concept individuation or that the experimental procedure is really a measure of concept possession.

I am thus setting my face against a variety of kinds of conceptual relativism, and it may be supposed that my doing so is itself merely dogmatic. But I think there are good grounds for taking a firm line on this issue. Certainly RTM is required to. I remarked in Chapter 1 that RTM takes for granted the centrality of intentional explanation in any viable cognitive psychology. In the cases of interest, what makes such explanations intentional is that they appeal to covering generalizations about people who believe that such-and-such, or people who desire that so-and-so, or people who intend that this and that, and so on. In consequence, the extent to which an RTM can achieve generality in the explanations it proposes depends on the extent to which mental contents are supposed to be shared. If everybody else's concept WATER is different from mine, then it is literally true that only I have ever wanted a drink of water, and that the intentional generalization 'Thirsty people seek water' applies only to me. (And, of course, only I can state that generalization; words express concepts, so if your WATER concept is different from mine, 'Thirsty people seek water' means something different when you say it and when I do.) Prima facie, it would appear that any very thoroughgoing conceptual relativism would preclude intentional generalizations with any very serious explanatory power. This holds in spades if, as seems likely, a coherent conceptual relativist has to claim that conceptual identity can't be maintained even across time slices of the same individual.

There is, however, a widespread consensus (and not only among conceptual relativists) that intentional explanation can, after all, be preserved without supposing that belief contents are often—or even ever—literally public. The idea is that a robust notion of content *similarity* would do just as well as a robust notion of content *identity* for the cognitive scientist's purposes. Here, to choose a specimen practically at random, is a recent passage in which Gil Harman enunciates this faith:

> Sameness of meaning from one symbol system to another is a similarity relation rather than an identity relation in the respect that sameness of meaning is not transitive . . . I am inclined to extend the point to concepts, thoughts, and beliefs . . . The account of sameness of content appeals to the best way of translating between two systems, where goodness in translation has to do with preserving certain aspects of usage, with no appeal to any more 'robust' notion of content or meaning identity . . . [There's no reason why] the resulting notion of sameness of content should fail to satisfy the purposes of intentional explanation. (1993: 169–79)[4]

It's important whether such a view can be sustained since, as we'll see, meeting the requirement that intentional contents be literally public is non-trivial; like compositionality, publicity imposes a substantial constraint upon one's theory of concepts and hence, derivatively, upon one's theory of language. In fact, however, the idea that content similarity is the basic notion in intentional explanation is affirmed a lot more widely than it's explained; and it's quite unclear, on reflection, how the notion of similarity that such a semantics would require might be unquestion-beggingly developed. On one hand, such a notion must be robust in the sense that it preserves intentional explanations pretty generally; on the other hand, it must do so *without itself presupposing a robust notion of content identity*. To the best of my knowledge, it's true *without exception* that all the construals of concept similarity that have thus far been put on offer egregiously fail the second condition.

Harman, for example, doesn't say much more about content-similarity-cum-goodness-of-translation than that it isn't transitive and that it "preserves certain aspects of usage". That's not a lot to go on. Certainly it leaves wide open whether Harman is right in denying that his account of content similarity presupposes a "'robust' notion of content or meaning identity". For whether it does depends on how the relevant "aspects of

[4] See also Smith, Medin, and Rips: "what accounts for categorization cannot account for stability [publicity] . . . [a]s long as *stability of concepts* is equated with *sameness of concepts* . . . But there is another sense of stability, which can be equated with *similarity of mental contents* . . . and for this sense, what accounts for categorization may at least partially account for 'stability'"(1984: 268). Similar passages are simply ubiquitous in the cognitive science literature; I'm grateful to Ron Mallon for having called this example to my attention.

usage" are themselves supposed to be individuated, and about this we're told nothing at all.

Harman is, of course, too smart to be a behaviourist; 'usage', as he uses it, is itself an intentional-cum-semantic term. Suppose, what surely seems plausible, that one of the 'aspects of usage' that a good translation of 'dog' has to preserve is that it be a term that implies *animal*, or a term that doesn't apply to ice cubes, or, for matter, a term that means *dog*. If so, then we're back where we started; Harman needs notions like *same* implication, *same* application, and *same* meaning in order to explicate his notion of content similarity. All that's changed is which shell the pea is under.

At one point, Harman asks rhetorically, "What aspects of use determine meaning?" Reply: "It is certainly relevant what terms are applied to and the reasons that might be offered for this application . . . it is also relevant how some terms are used in relation to other terms" (ibid.: 166). But I can't make any sense of this unless some notion of 'same application', 'same reason', and 'same relation of terms' is being taken for granted in characterizing what good translations *ipso facto* have in common. NB on pain of circularity: *same* application (etc.), not *similar* application (etc.). Remember that *similarity of semantic properties* is the notion that Harman is trying to explain, so his explanation mustn't *presuppose* that notion.

I don't particularly mean to pick on Harman; if his story begs the question it was supposed to answer, that is quite typical of the literature on concept similarity. Though it's often hidden in a cloud of technical apparatus (for a detailed case study, see Fodor and Lepore 1992: ch. 7), the basic problem is easy enough to see. Suppose that we want the following to be a prototypical case where you and I have different but similar concepts of George Washington: though we agree about his having been the first American President, and the Father of His Country, and his having cut down a cherry tree, and so on, you think that he wore false teeth and I think that he didn't. The similarity of our GW concepts is thus some (presumably weighted) function of the number of propositions about him that we both believe, and the dissimilarity of our GW concepts is correspondingly a function of the number of such propositions that we disagree about. So far, so good.

But the question now arises: what about the shared beliefs themselves; are they or aren't they *literally* shared? This poses a dilemma for the similarity theorist that is, as far as I can see, unavoidable. If he says that our agreed upon beliefs about GW are literally shared, then he hasn't managed to do what he promised; viz. introduce a notion of similarity of content that dispenses with a robust notion of publicity. But if he says

that the agreed beliefs aren't literally shared (viz. that they are only required to be similar), then his account of content similarity begs the very question it was supposed to answer: his way of saying what it is for concepts to have similar but not identical contents presupposes a prior notion of beliefs with similar but not identical contents.

The trouble, in a nutshell, is that all the obvious construals of *similarity of beliefs* (in fact, all the construals that I've heard of) take it to involve *partial overlap* of beliefs.[5] But this treatment breaks down if the beliefs that are in the overlap are themselves construed as similar but not identical. It looks as though a robust notion of content similarity *can't but* presuppose a correspondingly robust notion of content identity. Notice that this situation is not symmetrical; the notion of content identity doesn't require a prior notion of content similarity. Leibniz's Law tells us what it is for the contents of concepts to be identical; Leibniz's Law tells us what it is for *anythings* to be identical.

As I remarked above, different theorists find different rugs to sweep this problem under; but, as far as I can tell, none of them manages to avoid it. I propose to harp on this a bit because confusion about it is rife, not just in philosophy but in the cognitive science community at large. Not getting it straight is one of the main things that obscures how very hard it is to construct a theory of concepts that works, and how very much cognitive science has thus far failed to do so.

Suppose, for example, it's assumed that your concept PRESIDENT is similar to my concept PRESIDENT in so far as we assign similar subjective probabilities to propositions that contain the concept. There are plenty of reasons for rejecting this sort of model; we'll discuss its main problems in Chapter 5. Our present concern is only whether constructing a probabilistic account of concept similarity would be a way to avoid having to postulate a robust notion of content identity.

Perhaps, in a typical case, you and I agree that p is very high for 'FDR is/was President' and for 'The President is the Commander-in-Chief of the Armed Forces' and for 'Presidents have to be of voting age', etc.; but, whereas you rate 'Millard Fillmore is/was President' as having a probability close to 1, I, being less well informed, take it to be around p = 0.07 (*Millard Fillmore*???). This gives us an (arguably) workable construal of the idea that we have similar but not identical PRESIDENT concepts. But it does so only by helping itself to a prior notion of belief identity, and to the assumption that there are lots of thoughts of which

[5] 'Why not take content similarity as primitive and *stop trying* to construe it?' Sure; but then why not take content *identity* as primitive and stop trying to construe *it*? In which case, what is semantics *for*?

our respective PRESIDENTs are constituents that we literally share. Thus, you and I are, by assumption, both belief-related to the thoughts that Millard Fillmore was President, that Presidents are Commanders-in-Chief, etc. The difference between us is in the *strengths* of our beliefs, not in their contents.[6] And, as usual, it really does seem to be *identity* of belief content that's needed here. If our respective beliefs about Presidents having to be of voting age were supposed to be merely *similar*, circularity would ensue: since content similarity is the notion we are trying to explicate, it mustn't be among the notions that the explication presupposes. (I think I may have mentioned that before.)

The same sort of point holds, though even more obviously, for other standard ways of construing conceptual similarity. For example, if concepts are sets of features, similarity of concepts will presumably be measured by some function that is sensitive to the amount of overlap of the sets. But then, the atomic feature assignments must themselves be construed as literal. If the similarity between your concept CAT and mine depends (*inter alia*) on our agreement that '+ has a tail' is in both of our feature bundles, then the assignment of that feature to these bundles must express a literal consensus; it must literally be the property of *having a tail* that we both literally think that cats literally have. (As usual, nothing relevant changes if feature assignments are assumed to be probabilistic or weighted; or if the feature assigned are supposed to be "subsemantic", though these red herrings are familiar from the Connectionist literature.)

Or, suppose that concepts are thought of as positions in a "multi-dimensional vector space" (see e.g. Churchland 1995) so that the similarity between your concepts and mine is expressed by the similarity of their positions in our respective spaces. Suppose, in particular, that it is constitutive of the difference between our NIXON concepts that you think Nixon was even more of a crook than I do. Once again, a robust notion of content identity is presupposed since each of our spaces is required to have a dimension that expresses crookedness; a fortiori, both are required

[6] Alternatively, a similarity theory might suppose that what we share when our PRESIDENT concepts are similar are similar beliefs about the probabilities of certain propositions: you believe that p(presidents are CICs) = 0.98; I believe that p(presidents are CICs) = 0.95; Bill believes that p(Presidents are CICs) = 0.7; so, all else equal, your PRESIDENT concept is more like mine than Bill's is.

But this construal does nothing to discharge the basic dependence of the notion of content similarity on the notion of content identity since what it says makes our beliefs similar is that they make similar estimates of the probability *of the very same proposition*; e.g. of the proposition that presidents are CICs. If, by contrast, the propositions to which our various probability estimates relate us are themselves supposed to be merely similar, then it does *not* follow from these premisses that *ceteris paribus* your PRESIDENT concept is more like mine than like Bill's.

to have dimensions which express degrees of *the very same property*. That should seem entirely unsurprising. Vector space models identify the dimensions of a vector space *semantically* (viz. by stipulating what the location of a concept along that dimension is to *mean*), and it's just a truism that the positions along dimension *D* can represent degrees of *D*-ness only in a mind that possesses the concept of being *D*. You and I can argue about whether Nixon was merely crooked or very crooked only if the concept of *being crooked* is one that we have in common.

It may seem to you that I am going on about such truisms longer than necessity demands. It often seems that to me, too. There are, however, at least a zillion places in the cognitive science literature, and at least half a zillion in the philosophy literature, where the reader is assured that some or all of his semantical troubles will vanish quite away if only he will abandon the rigid and reactionary notion of content identity in favour of the liberal and laid-back notion of content similarity. But in none of these places is one ever told how to do so. That's because nobody has the slightest idea how. In fact, it's all just loose talk, and it causes me to grind my teeth.

Please note that none of this is intended to claim that notions like belief similarity, content similarity, concept similarity, etc. play less than a central role in the psychology of cognition. On the contrary, for all I know (certainly for all I am prepared non-negotiably to assume) it may be that every powerful intentional generalization is of the form "If *x* has a belief similar to *P*, then . . ." rather than the form "If *x* believes *P*, then . . .". If that is so, then so be it. My point is just that assuming that it is so doesn't exempt one's theory of concepts from the Publicity constraint. To repeat one last time: all the theories of content that offer a robust construal of conceptual similarity do so by presupposing a correspondingly robust notion of concept identity. As far as I can see, this is unavoidable. If I'm right that it is, then the Publicity constraint is *ipso facto* non-negotiable.

OK, so those are my five untendentious constraints on theories of concepts. In succeeding chapters, I'll consider three stories about what concepts are; viz. that they are definitions; that they are prototypes/stereotypes; and (briefly) something called the 'theory theory' which says, as far as I can make out, that concepts are abstractions from belief systems. I'll argue that each of these theories violates at least one of the non-negotiable constraints; and that it does so, so to speak, not a little bit around the edges but egregiously and down the middle. We will then have to consider what, if any, options remain for developing a theory of concepts suitable to the purposes of an RTM.

Before we settle down to this, however, there are a last couple of preliminary points that I want to put in place.

Here is the first: although I'm distinguishing three theories of concepts for purposes of exposition and attack, and though supporters of each of these theories have traditionally wanted to distance themselves as much as possible from supporters of the others, still all three theories are really versions of one and the same idea about content. I want to stress this since I'm going to argue that it is primarily because of what they agree about that all three fail.

The theories of concepts we'll be considering all assume a metaphysical thesis which, as I remarked in Chapter 1, I propose to reject: namely, that primitive concepts, and (hence) their possession conditions, are at least partly constituted by their inferential relations. (That complex concepts—BROWN COW, etc.—and their possession conditions are exhaustively constituted by their inferential relations to their constituent concepts is not in dispute; to the contrary, compositionality requires it, and compositionality isn't negotiable.) The current near-universal acceptance of Inferential Role Semantics in cognitive science marks a radical break with the preceding tradition in theories about mind and language: pre-modern theories typically supposed that primitive concepts are individuated by their (e.g. iconic or causal) relations to things in the world. The history of the conversion of cognitive scientists to IR semantics would make a book by itself; a comedy, I think, though thus far without a happy ending:

—In philosophy, the idea was pretty explicitly to extend the Logicist treatment of logical terms into the non-logical vocabulary; if IF and SOME can be identified with their inferential roles, why not TABLE and TREE as well?

—In linguistics, the idea was to extend to semantics the Structuralist notion that a level of grammatical description is a 'system of differences': if their relations of equivalence and contrast are what bestow phonological values on speech sounds, why shouldn't their relations of implication and exclusion be what bestow semantic values on forms of words?

—In AI, the principle avatar of IRS was 'procedural semantics', a deeply misguided attempt to extend the principle of 'methodological solipsism' from the theory of mental processes to the theory of meaning: if a mental process (thinking, perceiving, remembering, and the like) can be 'purely computational' why can't conceptual content be purely computational too? If computers qua devices that perform inferences can *think*, why can't computers qua devices that perform inferences *mean*?

—I don't know how psychology caught IRS; perhaps it was from philosophy, linguistics, and AI. (I know one eminent developmental psychologist who certainly caught it from Thomas Kuhn.) Let that be an object lesson in the danger of mixing disciplines. Anyhow, IRS got to be

the fashion in psychology too. Perhaps the main effect of the "cognitive revolution" was that espousing some or other version of IRS became the received way for a psychologist not to be a behaviourist.

So, starting around 1950, practically everybody was saying that the '"Fido"–Fido fallacy' is fallacious,[7] and that concepts (/words) are like chess pieces: just as there can't be a rook without a queen, so there can't be a DOG without an ANIMAL. Just as the value of the rook is partly determined by its relation to the queen, so the content of DOG is partly determined by its relation to ANIMAL. Content is therefore a thing that can only happen internal to *systems* of symbols (or internal to languages, or, on some versions, internal to forms of life). It was left to 'literary theory' to produce the *reductio ad absurdum* (literary theory is good at that): content is constituted *entirely* by intra-symbolic relations; just as there's nothing 'outside' the chess game that matters to the values of the pieces, so too there's nothing outside the text that matters to what it means. Idealism followed, of course.

It is possible to feel that these various ways of motivating IRS, historically effective though they clearly were, are much less than overwhelmingly persuasive. For example, on reflection, it doesn't seem that languages are a lot like games after all: queens and pawns don't mean anything, whereas 'dog' means *dog*. That's why, though you can't translate the queen into French (or, a fortiori, into checkers), you can translate 'dog' into 'chien'. It's perhaps unwise to insist on an analogy that misses so glaring a difference.

Phonemes don't mean anything either, so prima facie, *pace* Saussure, "having a phonological value" and "having a semantic value" would seem to be quite different sorts of properties. Even if it were right that phonemes are individuated by their contrasts and equivalences—which probably they aren't—that wouldn't be much of a reason to claim that words or concepts are also individuated that way.

If, in short, one asks to hear some serious arguments for IRS, one discovers, a bit disconcertingly, that they are very thin upon the ground. I think that IRS is most of what is wrong with current theorizing in cognitive science and the metaphysics of meaning. But I don't suppose for a minute that any short argument will, or should, persuade you to consider junking it. I expect that will need a long argument; hence this long book. Long arguments take longer than short arguments, but they do sometimes create conviction.

Accordingly, my main subject in what follows will be not the history of

[7] That is, the "fallacy" of assuming that the meaning of the word is the eponymous dog.

IR semantics, or the niceties of its formulation, or its evidential status, but rather its impact on empirical theories of concepts. The central consideration will be this: If you wish to hold that the content of a concept is constituted by the inferences that it enters into, you are in need of a principled way of deciding *which inferences constitute which concepts*. What primarily distinguishes the cognitive theories we'll consider is how they answer this question. My line will be that, though as far as anybody knows the answers they offer exhaust the options, pretty clearly none of them can be right. Not, NB, that they are incoherent, or otherwise confused; just that they fail to satisfy the empirical constraints on theories of concepts that I've been enumerating, and are thus, almost certainly, false.

At that point, I hope that abandoning IRS in favour of the sort of atomistic, informational semantics that I tentatively endorsed in Chapter 1 will begin to appear to be the rational thing to do. I'll say something in Chapter 6 about what this sort of alternative to IRS might be like.

So much for the first of my two concluding addenda. Here is the second:

I promised you in Chapter 1 that I wouldn't launch yet another defence of RTM; I proposed—aside from my admittedly tendentious endorsement of informational semantics—simply to take RTM for granted as the context in which problems about the nature of concepts generally arise these days. I do mean to stick to this policy. Mostly. But I can't resist rounding off these two introductory chapters by remarking how nicely the pieces fit when you put them all together. I'm going to exercise my hobby-horse after all, but only a little.

In effect, in these introductory discussions, we've been considering constraints on a theory of cognition that emerge from two widely different, and largely independent, research enterprises. On the one hand, there's the attempt to save the architecture of a Fregean—viz. a purely referential—theory of meaning by taking seriously the idea that concepts can be distinguished by their 'modes of presentation' of their extensions. It's supposed to be modes of presentation that answer the question 'How can coreferential concepts be distinct?' Here Frege's motives concur with those of Informational Semantics; since both are referential theories of content, both need a story about how thinking about the Morning Star could be different from thinking about the Evening Star, given that the two thoughts are connected with the same 'thing in the world'.

The project of saving the Frege programme faces two major hurdles. First, 'Mates cases' appear to show that modes of presentations can't be senses. Frege to the contrary notwithstanding, it looks as though practically any linguistic difference between prima facie synonymous expressions, merely syntactic differences distinctly included, can be recruited to block their substitution in some Mates context or other. In the

current jargon, the individuation of the propositional attitudes apparently slices them about as thin as the syntactic individuation of forms of words, hence not only thinner than reference can, *but also thinner than sense can*.

The other obstacle to saving the Frege programme was that it took for granted that the semantic question 'How can coreferential concepts fail to be synonyms?' gets the same answer as the psychological question 'How can there be more than one way of grasping a referent?' The postulation of senses was supposed to answer both questions. I argued, however, that given Frege's Platonism about senses, it's by no means obvious why his answer to the first would constitute an answer to the second; in effect, Frege simply stipulates their equivalence. I supposed the moral to be that Frege's theoretical architecture needs to be explicitly psychologized. Modes of presentation need to be 'in the head'.

The short form is: the Frege programme needs something that is both in the head and of the right kind to distinguish coreferential concepts, and the Mates cases suggest that whatever is able to distinguish coreferential concepts is apt for syntactic individuation. Put all this together and it does rather suggest that modes of presentation are syntactically structured mental particulars. Suggestion noted.

The other research programme from which my budget of constraints on theories of concepts derived is the attempt, in cognitive science, to explain how a finite being might have intentional states and capacities that are productive and systematic. This productivity/systematicity problem again has two parts: 'Explain how there can be infinitely many propositional attitudes each with its distinctive propositional object (i.e. each with its own content)' and: 'Explain how there can be infinitely many propositional attitudes each with its distinctive causal powers (i.e. each with its own causal role in mental processes).' Here I have followed what Pylyshyn and I (Fodor and Pylyshyn 1988) called the 'Classical' computational tradition that proceeds from Turing: mental representations are syntactically structured. Their conditions of semantic evaluation and their causal powers both depend on their syntactic structures; the former because mental representations have a compositional semantics that is sensitive to the syntactic relations among their constituents; the latter because mental processes are *computations* and are thus syntactically driven by definition. So the Classical account of productivity/systematicity points in much the same direction as the psychologized Frege programme's account of the individuation of mental states: viz. towards syntactically structured mental particulars whose tokenings are matched, case for case, with tokenings of the de dicto propositional attitudes.

Syntactically structured mental particulars whose tokenings are matched, case for case, with tokenings of the de dicto propositional

attitudes are, of course, exactly what RTM has for sale. So RTM seems to be what both the Frege/Mates problems and the productivity/systematicity problems converge on. If beliefs (and the like) are relations to syntactically structured mental representations, there are indeed two parameters of belief individuation, just as Frege requires: Morning Star beliefs have the same conditions of semantic evaluation as Evening Star beliefs, but they implicate the tokening of different syntactic objects and are therefore different beliefs with different causal powers. That believing *P* and believing *Q* may be different mental states even if '*P*' and '*Q*' have the same semantic value shows up in the Mates contexts. That believing *P* and believing *Q* may have different causal powers even if '*P*' and '*Q*' have the same semantic value shows up in all those operas where the soprano dies of mistaken identity.

So RTM looks like a plausible answer to several questions that one might have supposed to be unrelated. I hope that isn't an accident. This book runs on the assumption that it isn't, hence that we need RTM a lot. RTM, in turn, needs a theory of concepts a lot since compositionality says that the contents and causal powers of mental representations are both inherited, eventually, from the contents and causal powers of their minimal constituents; viz. from the primitive concepts that they contain. RTM is simply *no good* without a viable theory of concepts.

So be it, then. Let's see what there might be on offer.