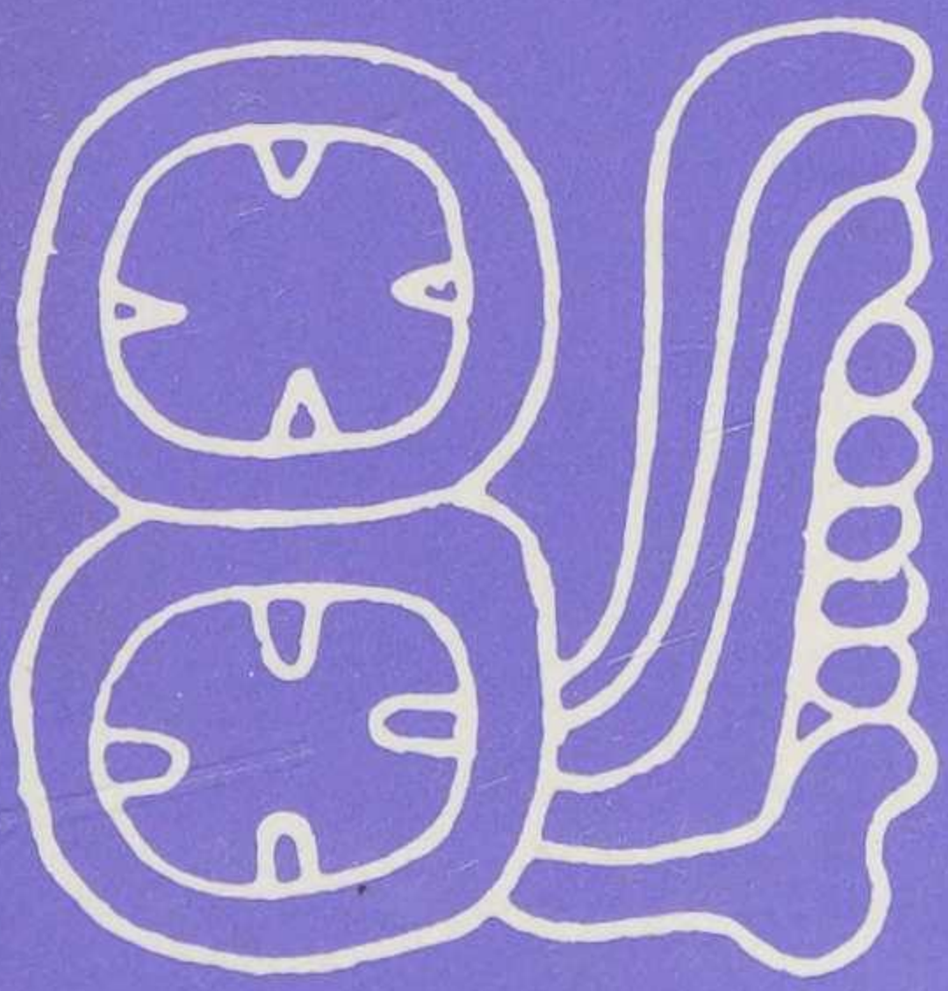
Philosophical Issues, 6, 1995

CONTENT

edited by

Enrique Villanueva
Sociedad Filosófica Ibero Americana



CONTENT

Philosophical Issues, 6, 1995

PHILOSOPHICAL ISSUES

Edited by Enrique Villanueva
(Universidad Nacional Autónoma de México)

EDITORIAL ADVISORY BOARD

Previously published volumes:

CONSCIOUSNESS
(*Philosophical Issues*, 1, 1991)

RATIONALITY IN EPISTEMOLOGY
(*Philosophical Issues*, 2, 1992)

SCIENCE AND KNOWLEDGE
(*Philosophical Issues*, 3, 1993)

NATURALISM AND NORMATIVITY
(*Philosophical Issues*, 4, 1993)

TRUTH AND RATIONALITY
(*Philosophical Issues*, 5, 1994)

Forthcoming volumes:

PERCEPTION
(*Philosophical Issues*, 7, 1996)

Philosophical Issues, 6, 1995

CONTENT

edited by

Enrique Villanueva

Sociedad Filosófica Ibero Americana

Ridgeview Publishing Company • Atascadero, California

Paper text: ISBN 0-924922-22-2
Cloth (library edition): ISBN 0-924922-72-9

The typesetting was done by José Luis Olivares.

Published in the United States of America
by Ridgeview Publishing Company
Box 686
Atascadero, California 93423

Printed in the United States of America
by Thomson-Shore, Inc.

Contents

Preface
Enrique Villanueva vii

1 Concepts: A Potboiler
Jerry Fodor 1

2 Fodor's Concepts
James Higginbotham 25

3 On What It Is Like to Grasp a Concept
Joe Levine 38

4 Can Semantic Properties Be Non–causal?
Pierre Jacob 44

5 Reference from the First Person Perspective
Brian Loar 53

6 Disquotation and Cause in the Theory of Reference
Paul Horwich 73

7 Reference from a Perspective *versus* Reference
David Sosa 79

8 Fregean Reference Defended
Ernest Sosa 91

9 On Sosa's "Fregean Reference Defended"
William Lycan 100

10 Doubts about Fregean Reference
Manuel García–Carpinteiro 104

11 More on Fregean Reference
Ernest Sosa 113

12 Mental Causation: What? Me Worry?
Jaegwon Kim 123

13 Mental Causation: A Query for Kim
James Tomberlin 152

14 I'm a Mother, I Worry
Louise Antony 160

15 Kim on the Exclusion Problem
Manuel Campos 167

16 Ruritania Revisited
Ned Block 171

17 Ruritania and Ecology
Josefa Toribio 188

18 Can There Be a Rationally Compelling Argument for Anti-realism about Ordinary ("Folk") Psychology?
Crispin Wright 197

19 A Note on Boghossian's Master Argument
Roger Gibson 222

20 Content, Computation and Externalism
Christopher Peacocke 227

21 Can We Knock Off the Shackles of Syntax?
Daniel Andler 265

22 Content Preservation
Tyler Burge 271

23 Testimony and *a priori* Knowledge
John Biro 301

24 Can Peter Be Rational?
Lourdes Valdivia 311

Contributors 325

Preface

The papers in this volume are concerned with issues on content, concepts, holism, reference, *a priori* knowledge, computation, externalism, mental causation, realism and other subtopics. Those issues revolve on the frontiers of Metaphysics, Philosophy of Mind and Language and Epistemology which have become prevalent in the volumes that form this *Philosophical Issues* series. It is good to remember one of the main aims of our Society in organizing the conferences that give birth to these volumes, namely, to bring the main philosophers of our time to discuss the central issues of Philosophy.

Most of the papers in this volume were presented at the Seventh Annual Philosophy Conference of our Sociedad Filosófica Ibero Americana, SOFIA, held in Lisbon, Portugal, May 22-24. In our Society's name I want to express our gratitude to Professor Joao Paulo Monteiro who co-organized for a second time a conference of ours under the sponsorship of *Lisbon 94 Capital Cultural of Europe*, *Luso-American Foundation for Development*, *National Council for Scientific and Technological Research*, *Department of Philosophy of the University of Lisbon* and under the High Auspices of His Excellency the President of the Republic of Portugal, Mario Soares. It was again a delightful experience to work with Joao Paulo both before and during that Conference.

I want to express my gratitude both to Ernie Sosa and Jim Tomberlin for their generous assistance through the organization of this conference.

Once more I want to express my gratitude to my university, the Universidad Nacional Autónoma de México who provided a sabbatical leave and support through Dirección de Asuntos del Personal Académico. To Mtro. Antonio Gago who renewed support for our Society (through convenio 94-01-09-160-212) at the Subsecretaría de Educación Superior e Investigación Científica, SEP. CONACYT (proyecto 0754-H9110) also provided support.

For a second year I am glad to thank the Ph D Program in Philosophy at the Graduate Center at CUNY, Chair Mr. Richard Mendelson who generously provided a room from where I kept organizing this Conference and later getting the papers for publication. Thanks also to Mr Peter Klein, Chair at the Department of Philosophy at Rutgers, for renewing my status as a visiting fellow.

Thanks again to José Luis Olivares who did the typesetting.

Enrique Villanueva
Highland Park, March 1995

PHILOSOPHICAL ISSUES, 6
Content, 1995

Concepts; A Potboiler

Jerry Fodor

Introduction: The centrality of concepts

What's ubiquitous goes unremarked; nobody listens to the music of the spheres (or to me, for that matter). I think a certain account of concepts is ubiquitous in recent discussions about minds; not just in philosophy but also in psychology, linguistics, AI, and the rest of the cognitive sciences; and not just this week, but for the last fifty years or so. And I think this ubiquitous theory is quite probably untrue. This paper aims at consciousness raising; I want to get you to see that there is this ubiquitous theory and that, very likely, you yourself are among its adherents. What to do about the theory's not being true (if it's not) —what our cognitive science would be like if we were to throw the theory overboard— is a long, hard question, and one that I'll mostly leave for another time.

The nature of concepts is the pivotal theoretical issue in cognitive science; it's the one that all the others turn on. Here's why:

Cognitive science is fundamentally concerned with a *certain mind–world* relation; the goal is to understand how its mental processes can cause a creature to behave in ways which, in normal circumstances, reliably comport with its utilities. There is, at present, almost[1] uni-

[1]The caveat is because it's moot how one should understand the relation between main–line cognitive science and the Gibsonian tradition. For discussion, see Fodor and Pylyshyn (1981).

versal agreement that theories of this relation must posit mental states some of whose properties are *representational*, and some of whose properties are *causal*. The representational (or, as I'll often say, *semantic*) properties of a creature's mental states are supposed to be sensitive to, and hence to carry information about, the character of its environment.[2] The causal properties of a creature's mental states are supposed to determine the course of its mental processes, and, eventually, the character of its behavior. Mental entities that exhibit both semantic and causal properties are generically called 'mental representations', and theories that propose to account for the adaptivity of behavior by reference to the semantic and causal properties of mental representations are called 'representational theories of the mind'.

Enter concepts. Concepts are the least complex mental entities that exhibit both representational and causal properties; all the others (including, particularly, beliefs, desires and the rest of the 'propositional attitudes') are assumed to be *complexes* whose constituents are concepts, and whose representational and causal properties are determined, wholly or in part, by those of the concepts that they're constructed from.

This account subsumes even the Connectionist tradition which is, however, often unclear, or confused, or both about whether and in what sense it is committed to *complex* mental representations. There is a substantial literature on this issue, provoked by Fodor and Pylyshyn (1988). (See, for example, Smolensky, (1988); Fodor and McLaughlin (1990) etc.) Suffice it for present purposes that connectionists clearly assume that there are *elementary* mental representations (typically labelled nodes), and that these have both semantic and causal properties. Roughly, the semantic properties of a node in a network are specified by the node's label, and its causal properties are determined by the character of its connectivity. So even connectionists think there are concepts as the present discussion understands that notion.

On all hands, then, concepts serve both as the domains over which the most elementary mental processes are defined, and as the most

---

[2]There is no general agreement, either in cognitive science or in philosophy, about how the representational/semantic properties of mental states are to be analyzed; they are, in general, simply taken for granted by psychologists when empirical theories of cognitive processes are proposed. This paper will *not* be concerned, other than tangentially, with these issues in the metaphysical foundations of semantics. For recent discussion, however, see Fodor (1990) and references cited there.

primitive bearers of semantic properties. Hence their centrality in representational theories of mind.

1 Ancient History: The Classical Background

The kind of concept–centered psychological theory I've just been sketching should seem familiar, not only from current work in cognitive science, but also from the philosophical tradition of Classical British Empiricism. I want to say a bit about Classical versions of the representational theory of mind because, though their general architecture conforms quite closely to what I've just outlined, the account of concepts that they offered differs, in striking ways, from the ones that are now fashionable. Comparison illuminates both the Classical and the Current kinds of representational theories, and reveals important respects in which the older story was closer to being right about the nature of concepts than ours is. So, anyhow, I am going to argue.

Here's a stripped–down version of a Classical Representational theory of concepts:

Concepts are mental images. They get their causal powers from their associative relations to one another, and they get their semantic properties from their resemblance to things in the world. So, for example: The concept DOG applies to dogs because dogs are what (tokens of) the concept looks like. Thinking about dogs often makes one think about cats because dogs and cats often turn up together in experience, and it's the patterns in one's experience, and only these, that determine the associations among one's Ideas. Because association is the only causal power that Ideas have, and because association is determined only by experience, any Idea can, in principle, become associated to any other, depending on which experiences one happens to have had. *Classical Ideas cannot, therefore be defined by their relations to one another.* Though DOG–thoughts call up CAT–thoughts, LEASH–thoughts, BONE–thoughts, BARK–thoughts and the like in most *actual* mental lives, there are *possible* mental lives in which that very same concept reliably calls up, as it might be, PRIME NUMBER–thoughts or TUESDAY AFTERNOON–thoughts or KETCHUP–thoughts. It depends entirely on how often you've come across prime numbers of dogs covered with ketchup on Tuesday afternoons.

So much by way of a reminder of what Classical theorists said about concepts. I don't want to claim much for the historical accuracy of my exegesis (though it may be that Hume held a view within

hailing distance of the one I've sketched; for purposes of exposition, I'll assume he did). But I do want to call your attention to a certain point about the *tactics* of this kind of theory construction; a point that's essential but easy to overlook.

Generally speaking, if you know *what an X is*, then you also know *what it is to have an X*. And ditto the other way around. No doubt, this applies to concepts. If, for example, your theory is that concepts are pumpkins, then it has to be a part of your theory that having a concept is having a pumpkin; and if your theory is that having a concept is having a pumpkin, then it has to be a part of your theory that pumpkins are what concepts are. I take it that this is just truistic.

Sometimes it's clear in which direction the explanation should go, and sometimes it isn't. So, for example, one's theory about *having* a cat ought surely to be parasitic on one's theory about *being* a cat; first you say what a cat is, and then you say that having a cat is just: *having one of those.* With jobs, pains, and siblings, however, it goes the other way 'round. First you say what it is to *have* a job, or a pain, or a sibling, and then the story about what jobs, pains and siblings *are* is a spin off.

These examples are, I hope, untendentious. But decisions about the proper order of explanation can be unobvious, important, and extremely difficult. To cite a notorious case: ought one first explain what the number three is and then explain what it is for a set to have three members? Or do you first explain what sets are, and then explain what numbers are in terms of them? Or are the properties of sets and of numbers both parasitic on those of something quite else, (like counting, for example). If I knew, I would be I would be rich and famous; well, famous.

Anyhow, Classical representational theorists uniformly took it 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 '*identity conditions*' for the concept—, and then the story about concept *possession* follows accordingly. Fine, but *how* do you identify a concept if not by giving its possession conditions? Answer: You identify a concept by saying *what it is the concept of.* The concept DOG, for example, is the concept *of dogs*; that's to say, it's the concept that you use to think about dogs with. Correspondingly, *having* the concept DOG is just *having a concept to think about dogs with.*

Similarly, of course, for concepts of other than canine content: What individuates the concept *X* is that it's the concept *of Xs.* Possession conditions are parasitic; *having* the concept *X* is just

having a concept to think about Xs with. (More precisely, it's having a concept to think about *X*s 'as such' with. The context 'thinks about...' is intentional for the '...' position. We'll return to this presently.)

So much for the explanatory tactics of Classical Representational Theories of Mind. Without exception, however, Current theorizing about concepts reverses the Classical direction of analysis. The substance of Current theories lies in what they say about the *possession conditions* for concepts. It's the story about concept *individuation* that they treat as parasitic: The concept *X* is just *whatever it is that a creature has* when it satisfies the possession conditions for that concept. (See, for example, Peacocke (1992), which is illuminatingly explicit on this point.) This subtle, and largely inarticulate, difference between Contemporary representational theories and their Classical forebears has had, so I'll argue, the most profound implications for our cognitive science. To a striking extent, it determines the kinds of problems we work on and the kinds of theories that we offer as solutions to our problems. I fear that it was a wrong turn —on balance, a catastrophe— and that we shall have to go back and do it all again.

First, however, just a little about why the Classical Representational view was abandoned. There were, I think, three kinds of reasons: methodological, metaphysical and epistemological. We'll need to keep them all in mind when we turn to discussing Current accounts of concepts.

Methodology: Suppose you're a behaviorist of the kind who thinks there are no concepts. In that case, you will feel no need for a theory about what concepts are, Classical or otherwise. Behaviorist views aren't widely prevalent now, but they used to be; one of the things that killed the Classical theory of concepts was simply that concepts are mental entities,[3] and mentalism went out of fashion.

Metaphysics: A Classical theory individuates concepts by specifying their contents; the concept *X* is the concept *of Xs*. This seemed ok —it seemed not to beg any principled questions— because Classical theorists thought that they had *of–ness* under control; they thought that the image theory of mental representation explained it. It's now clear that they were wrong to think this. Even if con-

---

[3]Terminological footnote: Here and elsewhere in this paper, I follow the psychologist's usage rather than the philosopher's; for philosophers, concepts are generally *abstract* entities, hence, of course, *not* mental. The two ways of talking are compatible. The philosopher's concepts can be viewed as the types of which the psychologist's concepts are tokens.

cepts are mental images (which they aren't) and even if the concept DOG looks like a dog (which it doesn't) still, it isn't *because* it looks like a dog that it's the concept *of dogs*. *Of*–ness ('content', 'intentionality') does not reduce to resemblance, and it is now widely, and rightly, viewed as problematic. It doesn't follow either that Classical theorists were wrong to hold that the story about concept possession should be parasitic on the story about concept identification, or that they were wrong to hold that concepts should be individuated by their contents. But it's true that if you want to defend the Classical order of analysis, you need an alternative to the picture theory of meaning.

Epistemology: The third of the standard objections to the Classical account of concepts, though at least as influential as the others, is distinctly harder to state. Roughly, it's that Classical theories aren't adequately 'ecological'. Used in this connection, the term has a Gibsonian ring; but I'm meaning it to pick out a much broader critical tradition. (In fact, I suspect Dewey was the chief influence; see the next footnote). Here's a rough formulation:

What cognitive science is trying to understand is something that happens *in the world*; it's the interplay of environmental contingencies and behavioral adaptations. Viewing concepts primarily as the vehicles of *thought* puts the locus of this mind/world interaction (metaphorically and maybe literally) not in the world but in the head. Having put it in there, Classical theorists are at a loss for how to get it out again. So the ecological objection goes.

This kind of worry comes in many variants, the epistemological being, perhaps, the most familiar. If concepts are internal mental representations, and thought is conversant only with concepts, how does thought ever contact the external world that the mental representations are supposed to represent? If there is a 'Veil of Ideas' between the mind and the world, how can the mind see the world through the veil? Isn't it, in fact, inevitable that the Classical style of theorizing eventuates either in solipsism ('thought never does connect with the world, only with our Idea of it') or in Idealism ('it's ok if our thoughts never get out of our heads because the world is in there with them')?[4] And, surely, solipsism and idealism are both refutations of theories that entail them.

---

[4] "Experience to them is not only something extraneous which is occasionally superimposed upon nature, but it forms a veil or screen which shuts us off from nature, unless in some way it can be 'transcended'(p. 1a)" "Other [philosophers' methods] begin with results of a reflection that has already torn in two the subject–matter and the operations and states of experiencing. The problem is then to get together again what has been sundered..."(p. 9). Thus Dewey (1958).

Notice that this ecological criticism of the Classical story is different from the behaviorist's eschewal of intentionality as such. The present objection to 'internal representations' is not that they are representations, but that they are internal. (In fact, this sort of objection to the Classical theory predates behaviorism by a lot. Reid used it against Hume, for example.) Notice too that this objection *survives* the demise of the image theory of concepts; treating mental representation as, say, discursive rather than iconic, doesn't help. What's wanted isn't either pictures of the world *or* stories about the world; what's wanted is what they call in Europe *being in* the world. (I'm told this sounds even better in German.)

This is all, as I say, hard to formulate precisely; I think, in fact, that it is extremely confused. But even if the 'ecological' diagnosis of what's wrong with Classical concepts is a bit obscure, it's clear enough what cure was recommended, and this brings us back to our main topic. If what we want is to get thought out of the head and into the world, we need to reverse the Classical direction of analysis, precisely as discussed above; we need to take *having a concept* as the fundamental notion and define concept individuation in terms of it. This is a true Copernican revolution in the theory of mind, and we are still living among the debris.

Here, in roughest outline, is the new theory about concept possession: *Having a concept is having certain epistemic capacities.* To have the concept *of* X is to be able to recognize Xs, and/or to be able to reason about Xs in certain kinds of ways. (Compare the Classical view discussed above: Having the concept *of* X is just being able to have thoughts *about* X*s*). It is a paradigmatically Pragmatist idea that having a concept is being able to *do* certain things rather than being able to *think* certain things. Accordingly, in the discussion that follows, I will contrast Classical theories of concepts with 'Pragmatic' ones. I'll try to make it plausible that all the recent and current accounts of concepts in cognitive science really are just variations on the Pragmatist legacy.

In particular, I propose to consider (briefly, you'll be pleased to hear) what I take to be five failed versions of pragmatism about concepts. Each evokes its proprietary nemesis; there is, for each, a deep fact about concepts by which it is undone. The resulting symmetry is gratifyingly Sophoclean. When we've finished with this

---

The remedy he recommends is resolutely to refuse to recognize the distinction between experience and its object. "[Experience] recognizes in its primary integrity no division between act and material, subject and object, but contains them both in an unanalyzed totality".

catalogue of tragic flaws, we'll have exhausted all the versions of concept pragmatism I've heard of, or can think of, and we'll also have compiled a must–list for whatever theory of concepts Pragmatism is eventually replaced by.

1.1 BEHAVIORISTIC PRAGMATISM; (AND THE PROBLEM OF INTENTIONALITY)

I remarked above that behaviorism can be a reason for ruling all mentalistic notions out of psychology, concepts included. However, not all behaviorists were eliminativists; some were reductionists instead. Thus Ryle, and Hull (and even Skinner about half the time) are perfectly content to talk of concept–possession, so long as the 'criteria' for having a concept can be expressed in the vocabulary of behavior and/or in the vocabulary of dispositions to behave.

Do not ask what criteria are; there are some things we're not meant to know. Suffice it that criterial relations are supposed to be sort–of–semantical rather than sort–of–empirical.

So, then, *which* behaviors are supposed to be criterial for concept possession? Short answer: sorting behaviors. Au fond, according to this tradition, having the concept X is being able to discriminate Xs from non–Xs; it's being able to sort things into the ones that are X and the ones that aren't. This approach gets concepts into the world with a vengeance: having a concept is responding selectively, or being disposed to respond selectively, to the things in the world that the concept applies to; and paradigmatic responses are overt behaviors 'under the control' of overt stimulations. (Though behaviorist in essence, this identification of possessing a concept with being able to discriminate the things it applies to survived well into the age of computer models; see, for example, 'procedural' semanticists like Woods, (1975). Lots of philosophers *still* think there must be *something* to it; see, for example, Peacocke (*op. cit.*).)

I don't want to bore you with ancient recent history, and I do want to turn to less primitive versions of Pragmatism about concepts. So let me just briefly remind you of what proved to be the decisive argument against the behavioristic version: Concepts can't be *just* sorting capacities, for if they were, then coextensive concepts —concepts that apply to the same things— would have to be identical. And coextensive concepts aren't, in general, identical. Even *necessarily* coextensive concepts —like TRIANGULAR and TRILATERAL, for example— may perfectly well be distinct.

To put this point another way, sorting is something that happens *under a description*; it's always relative to some or other way of conceptualizing the things that are being sorted. Though their behaviors may *look* exactly the same, and though they may end up with the very same things in their piles, the creature that is sorting triangles is in a different mental state, and is behaving in a different way, from the creature that is sorting trilaterals; and only the first is exercising the concept TRIANGLE. (For a clear statement of this objection, see Dennett, 1978).[5]

Behaviorists had a bad case of mauvais fois about this; they would dearly have liked to deny the intentionality of sorting outright. In this respect, articles like Kendler (1952), according to which 'what is learned, [is] a pseudoproblem in psychology' make fascinating retrospective reading. Suppose, however, that you accept the point that sorting is always relative to a concept, but you wish, nonetheless, to cleave to some kind of pragmatist reduction of concept individuation to concept possession and of concept possession to having epistemic capacities. The question then arises: *what difference in their epistemic capacities* could distinguish the creature that is sorting triangles from the creature that is sorting trilaterals? What could the difference between them be, if it isn't in the piles that they end up with?

The universally popular answer has been that the difference between *sorting under the concept TRIANGLE* and *sorting under the concept TRILATERAL* lies in *what the sorter is disposed infer* from the sorting he performs. To think of something as a *triangle* is to think of it *as having angles*, to think of something as a *trilateral* is to think of it as *having sides*. The guy who is collecting triangles must therefore accept *that the things in his collection have angles* (whether or not he has noticed that they have sides); and the guy who is collecting trilaterals must accept that the things in his collection have sides (even if he hasn't noticed that they have angles).

---

[5]I'm sometimes asked whether informational semantics —a view that I'm enthusiastic for— isn't itself a version of the *concept possession = sorting capacity* story; and, in fact, I have myself stressed the similarities between Skinnerian and Dreteskian approaches to naturalizing content (see TOC). There is, however, this crucial difference: whereas sorting is an intentional notion, 'carries information about —' is extensional at the '—' position. By taking sorting for granted, Skinnerian semantics begs the problems intentionality from the outset. Whatever may be said against informational semantics, at least it doesn't do that.

I'm grateful to Georges Rey for having raised this issue (in conversation).

The long and short is: having concepts is having a mixture of *abilities to sort* and *abilities to infer.*[6] Since inferring is presumably neither a behavior nor a behavioral capacity, this formulation is, of course, not one that a *behavioristic* pragmatist can swallow. So much the worse for behaviorists, as usual. But notice that pragmatists as such are still ok: even if having a concept isn't just knowing how to sort things, it still may be that having a concept is *some* kind of knowing how, and that theories of concept possession are prior to theories of concept individuation.

We are now getting very close to the Current scene. All nonbehaviorist versions of Pragmatism hold that for concept possession is constituted, at least in part, by inferential dispositions and capacities. They are thus all required to decide *which inferences constitute which concepts.* Contemporary theories of concepts, though without exception pragmatist, are distinguished by the ways that they approach this question. Of nonbehavioristic pragmatist theories of concepts there are, by my reckoning, exactly four. Of which the first is:

1.2 ANARCHIC PRAGMATISM (AND THE REALISM PROBLEM)

Anarchic pragmatism is the doctrine that though concepts are constituted by inferential dispositions and capacities, there is no fact of the matter about which inferences constitute which concepts. This is a very laid–back point of view and California is, of course, its *locus classicus.* But no doubt there are those even on the East Coast who believe it in their hearts.

I'm not going to discuss anarchic pragmatism. If there are no facts about which inferences constitute which concepts, then there are no facts about which concepts are which. And if there are no facts about which concepts are which, then there are no facts about which beliefs and desires are which (for, by assumption, beliefs and desires

---

[6]The idea that concepts are (at least partially) constituted by inferential capacities receives what seems to be independent support from the success of logicist treatments of the 'logical' concepts (AND, ALL, etc.). For many philosophers (though not for many psychologists) thinking of concepts as inferential capacities is a natural way of extending the logicist program from the logical vocabulary to TREE or TABLE. So, when these philosophers tell you what it's like to analyze a concept, they always like to start with AND. (Here again, Peacocke (1992) is paradigmatic).

It should, however, strike you as *not obvious* that the analysis of AND is a plausible model for the analysis of TREE or TABLE.

are complexes of which concepts are the constituents). And if there are no facts about which beliefs and desires are which, there is no intentional cognitive science, for cognitive science is just belief/desire explanation made systematic. And if there is no cognitive science, we might as well stop worrying about what concepts are and have a nice long soak in a nice warm hot tub instead.

I'm also not going to consider a doctrine that is closely related to anarchic pragmatism: Namely, that while nothing systematic can be said about concept *identity*, it may be possible to provide a precise account of when, and to what degree, and in what respects, two concepts are *similar*. Some such thought is often voiced informally in the cognitive science literature, but there is, to my knowledge, not even a rough account of how such a similarity relation over concepts might be defined. I strongly suspect this is because a robust notion of similarity is possible only where there is a correspondingly robust notion of identity. For a discussion, see Fodor and Lepore (1992), Chapter 7.

1.3 Definitional pragmatism (and the analyticity problem)

Suppose the English word 'bachelor' means the same as the English phrase 'unmarried male'. Synonymous terms presumably express the same concept (this is a main connection between theories about concepts and theories about language), so it follows that you couldn't have the concept BACHELOR and fail to have the concept UNMARRIED MALE. And from that, together with the intentionality of sorting (see 2.1), it follows that you couldn't be collecting bachelors *so described* unless you take yourself to be collecting unmarried males; i.e. unless you accept the inference that if something belongs in your bachelor collection, then it is something that is male and unmarried.

Maybe this treatment generalizes; maybe, having the concept X *just is* being able to sort Xs *and being disposed to draw the inferences that define* X*–ness.*

The idea that it's the disposition to draw *defining* inferences that counts for concept possession is now almost as unfashionable as behaviorism. Still, the departed deserves a word or two of praise. For one thing, the definition story offered a plausible (though partial) account of the *acquisition* of concepts. If the concept BACHELOR *is* the concept UNMARRIED MALE, then it's not hard to imagine how a creature that has the concept UNMARRIED and has the concept MALE could put them together and thereby achieve the

concept BACHELOR. (Of course the theory that complex concepts are acquired by constructing them from their elements *presupposes* the availability of the elements. About the acquisition of these, definitional pragmatism tended to be hazy.) This processes of assembling concepts can be —indeed, was— studied in the laboratory; cf. Bruner, Goodnow and Austin (1956) and the large experimental literature that it inspired.

Other significant virtues of the definition story will suggest themselves presently, when we discuss concepts as prototypes. But alas, despite its advantages, the definition theory didn't work. Concepts can't *be* definitions because most concepts don't *have* definitions. At a minimum, to define a concept is to provide necessary and sufficient conditions for something to be in its extension (that is, for being among the things that the concept applies to). As it turns out, for most concepts, this condition can't be met. Maybe being male and unmarried is necessary and sufficient for being a bachelor; but try actually filling in the blanks in 'x is a dog iff x is a . . . ' without using words like 'dog' or 'canine' or the like on the right hand side.

To be sure, you might wonder just what's wrong with using words like 'dog' or 'canine' on the right hand side. Unlike so much else in what linguists call lexical semantics, ''dog' means *dog* has, at a minimum, the advantage of being *true*'. Correspondingly, if you're a psychologist wondering what 'features' of dogs are reliable enough to construct the concept DOG from, *being dogs* does rather suggest itself. 'Dogs bark' and 'dogs have tails' will let you down now and then; but 'dogs are dogs' never will. I think it's remarkable —one of the all time curiousities of our intellectual tradition— that the force of these truisms isn't widely appreciated. Take them seriously, and you're on the road to conceptual atomism. I think that's the right road to be on, but nevermind; that's matter for a different paper.

Short of defining 'dog' as *dog* there might be one other way to provide necessary and sufficient condtitions for applying it: if you could make a *list* of all and only the dogs (Rover, Lassie, Spot,... etc.), then a thing's being on the list would be necessary and sufficient for 'dog' to be true of it. That there is this option is, however, no comfort for the theory that concepts are definitions. Rather, what it shows is that being a necessary and sufficient condition for the application of a concept is not a sufficient condition for being a definition of the concept.

This point generalizes beyond the case of lists. *Being a creature with a backbone* is necessary and sufficient for *being a creature with a heart* (so they tell me). But it isn't the case that 'creature with a backbone' defines 'creature with a heart' or vice versa. Quite

generally, it seems that Y doesn't define X unless Y applies to all and only the *possible* Xs (as well, of course, as all and only the *actual* Xs). It is, then, the modal notion —possibility— that's at the heart of the idea that concepts are definitions. Correspondingly, what killed the definition theory of concepts in philosophy, is that nobody was able to explicate the relevant sense of 'possible'.

I suppose it's not required, for this audience, to rehearse the vicissitudes of philosophical attempts to construe this notion. But it's important to bear in mind the extent to which the catastrophe was interdisciplinary. What killed the definition theory in cognitive psychology, for example, wasn't scruples about modality. It was that that the experiments that were supposed to warrant the identification of concepts with definitions uniformly didn't work. Definitions turn out to contribute vanishingly little to explaining what subjects do in tasks that involve applying concepts to things that fall under them. To put it another way: The cases that are supposed to possess the special kind of modality that definitional relations are supposed to engender *don't seem to be a natural kind*; they don't, as it were, crop up anywhere else in cognitive science.

Actually, the number and variety of psychological phenomena that definition theories do *not* predict, and the reliability with which they do not predict them, is about as impressive as anything known in cognitive science. Definitions don't predict the relative accessibility of concepts or the relative difficulty of concept application tasks; they don't predict the order of acquisition of concepts, or of the words that the concepts are expressed by; they don't predict the demands entertaining a concept make on memory or attention; they don't predict relations of conceptual inter–facilitation in priming or related tasks. In fact, to my knowledge, there is *no* experimental environment that distinguishes the consequences of definitional connections from those of association or empirical centrality; not even in what are supposed to be the clearest cases of definable concepts.

If you want to predict the effects of definitional structure in psychological tasks, the rule is simple and reliable: *It always acts exactly as though it wasn't there.* Such friends as definitions still have spend a lot of time trying to explain away these empirical failures. I suspect special pleading.

It's often (and rightly) said that nobody has *proved* that there can't be a viable analytic/synthetic distinction; and that, barring such a proof, there's still hope for such related constructs as definitions, conceptual necessities, criteria and the like. With research strategies, as with horse races, you pays your money and you makes your bets; and, no doubt, you can sometimes get rich betting against

a run. The rest of this paper is for those who want a robust notion of *concept* but doubt that appeal to *conceptual necessity* will buy it for them. Or who are willing to suspend their disbelief.

1.4 STEREOTYPES AND PROTOTYPES (AND THE PROBLEM OF COMPOSITIONALITY)

Because it was pragmatist, the definition story treated having a concept as having a bundle of inferential capacities, and was faced with the usual problem about which inferences belong to which bundles. The notion of a *defining* inference was supposed to bear the burden of answering this question, and the project foundered because nobody knows what makes an inference defining, and nobody has any idea how to find out. 'Well', an exasperated pragmatist might nonetheless reply, 'even if I don't know what makes an inference definitional, I do know what makes one statistically reliable. So why couldn't the theory of concept possession be statistical rather than semantic? Why couldn't I exploit the notion of a *reliable* inference to do what definitional pragmatism tried and failed to do with the notion of an *analytic* inference?'

We arrive, at last, at Modern Times. For lots of kinds of Xs, people are in striking agreement about what properties an arbitrarily chosen X is likely to have. (An arbitrarily chosen bird is likely to be able to fly; an arbitrarily chosen conservative is likely to be a Republican; an arbitrarily chosen dog is likely to be less than a light year long.) Moreover, for lots of kinds of Xs, people are in striking agreement about which Xs are prototypic of the kind (diamonds for jewels; red for colors; not dachshunds for dogs). And, sure enough, the Xs that are judged to be prototypical are generally ones that have lots of the properties that an arbitrary X is judged likely to have; and the Xs that are judged to have lots of the properties that an arbitrary X is likely to have are generally the ones that are judged to be prototypical.

Notice, in passing, that the theory that concepts are stereotypes shares one of the most agreeable features of the theory that concepts are definitions: it makes the learning of (complex) concepts intelligible. If the concept of an X is the concept of something that is reliably Y and Z, then you can learn the concept X if you have the concepts Y and Z together with enough statistics to recognize reliability when you see it. It would be ok, for this purpose, if the available statistical procedures were analogically (rather than explicitly) represented in the learner. Qua learning models, 'neural networks' are analog computers of statistical dependencies, so it's hardly surprising that pro-

totype theories of concepts are popular among connectionists (See, for example, McClelland and Rummelhart, (1986).

So, then, why shouldn't having the concept of an X be having the ability to sort by Xness, together with a disposition to infer from something's being X to its having the typical properties of Xs? I think, in fact, that this is probably the view of concepts that the prototypical cognitive scientist holds these days.

To see why it doesn't work, let's return one last time to the defunct idea that concepts are definitions. It was a virtue of that idea that it provides for the *compositionality* of concepts, and hence for the productivity and systematicity of thought. This, we're about to see, is no small matter.

In the first instance, productivity and systematicity are best illustrated by reference to features (not of minds but) of natural languages. To say that languages are productive is to say that there is no upper bound to the number of well–formed formulas that they contain. To say that they are systematic is to say that if a language can express the proposition P, then it will be able to express a variety of other propositions that are, in one way or another, semantically related to P. (So, if a language can say that P and that $-Q$, it will also be able to say that Q and that $-P$; if it can say that John loves Mary, it will be able to say that Mary loves John... and so forth.) As far as anybody knows, productivity and systematicity are universal features of human languages.

Productivity and systematicity are also universal features of human *thought* (and, for all I know, of the thoughts of many infra–human creatures). Assuming the usual distinction between cognitive 'competence' and cognitive 'performance', there is no upper bound to the number of thoughts that a person can entertain. And likewise, if a mind can think the thought that P and any negative thoughts, it can also think the thought that $-P$; if it can think the thought that Mary loves John, it can think the thought that John loves Mary... etc.

It is extremely plausible that the productivity and the systematicity of language and thought are both to be explained by appeal to the systematicity and productivity of mental representations, and that mental representations are systematic and productive because they are compositional. The idea is that mental representations are constructed by the application of a finite number of combinatorial principles to a finite basis of (relatively or absolutely) primitive concepts. (So, for example, the very same construction that gets you the concept MISSILE from the concept ANTIMISSILE, also gets you the concept ANTIMISSILE from the concept ANTIANTIMISSILE.)

Productivity follows because the application of these constructive principles can iterate without bound. Systematicity follows because the concepts and principles you need to construct the thoughts that P and $-Q$ are the very same ones that you need to construct the thoughts that Q and $-P$; and the concepts and principles you need to construct the thought that John loves Mary are the very same ones that you need to construct the thought that Mary loves John.

This sort of treatment of compositionality is familiar, and I will assume that it is essentially correct. I want to emphasize that it places a heavy constraint on both theories of concept possession and theories of concept individuation. If you accept compositionality, then you are required to say that whatever the concept DOG is that occurs in the thought that *Rover is a dog*, that *very same* concept DOG also occurs in the thought that *Rover is a brown dog*; and, whatever the concept BROWN is that occurs in the thought that *Rover is brown*, the very same concept BROWN also occurs in the thought that *Rover is a brown dog*. Compositionality requires, in effect, that constituent concepts must be *insensitive* to their host; a constituent concept contributes the same content to all the complex representations it occurs in. It's on these assumptions that compositionality explains how being able to think that Rover is brown and that Rover is a dog is sufficient for being able to think that Rover is a brown dog.

And compositionality further requires that the content of a complex representation is *exhausted* by the contributions that its constituents make. Whatever the content of the concept of BROWN DOG may be, it must be completely determined by the content of the constituent concepts BROWN and DOG, together with the combinatorial apparatus that sticks these constituents together; if this were not the case, your grasp of the concepts BROWN and DOG wouldn't explain your grasp of the concept BROWN DOG. In short, when complex concepts are compositional, the whole must *not* be more than the sum of its parts, otherwise compositionality won't explain productivity and systematicity. And if compositionality doesn't, nothing will. If the constraints that the compositionality of concepts imposes on the notion of conceptual content strike you as a bit austere, it may be some comfort that systematicity and productivity are compatible with compositionality failing in any finite number of cases. They allow, for example, that finitely many thoughts (hence, a fortiori, finitely many of the linguistic expressions used to express them) are idiomatic or metaphoric, *so long as there are infinitely many that are neither.*

We can now see why, though concepts might have turned out to

be definitions, they couldn't possibly turn out to be stereotypes or prototypes: Given the constraints compositionality imposes, there are no candidates for identification with concepts *except* defining properties.

Concepts contribute their *defining* properties to the complexes of which they are constituents, and the *defining* properties of complex concepts are exhaustively determined by the defining properties that their constituents contribute. Since bachelors are, by definition, unmarried men, tall bachelors are, by the same definition, tall unmarried men; and very tall bachelors are very tall unmarried men, and very tall bachelors from Hoboken are very tall unmarried men from Hoboken... and so on. Correspondingly, there is nothing more to the *definition* of 'very tall bachelor from Hoboken' than *very tall unmarried man from Hoboken*; that is, there is nothing more to the definition of the phrase than what the definitions of its constituents contribute.

So, then, if concepts were definitions, we could see how thought could be compositional, and hence productive and systematic. Concepts aren't definitions, of course. It's just that, from the present perspective, it's rather a pity that they're not.

For, stereotypes, alas, don't work the way that definitions do. Stereotypes aren't compositional. For one thing, 'ADJECTIVE *X*' can be a perfectly good concept even if there is no *adjective X* stereotype. And even if there are stereotypic *adjective X*s, they don't have to be either stereotypic *adjectives* or stereotypic *X*s. I doubt, for example, that there is a stereotype of very tall men from Hoboken; but, even if there were, there is no reason to suppose that it would approximate either a the stereotype for tall men, or the stereotype for men from Hoboken, or the stereotype for men. To the contrary: often enough, the adjective in 'ADJECTIVE *X*' is there precisely to mark a way that adjective *X*s *depart* from stereotypic *X*s. Fitzgerald made this point about stereotypes to Hemingway when he said, 'The rich are different from the rest of us'. Hemingway replied by making the corresponding point about definitions; 'yes', he said, 'they have more money'.

In fact, this observation about the uncompositionality of stereotypes generalizes in a way that seems to me badly to undermine the whole pragmatist program of identifying concept possession with inferential dispositions. I've claimed that knowing what is typical of things that are Adjective and what is typical of things that are *X* doesn't, in the general case, tell you what is typical of things that are Adjective *X*s. The reason it doesn't is perfectly clear; though some of your beliefs about Adjective *X*s are compositionally inher-

ited from your beliefs about Adjectives, and some are compositionally inherited from your beliefs about Xs, *some are beliefs that you have acquired about Adjective Xs as such*, and these aren't compositional at all.

The same applies, of course, to the inferences that your beliefs about Adjective Xs dispose you to draw. Some of the inferences you are prepared to make about green apples follow just from their being green and from their being apples. That is to say: they derive just from the constituency and structure of your GREEN APPLE concept. But others depend on information (or misinformation) that you have picked up about green apples as such: that green apples go well in apple pie; that they are likely to taste sour; that there are kinds of green apples that you'd best not eat uncooked, and so forth. Patently, these inferences are not definitional; they belong to what you know about green apples, not to what you know about the corresponding concepts. And, since they aren't definitional, they aren't compositional either; they're among the GREEN APPLE inferences that *aren't* inherited either from GREEN inferences or from APPLE inferences.

Similarly, mutatis mutandis, for the words that express these concepts: You learned that "green apple" means *green and apple* when you learned English at your mother's knee. But probably you learned that green apples mean apple pies from the likes of Julia Child.

Let me pause to make this line of argument perspicuous. If, as I'm supposing, it really is their compositionality that explains why concepts are systematic and productive, then whatever about concepts is *not* compositional determined is ipso facto not their content. But, as we've just been seeing, inferential roles aren't compositional, except when the inferences are definitional. So concepts can't be inferential roles unless there are definitions.

This puts your paradigmatic cognitive scientist in something of a pickle. On the one hand, he has (rightly, I think) *rejected* the idea that concepts are definitions. On the other hand, he cleaves (wrongly, I think) to the idea that having concepts is having certain inferential dispositions. But, on the third hand (as it were), *only defining inferences are compositional* so if there are no definitions, then having concepts *can't* be having inferential capacities.

This line of argument was first set out in Fodor and Lepore (1992). I think that is very close to being a proof that either there are definitions or the Pragmatist notion of what it is to have a concept must be false. Philosophical reaction has been mostly to opt for the first horn: If the price of the pragmatist account of concepts is reviving the notion that there are analytic/definitional inferences, then

there must indeed be analytic/definitional inferences. My own view is that cognitive science is right about being no definitions, and it's the analysis of having concepts in terms of drawing inferences that is mistaken. Either way, it seems clear that the current situation is unstable. Something's gotta give.

I return briefly to my enumeration of the varieties of pragmatist theories of concept possession. It should now seem unsurprising that none of them work. In light of the issues about compositionality that we've just discussed, it appears there are principled reasons why none of them could.

1.5 The 'theory theory' of concepts (and the problem of holism)

Pragmatists think that having a concept is having certain epistemic capacities; centrally it's having the capacity to draw certain inferences. We've had trouble figuring out *which* inferences constitute which concepts; well, maybe that's because we haven't been taking the *epistemic* bit sufficiently seriously.

Concepts are typically parts of beliefs; but they are also, in a different sense of 'part', typically parts of theories. This is clearly true of sophisticated concepts like ELECTRON, but perhaps its *always* true. Even every day concepts like HAND or TREE or TOOTHBRUSH figure in complex, largely inarticulate knowledge structures. To know about hands is to know, inter alia, about arms and fingers; to know about toothbrushes is, inter alia, to know about teeth and the brushing of them. Perhaps, then, concepts are just *abstractions from* such formal and informal knowledge structures. On this view, to have the concept ELECTRON is to know what physics has to say about electrons; and to have the concept TOOTHBRUSH is to know what dental folklore has to say about teeth.

Here are some passages in which the developmental cognitive psychologist Susan Carey (1985) discusses the approach to concepts that she favors: "... [young] children represent only a few theory–like cognitive structures, in which their notions of causality are embedded and in terms of which their deep ontological commitments are explicated. Cognitive development consists, in part, in the emergence of new theories out of these older ones, with the concomitant reconstructing of the ontologically important concepts and emergence of new explanatory notions (14)". "...successive theories differ in three related ways: in the domain of phenomena accounted for, the nature of explanations deemed acceptable, and even in the individual concepts at the center of each system... Change of one kind cannot

be understood without reference to the changes of the other kinds (4–5)". The last two sentences are quoted from Carey's discussion of theory–shifts in the history of science; her proposal is, in effect, that these are paradigms for conceptual changes in ontogeny.

The version of Pragmatism according to which concepts are abstractions from knowledge structures corresponds exactly to the version of Positivism according to which terms like "electron" are 'defined implicitly' by reference to the theories they occur in. Both fail, and for the same reasons.

Suppose you have a theory about electrons (viz that they are X) and I have a different theory about electrons (viz that they are Y). And suppose, in both cases, that our use of the term "electron" is implicitly defined by the theories we espouse. Well, the 'theory theory' says that you have an essentially different *concept* of electrons from mine if (and only if?) you have an essentially different *theory* of electrons from mine. The problem of how to individuate concepts thus reduces to the problem of how to individuate theories, according to this view.

But, of course, nobody knows how to individuate theories. Roughly speaking, theories are bundles of inferences, just as concepts are according to the Pragmatist treatment. The problem about which inferences constitute which concepts has therefore an exact analagon in the problem which inferences constitute which theories. Unsurprisingly, these problems are equally intractable. Indeed, according to the pragmatist own view, they are interdefined. Theories are essentially different if they exploit essentially different concepts; concepts are essentially different if they are exploited by essentially different theories. It's hard to believe it matters much which of these shells you keep the pea under.

One thing does seem clear: if your way out of the shell game is to say that a concept is constituted by the *whole* of the theory it belongs to, you will pay the price of extravagant paradox. For example: it turns out that you and I can't disagree about dogs, or electrons, or toothbrushes since we have no common conceptual apparatus in which to couch the disagreement in. You utter 'Some dogs have tails'. 'No dogs have tails' I reply. We seem to be contradicting one another, but in fact we're not. Since tailessness is part of my *theory* of dogs, it is also part of my *concept* DOG according to the present, holist account of concept individuation. Since you and I have different concepts of dogs, we mean different things when we say 'dog'. So the disagreement between us is, as comfortable muddleheads like to put it, 'just semantic'. You might have thought that our disagreement was about the facts and that you could refute what I said by

producing a dog with a tail. But it wasn't and you can't, so don't bother trying; 'you have your idea of dogs and I have mine'. (What, one wonders, makes them both ideas *of dogs*?) First the Pragmatist theory of concepts, then the theory theory of concepts, then holism, then relativism. So it goes. Or so, at least, it's often gone.

Two caveats: The first is that I'm *not* accusing Carey of concept holism, still less of the slide from concept holism to relativism. Carey thinks that only the 'central' principles of a theory individuate its concepts. The trouble is that she has no account of centrality, and the question 'which of the inferences a theory licenses are *central*?' sounds suspiciously similar to the question 'which of the inferences that a concept licenses are *constitutive*?' Carey cites with approval Kuhn's famous distinction between theory changes that amount to paradigm shifts and those that don't (Kuhn, 1962). If you have caught onto how this game is played, you won't be surprised to hear that nobody knows how to individuate paradigms either. *Where is this buck going to stop?*

My second caveat is that holism about *the acquisition of beliefs* and about *the confirmation of theories* might well both be true even if holism about the *individuation of concepts* is, as I believe, hopeless. There is no contradiction between Quine's famous dictum that it's only as a totality that our beliefs "face the tribunal of experience", and Hume's refusal to construe the content of one's concepts as being determined by the character of one's theoretical commitments. There is, to be sure, a deep, deep problem about how to get a theory of confirmation and belief fixation if you don't invoke a notion of conceptual necessity. But there is also a deep, deep problem about how to get a theory of confirmation and belief fixation if you are prepared to invoke a notion of conceptual necessity. So far as I know, there's no reason to suppose that the first of these problems is worse than the second.

So much for caveats. It's worth noticing that the holistic account of concepts at which we've now dead–ended is diametrically opposite to the Classical view that we started with. We saw that, for the likes of Hume, any concept could become associated to any other. This was a way of saying that the identity of a concept is independent of the theories one holds about the things that fall under it; it's independent, to put it in our terms, of the concept's inferential role. In Classicial accounts, concepts are individuated by what they are concepts of, and not by what theories they belong to. Hume was thus a radical atomist about concepts just where contemporary cognitive scientists are tempted to be radically holist. In this respect, I think that Hume was closer to the truth than we are.

Here's how the discussion has gone: Modern representational theories of mind are devoted to the pragmatist idea that having concepts is having epistemic capacities. But not just sorting capacities since sorting is itself relativized to concepts. Maybe, then, inferential capacities as well? So be it, but *which* inferential capacities? At a minimum, inferential capacities that respect the compositionality of mental representations. Defining inferences are candidates since they do respect the compositionality of mental representations. Or, rather, they would if there were any definitions, but there aren't any definitions to speak of. Statistical inferences, aren't candidates because they aren't compositional. It follows that concepts can't be stereotypes. The 'theory theory' merely begs the problem it is meant to solve since the individuation of theories *presupposes* the individuation of the concepts they contain. Holism would be a godsend and the perfect way out except that it's preposterous on the face of it. What's left, then, for a Pragmatist to turn to?

I suspect, in fact, that there is nothing left for a Pragmatist to turn to and that our cognitive science is in deep trouble. Not that there aren't mental representations, or that mental representations aren't made of concepts. The problem is, rather, that Hume was right: concepts aren't individuated by the roles that they play in inferences, or, indeed, by their roles in any other mental processes. If, by stipulation, semantics is about what constitutes concepts and psychology is about the nature of mental processes, then the view I'm recommending is that *semantics isn't part of psychology.*

If semantics isn't part of psychology, you don't need to have a sophisticated theory of mental processes in order to get it right about what concepts are. Hume, for example, did get it right about what concepts are, even though his theory of mental processes was associationistic and therefore hopelessly primitive: Concepts are the constituents of thoughts; as such, they're the most elementary mental objects that have both causal and representational properties. Since, however, concepts are individuated by their representational *and not* by the their causal properties, all that has to be specified in order to identify a concept is what it is the concept of. The whole story about the individuation of the concept DOG is that it's the concept that represents dogs, as previously remarked. (Well, *almost* the whole story; see Fodor, TOC, EE).

But if 'What individuates concepts?' is easy, that's because its the wrong question, according to the present view. The right questions are: 'How do mental representations represent?' and 'How are we to reconcile atomism about the individuation of concepts with the holism of such key cognitive processes as inductive inference and the

fixation of belief?' Pretty much all we know about the first question is that here Hume was, for once, wrong; mental representation doesn't reduce to mental imaging. What we know about the second question is, as far as I can tell, pretty nearly nothing at all. The project of constructing a representational theory of the mind is among the most interesting that empirical science has ever proposed. But I'm afraid we've gone about it all wrong.

Another way to put all this:

Theories of concept individuation are plausibly required to satisfy two sorts of constraints: they have to say something about how concepts function as categories (how they apply, and are applied, to the world); and they have to say something about the compositional principles according to which the contents of complex concepts are determined by those of their constituents. Roughly, psychologists have been concerned, almost exclusively, with the first of these requirements, philosophers (and many linguists) with the second.

As it turns out, neither is very hard to satisfy on its own. If, as psychologists are wont to do, you largely ignore the problem of compositionality, then prototypes, stereotypes and the like give a pretty good account of how concepts function as categories. If, as philosophers and linguists are wont to do, you largely ignore the extreme paucity of actual examples, and the failure to find their effects in 'performance' studies of categorization, definitions do fine for explaining the compositionality of thought and inference. What I've been claiming is that no theory according to which concepts are identified with inferential capacities can satisfy *both* requirements. (If this is not widely recognized, that's because philosophers, linguists and psychologists have generally been careful not to talk to one another much; each rightly suspects that the others bear bad news.) Since, *qua* pragmatist, all the current accounts do insist on the identification of concepts with inferential capacities, the moral would seem to be that we have to rethink the theory of concepts *from the ground up*.

At the very end of Portnoy's Complaint, the client's two hundred pages of tortured, non–directive, self–analysis comes to a stop. In the last sentence of the book, the psychiatrist finally speaks: "So [*said the doctor*]. Now vee may perhaps to begin. Yes?"

References

Bruner, J., Goodnow, J. and Austin, G (1956). *A Study of Thinking* (Wiley: New York).

Carey, S. (1985). *Conceptual Change in Childhood* (MIT Press: Cambridge).

Dennett, D. (1978). "Skinner Skinned", in *Brainstorms* (MIT Press: Cambridge).
Dewey, J. (1958). *Experience and Nature* (Dover: N.Y.).
Fodor, J. (1990). *A Theory of Content and Other Essays* (MIT Press: ?).
Fodor, J. and Lepore, E. (1991). "Why meaning (probably) isn't conceptual role", *Mind and Language*, 6, 328-343.
Fodor, J. and Lepore, E. (1992). *Holism, a Shopper's Guide* (Blackwell: Oxford).
Fodor, J. and McLaughlin, B. "Connectionism and the problem of systematicity; why Smolensky's solution doesn't work", *Cognition*, 35, 183-204.
Fodor, J. and Pylyshyn, Z. (1981). "How direct is visual perception? Some reflection on Gibson's 'ecological approach", *Cognition*, 9, 139-196.
Kendler, H. (1952). "'What is learned?' — a theoretical blind alley", *Psychological Review*, 59: 269-277.
Kuhn, T. (1962). *The Structure of Scientific Revolutions* (University of Chicago Press: Chicago).
McClelland, J. and Rummelhart, D. (1986). "A distributed model of human learning and memory", in McClelland, J. and Rummelhart, D. (eds.), *Parallel Distributed Processing*, Vol. 2 (MIT Press: Cambridge).
Peacocke, C. (1992). *A Study of Concepts* (MIT Press: Cambridge).
Smolensky, P. (1988). "On the proper treatment of connectionism", *Behavioral and Brain Sciences*, 11, pp. 1–23.
Woods, W. (1975). "What's in a link?" in Bobrow, D. and Collins, A. (eds.), *Representation and Understanding* (Academic Press: New York).

PHILOSOPHICAL ISSUES, 6
Content, 1995

Fodor's Concepts

James Higginbotham

In his paper "Concepts; A Potboiler", Jerry Fodor contrasts two types of views about inquiry into the nature of concepts and their deployment by creatures that have them. On the first of these, which Fodor calls the Classical view, a concept is a mental entity, individuated by what it represents. So weak a statement leaves it quite open what sort of entity a concept, or a human concept, is, and how it comes to be endowed with representational powers. Nevertheless, the Classical view diverges from what Fodor calls the Current Pragmatist view (henceforth simply the Current view, with a capital 'C'), according to which the very notion of a concept is derivative upon the more fundamental notion of having a concept. On Current views the concept F, if there is such a thing, will emerge at the end of an inquiry into what it is to have the concept F, or to be as it were F–concepted.

For a version of the Classical view, Fodor hearkens back to Hume. Current views are of many sorts, and include even behaviorism, according to which doctrine to have the concept F is just to be disposed to behave in certain observable ways, so that the concept itself, though not what it is to have it, disappears on analysis. Behaviorism apart, the Classical and Current views are views about a common project, with the common background assumption that creatures

think because they represent the world they think about, the latter thesis being the representational theory of the mind as Fodor understands and advances it. Concepts, or primitive concepts, are then the simplest representations, the ultimate constituents of a creature's thoughts and attitudes.

Current theories approach the constituents of thought indirectly, by asking what it is to have them in one's repertoire. Enlarging the behavorist scope for explication, Current theory holds that one has the concept F if one is disposed, or disposed on certain grounds, to behave and to infer in certain ways. The burden of Fodor's argument is that, so understood, all Current views lead to blind alleys or conceptual circles. For, Current views owe an explication of what Classical views take for granted, namely the individuation of concepts; given that concepts themselves are derivative, the principles of individuation must flow just from the behavioral–*cum*–inferential criteria for having them; and such an explication has not shown itself to be possible. To have argued successfully against Current views is not of itself to have advanced the inquiry into the basis for mental representation and thought; but it may clear the way for such inquiry, by setting aside the puzzles that Current theory is heir to.

Assuming that the above summary is sufficiently faithful to Fodor's text and intentions, I will in these comments first of all sharpen somewhat the contrast between the explanatory burdens faced by the Classical and Current views, and consider some issues for the Classical view that may be argued to threaten to restore all the problems Fodor presents for Current views. I turn then to Fodor's criticisms of Current ideas, emphasizing a point that Fodor makes in passing, that non–behaviorist interpretations of what it is to have a concept tend to presuppose the availability of other concepts of a like nature. Finally, I consider the question how Fodor's notion of representation is to be understood, and the prospects for a Current theory of concepts that deflects Fodor's criticisms. The articulation of that theory that I sketch will take for its model an analogous theory of meaning. It will conflict with views about psychology that Fodor has expressed elsewhere; but if I am right then the exercise will at least have shown that the issues between Classical and Current views cannot be resolved just on the grounds that Fodor presents in his paper.

1 Classical Complications

On what Fodor calls the Classical view, I believe, a concept is a mental entity with a semantics. He says that a given mental entity

is the concept *dog* if the mind whose mental entity it is uses it (or could use it) to think about dogs with (p. 4). Once we have tidied up this gesture, noting that the concept *dog* must be such that one can use it to think about all dogs, and only dogs, and allowing that one uses the concept *dog* to think about dogs in general and perhaps no particular dog at all, we have a conception of the concept *dog* as a mental entity that is true of dogs, and nothing else. In short, the concept *dog* on the Classical view is a word in the mind, and therefore a word of a language in the mind, individuated by the conditions on its reference.

The knowledge that for an organism to have the predicative concept F is for it to have a mental entity that applies to all and only the Fs tells us nothing whatever about the circumstances under which any particular organism does have that concept, or what the having of it may be. Nevertheless, Fodor says that on the Classical view the problem of what it is to have a concept —the deep problem to which Current theories are addressed— disappears: once we have given identity conditions for a concept, then "the story about concept possession follows without further fuss" (p. 4). This statement calls forth an immediate objection, and at least one further difficulty. The immediate objection is that the relevant notion of "having" a concept has not been filled in, and it would appear that to raise the question how it is to be understood is just to embark on the deep problem that Current theories attempt to address. The difficulty is that the mere notion of having a concept distinguishes only grossly between cognitive states, whereas what is crucial for many purposes is not just whether a creature has a concept, but how far it is in command of the concept. I will argue that the difficulty is in fact a more serious problem than the immediate objection is. But the objection is forceful, and requires treatment.

There are two standard ways in which the notion of "having" an F might be filled in. One is illustrated by the notion of having a dog. You have a dog (in one contextual setting) if you own a dog; so in this case the relation of "having" goes proxy for a contentful relation. The other is illustrated by the notion of having a Polish grandmother. You have a Polish grandmother if there is a woman who is Polish who is your grandmother; so in this case the "having" dissolves on analysis. Whichever method is correct for the case of "having" a concept, the Classical view certainly owes some interpretation or other, but of itself doesn't seem to offer any.

The method of dissolution, applicable to the notion of having a Polish grandmother, is not properly applicable to the notion of having a concept: *concept*, unlike *grandmother*, is an absolute, non–relational

noun. Hence to have the concept F is to stand in some contentful relation to it. The nature of this relation is just the problem for Current theory, which is not advanced by having identity conditions for concepts fixed in advance. The "story about concept possession", therefore, remains as blank as before: such is the objection.

As stated, however, the objection would miss the point that on the Classical view interpretations of what it is to have a concept are essentially less restricted than they are on Current views. According to the criteria for concept possession that Fodor attributes to them these views can be successful only insofar as they bring forth analyses of the form (1):

(1) x has the concept F if and only if $\phi(x)$

where ϕ would be appropriately non–circular, and in particular would not refer to the concept F. On the Classical view, however, the description of what it is to have the concept F can perfectly well refer to this concept, just as the description of what it is to have a dog quantifies over dogs, and what it is to have a Polish grandmother quantifies over Polish grandmothers. Thus it is possible for the Classicist, though not for Current theorists as Fodor describes them, to assail the project of analyzing what it is to have the concept F in terms like (2):

(2) x has the concept F if and only if $\psi(x, \text{the concept } F)$

Still, it remains to be seen whether the greater freedom afforded for the interpretation of concept possession amounts to anything. Toward the end of his paper Fodor writes:

> But if 'What individuates concepts?' is easy, that's because it's the wrong question, according to the present view. The right questions are: 'How do mental representations represent?' and 'How are we to reconcile atomism about the individuation of concepts with the holism of such key cognitive processes as inductive inference and the fixation of belief?'

Even if the these remarks are correct, the right questions just enumerated cannot be the only ones. There is also this question: given a mental representation that is properly representing whatever–it–is, so that the concept is as it were stored somewhere in the mind, the having of the concept must consist in certain capacities for sorting, inferring and so forth; and which ones are these?

Following a familiar Wittgensteinian line of thought, or a line of thought that is at any rate attributed to Wittgenstein, the question

can be parlayed into an objection to Classicism. First, we cannot say that having a concept s simply having a representation stored somewhere in the mind, for in that case one might have a concept that one can do nothing with, and that state is indistinguishable from not having the concept at all. Second, supposing we know the behavioral and inferential capacities that give evidence that an organism has a concept, whether there are concepts in Fodor's sense; i.e., mental words with a syntax and a semantics, is at best a secondary empirical issue, concerning those properties of the mind (or brain) that are responsible for concept possession (a more militant version of the Wittgensteinian position would be that the very existence of concepts in Fodor's sense is irrelevant.)

The Classicist, however, has no reason to accept either thesis. From the Classical point of view there is nothing incoherent in the supposition that one might have a concept that one cannot exercise. And it begs the question to suppose that the existence of mental tokens must be secondary to the account of concept possession; why shouldn't the laws governing the mental tokens and their properties be part of the empirical evidence that one has the relevant concepts?

Consider now the further difficulty that the mere notion of having a concept is not sufficiently rich to distinguish cognitive states. Besides the state of having the concept F there is also the state of *commanding* or having *mastered* the concept, of being fully competent with respect to it. If it is allowed that one may have only partial command of a concept that one possesses, then even if concept possession follows without further fuss, concept mastery does not so follow. It does not follow, because mastery of a concept requires establishment of the appropriate links between it and other concepts. For example, as one might offer, you might have the concept *the number two* without realizing that the numbers form an infinite sequence, or that you can always go on counting. But you do not command that concept unless you do realize that you can always go on counting. Now, Current theorists do not just ask what it is merely to have the concept F; they ask what it is to command that concept. It is a further question, given that a person may have but not command a concept, how far she may fall short of command, and what produces the concept in her even though she does not command it. Since concept mastery requires the appropriate links to other concepts, Current theorists are generally holists, in the sense that they recognize the existence of sets of concepts such that one commands no concept in the set without commanding all of them. But the holism does not come about because of Current

theory. It lies in the nature of concept mastery itself. For this reason much of Current practice must survive even if Fodor's position is accepted. Above, I quoted Fodor as asking, "How are we to reconcile atomism about the individuation of concepts with the holism of such key cognitive processes as inductive inference and the fixation of belief?" To these questions one might add, "How are we to reconcile atomism about individuation with the holism of concept mastery?"

I argued above that the casual objection that what it is to have a concept was not trivial even on the Classical view is one that of itself carries little weight, because the Classicist has a degree of freedom that Current views lack, to refer to the concept in the course of explaining what it is to have it; and that the Wittgensteinian objection, as baldly stated, begged the question against the Classicist. The difficulty just scouted, that the Classical view doesn't yield a sufficiently robust classification of cognitive states is, I think, more serious than either. There is another, more ad hominem, reason for stressing this latter limitation of Classicism; namely, that Fodor's chief objection to Current theories (behaviorism aside) is that they are enmeshed in the problem of individuating concepts according to analytic connections, something which he despairs making sense of: but even if Fodor is right the problem of analytic connections resurfaces at the level of concept mastery, so there is a difficulty for everybody.

2 The Critique of Currency

I shall with Fodor set aside, on the grounds that simple sorting criteria cannot distinguish coextensive concepts, the behaviorist conception of concept possession (the conception that would lead behaviorism into the doctrine of indeterminacy). Cutting through some of the distinctions Fodor makes among non–behaviorist Current views, I will offer the interpretation that according to Fodor the difficulties they all face are two. First, there is the problem of distinguishing what is central to possession of a concept from what is peripheral to it; I will call this the problem of centrality. Second, there is the problem of giving a basis for concept possession that does not involve antecedent concept possession; I will call this the problem of circularity.

Most of Fodor's arguments are variations on the problem of centrality. Particularly significant is his objection to the use of semantic postulates. With semantic postulates one interprets 'being G is

central to the concept F' as: 'it is logically (or conceptually, or analytically) impossible that some F is not a G'. The objection is that the formula rests on an unclear notion of possibility.

Fodor considers a position, the stereotype theory, that endeavors to avoid the problem of fixing the boundaries of analytic impossibility by appealing to individuation of concepts based on propensities to belief. The objection is then pressed that stereotypes are not compositional, so that the appropriate finite basis for a repertoire of concepts cannot be found. However, the congeries of dispositions the stereotype theorist appeals to would seem to be real enough: for anyone who has a particular concept, say *dog*, there are background conditions K such that one will judge with degree of belief d that x is a dog given that x is a K, and there are anticipations of perception A such that one will judge with degree of belief d' that x is A given that x is a dog. This congeries, like compositional semantics, will have to have *some* finite basis. So even if the congeries associated with a complex concept is not composed of those associated with its parts there should be some solution to the basis problem, and whatever it is can be recruited by the stereotype theorist. It is therefore unclear why the non–compositionality of stereotypes should be a decisive objection to the view.

I suspect, however, that the stereotype theory should by Fodor's lights inherit the problem of characterizing analytic impossibility. Fodor has consistently held, and holds in the paper under discussion, that common belief and mutual disagreement presuppose common concepts, so that what concepts represent must have the intersubjectivity of Fregean senses. On this assumption, even within the stereotype theory there must be a distinction between what things concepts are true of and what one believes about the things they are true of, and the question therefore arises how far one may depart from applying the concept to the things to which it applies and still be said to be using that very concept. A departure that is too great to be tolerated will reflect the counterpart of an analytic impossibility, whose explication was said to be obscure. The problem of centrality, therefore, will reemerge.

The problem of circularity is seen most clearly in the case of what Fodor calls definitional pragmatism. Definitional pragmatism presupposes the availability of some concepts, and offers definitions for the rest. About how we come by the basic concepts, Fodor remarks, "definitional pragmatism tended to be hazy" (p. 12). Quite apart from the question whether a non–trivial definitional basis for typical human concepts exists, Fodor's remark points toward an objection to every variety of Current theory. Consider again the paradigm

x has the concept F if and only if $\phi(x)$

The problem is that the condition ϕ will typically contain reference to how x is disposed to *judge* of things; and therefore the possession of concepts representing the notions used in giving the contents of these judgments will be presupposed. This feature of Current theory is seen clearly in the version (S) of the paradigm for concept possession suggested in Peacocke (1992):

(S) F is the concept G such that x possess G iff $\chi(x, G)$

where the condition χ will, for ordinary empirical concepts, involve a formula like

> x finds it primitively compelling that $G(y)$ under circumstances where y is present in mode M

Finding something primitively compelling is coming to have an attitude toward a propositional content. We must therefore be able antecedently to think such contents. Even if there are concepts so simple that their possession conditions do not involve concepts at all, it does not seem credible that these would be adequate to constructing definitions of all other concepts.

3 The Represented

Fodor writes that concepts on the Classical view are individuated by what they represent, so that a mental token is of the concept *dog* just in case it represents dogs (or "represents dogs as such", as Fodor later modifies it, and as I shall understand it). This formulation gives rise to the obvious question how the predicate 'represents dogs' is to be understood. Its object position is of course not extensional. But it is not, in the sense of possible–worlds semantics, intensional either; or at least it is not intensional by Fodor's lights, since he holds that concepts are constituents of beliefs, and those constituents cannot be intensionally individuated. Having the concept *triangular* is to be distinguished from having the concept *trilateral* (p. 8); so the mental tokens that represent triangles cannot always coincide with those that represent trilaterals.

Are we to say then that 'represents . . . ' is a non–extensional, non–quotational context, and leave it go at that? If we include complex concepts as well as simple ones, then there are infinitely many concepts, between any two of which the question arises of the identity of what they represent. And since Fodor does not accept cointensivity

as sufficient for identity, there will be infinitely many questions of identity even within a given class of cointensive concepts.

The Current theorist could suggest as a criterion for identity, that concepts represent identical things if the concepts are cointensive and agents command them under the same conditions (on some conceptions of the sameness of conditions of concept mastery the cointensiveness of two concepts would even follow from their being commanded under the same conditions). Just as the problem of characterizing meanings is not fatal to the notion of synonymy, or the notion of translation, the problem of centrality is not fatal to the application of the Current theory's criterion of concept–identity. But the victory is a hollow one unless the problem of centrality can be fended off.

Suppose one concludes with Fodor that the problem of centrality as he describes it is intractable: there is no notion of analytic or conceptual impossibility that is sufficiently clear to distinguish what is central to a concept from what is peripheral to it. And suppose that one concludes also that the problem of circularity cannot be avoided: we cannot build up the set of human concepts from any simple basis. I will suggest that a model drawn from the theory of meaning can motivate a Current theory of concepts that can abide both consequences.

The question of the individuation of concepts has an exact analogy in the philosophy of language, in the question of the individuation of meanings. Suppose that

> the word 'dog' means dogs

Then the position of the object of 'means' is neither extensional nor intensional; or so I will assume. How then should attributions of meaning be understood? The Current theory of meaning, in analogy with the Current theory of concepts, would suggest that we should ask not for the meaning of the word, but rather for what a person knows who knows the meaning of the word. Taking this step, and assuming that knowledge of meaning is fundamentally knowledge of conditions on reference, we have such statements as (3):

> (3) If you know the meaning of the word 'dog', then you know that: 'dog' is true of x iff x is a dog

But we have also much more. It is obvious that if you know the meaning of the word 'dog' then you know that it refers to things that have four feet, bark, and are kept as pets, and you know that it is a law of nature that they have four feet and bark, and a contingent

social matter that they are kept as pets rather than, say, worshipped or kept as livestock. The Current theory of meaning does not have to distinguish between analytic and synthetic (though this distinction may be useful for other purposes). What one has to know to know the meaning of an expression need not comprise all the analytic (or even logical) truths involving it, and may include a number of contingent facts, and moreover the fact that they are contingent. A person can get hold of the word with its meaning even if she does not appreciate what that meaning is; i.e., even if she does not know what one is supposed to know about its reference. In this way the account makes room for the distinction between being able to use a word (the analogue of concept possession) and being competent in it (the analogue of concept mastery). Phrases will have the same meaning if what you have to know to know their meanings is the same; so the notion of synonymy is firmly available, whatever its extension turns out to be exactly.

Consider now the prospect of taking this Current theory of meaning as a model for the Current theory of concepts. The non–necessity of a distinction between analytic and synthetic carries over, provided that one can on other grounds distinguish disagreement about matters of fact from divergence of concepts. The natural thought is that the distinction is to be made, not at the level of what belongs narrowly to an individual's psychology, but in terms of social practice. For Fodor, this consequence would be grounds for rejecting the proposal; but the questions that then arise concerning the nature of psychology and psychological laws are beyond the scope of this discussion. The Current theory of meaning, however, presupposes the possession (although not the mastery) of the concepts used in describing what the competent speaker knows. Likewise, any formula

> If you have the concept *dog* then you know that ...

will presuppose the possession of whatever concepts figure in the complement clause '...'. Because of this, the strategy of the Current theory of meaning is problematic when transferred to the theory of concepts itself. Problematic, because depriving us of the possibility of a substantive foundational theory, but not therefore circular; for it may simply turn out that there is a very large set of human concepts, each of which has one or more of the others figuring in the conditions for its mastery.

Current theory, as I have just reconstructed it sketchily and in the large in response to Fodor's critique, has at least one good reason for putting the question when one has a concept ahead of the

question what a concept is, namely that the semantically obscure notion of representation does not figure in it. Notions like knowledge, and disposition to judgment, do figure in it, and like the notion of representation their contexts are neither extensional nor intensional. On the assumption that these notions, even if they acquire a technical overlay in the course of the development of a philosophical and psychological theory of concept possession, still touch base with what we are antecedently familiar with, so much the better for them, and so much the worse for taking representation as a primitive notion.

It is appropriate at this point to consider the possibility that the objects of representation are properties, broadly construed. Perhaps the context

represents Fs

can be understood as

represents the property F

with properties in their turn individuated on metaphysical grounds. An ad hoc metaphysics of properties does not solve the problem of concept individuation; it just begs the question. But it is possible that some principled metaphysics could be formulated that would individuate properties more finely than intensions, and insert the representations of properties satisfactorily into explanations of behavior. Indeed, the Current theorist must concede that social practice comes into play only for animals that have the relevant practices, a proposition that at first blush is doubtful for pandas, very doubtful for cockroaches, and absurd for paramecia. But all these animals can plausibly be held to represent the world by representing properties of it. On the other hand, (i) it may be that the concepts such animals have are so simple that what is analytic to them can be specified, and (ii) the application of the metaphysical theory to human beings remains doubtful, because what is represented by human beings is individuated by states of knowledge and belief, which are highly sensitive to the ways in which physical and other properties are specified.

I have appealed to the sensitivity of mental states to the notation used for expressing them. Fodor himself has elaborated elsewhere a syntactic theory of the epistemic states and propositional attitudes, according to which 'x knows that dogs have four feet' would owe its truth to x's standing in some relation to a synthesis of concepts in x's language of thought, that synthesis expressing that dogs have

four feet. Current theory need not dispute this; the explanatory priority, however, will go to explaining what it is to think that dogs have four feet, that information being used to identify the synthesis of concepts in question.

4 Conclusion

In this discussion I have given some reasons for supposing that the investigation of concept possession, and still more of concept mastery, in terms of Current theory will remain substantially intact even if the Classical theory, that to have the concept F is just to have a representation that represents Fs, should prove to be correct. Fodor's critique of Current theory, as I see it, ultimately rests on the same type of considerations that led Quine away from concepts altogether, namely the inability to specify what is central to a concept and what is only peripheral to it, or in Quinean terms what is analytic and what is mere widely shared collateral information. This is the problem of centrality, which can perhaps be overcome by rejecting the premise that what one needs to know to count as master of a concept is just what is analytic to it. The boundary instead will be between what one is required to know concerning the application of the concept, and what one is not required to know, a distinction that, it is hoped, will in turn serve in place of the analytic–synthetic distinction to ground the distinction between difference in belief and difference in conceptual scheme. The problem of circularity, which loomed very large when the hope was to found all concepts upon a narrow basis, is not overcome, but simply taken on board as characteristic of concepts, or at least of human concepts. Finally, the notion of representation is explicable in terms of Current theory, assuming that it can be completed in other respects.

I said above that there was at least one point where Fodor would demur at the revised Current theory sketched here: he would not accept that social practice could replace individual psychology as the basis for distinctions critical to the individuation of concepts. I have also assumed that there is a distinction between merely having a concept and being fully competent with respect to it, a tendentious thesis that has characteristically been rejected by a psychology of what Tyler Burge has called individualistically individuated mental states. Much of what I have said depends also on the view that the notion 'represents Fs' wants a semantics; that it will not do at least for human cognition to take the object of 'represents' as denoting a

property; and on my assumption that the contexts of knowledge and judgment, as they are made more precise and developed in technical directions in cognitive science, will remain clearer than representation itself is. In all these ways, Fodor's discussion seems to me to lead beyond its immediate confines, to issues of larger and continued significance.

References

Peacocke, C. (1992). *A Study of Concepts*. Cambridge, Mass.: The MIT Press. A Bradford book.

PHILOSOPHICAL ISSUES, 6
Content, 1995

On What It's Like to Grasp a Concept

Joseph Levine

In his paper, Fodor argues that the turn from the metaphysical concern with what a concept is to the epistemological concern with what it is to have a concept was misdirected, and has led to pragmatist theories with apparently insurmountable problems. I largely agree with him on this point. In particular, I think the major difficulties he identifies with the pragmatist theories he surveys are indeed serious and therefore a rethinking of the turn he describes is warranted. Yet, I find myself unsatisfied, thinking that there is a legitimate demand behind the epistemological turn that Fodor is neglecting. What I want to do is explore this demand, and see whether any sort of theory we can now imagine could satisfy it.

One source of the epistemological turn in the theory of concepts is clearly connected with the rise of behaviorist, pragmatist, and Wittgensteinian approaches to the philosophy of mind and language in general. Concerns that meanings and mental states, if not pinned down by "outward criteria" of some sort, are hopelessly illusive, subjective, private, and therefore ultimately mysterious, lead naturally to the idea that what is at issue in a theory of concepts is the question of what conditions justify the attribution of conceptual competence to a subject. On this way of looking at it, the turn from the metaphysical concern to the epistemological concern follows from a turn

from the first–person perspective to the third–person perspective. While Classical theorists took the data of their own mental lives as starting points for theory, the pragmatists and behaviorists took the primary data to be the behavior of others, and then assumed that activity of some sort, ultimately related to discriminative behavior, must be the ground for the attribution of concepts.

Today too we find this obsession with what would justify our attribution of conceptual competence to others. Many of the arguments that purport to show that some inference or other is constitutive of a concept involve asking the rhetorical question: could you really justify granting Jones competence with concept X if she couldn't infer Y? I believe that there are answers to these sorts of objections available to the conceptual atomist, and of course it is now generally conceded that a robust mentalism can survive the methodological scruples of objective science.

I say all this mostly to set it aside, since it is not this source of the epistemological turn that concerns me here. In fact, quite the contrary. I believe, though I make no historical claims in saying this, that another source of the epistemological concern is actually located in the first–person perspective itself. That is, from the inside we seem to have the experience when entertaining a thought of actually "grasping" its meaning in a way that seems to demand a certain kind of epistemological account. But just what is this demand, and is it possible to meet it?

When we entertain a thought —occurrently, consciously— we know what we're thinking. Perhaps this knowledge isn't incorrigible, and maybe it involves no special, privileged access of the sort that could serve as the foundation for all knowledge, but generally we know quite well what we're thinking and it's not likely we're wrong. This knowledge of what we're thinking involves the content of our thought. I know when I ponder the swaying green leaves I can see from my front porch as I write this that I'm pondering *leaves*, among other things. Is there a problem about this?

Well, some have thought so (cf. Bhogossian). In particular, it's thought to be a problem for externalist theories of content. If, for example, my mental representation LEAF refers to objects which have a fancy biological description that I don't know —and if I couldn't tell a leaf from a twin–leaf— then it looks like I really don't know what I'm thinking about. Of course one might not be worried about this sort of case because it may be thought that to a certain extent I don't know what it is I'm thinking about. If leaves are essentially objects with chlorophyll, say, and I don't know this, then to a certain degree it's true that when I think of a leaf I don't know what

constitutes the object of my thought, and therefore don't know what I'm thinking about. The fact that I couldn't discriminate genuine leaves from fake ones is testimony to this level of ignorance.

I think this response to the problem is only plausible if it goes along with a major retreat from externalism, to the view that the actual content of my thought —what it is I grasp in knowing what I'm thinking— is not the property of being a leaf, but some more narrowly delineated property, one that leaves and twin–leaves are sure to share. For then it is easy to admit ignorance of the identity conditions of leaves, and thus the wide contents of my thoughts about leaves, while holding on to what seems patently obvious: namely, that when I contemplate the swaying leaves in front of me I know what I'm thinking about.

Again, one could try to sidestep the problem by denying the data. Of course I can have thoughts about thoughts, and the contents of these second–order thoughts may be to the effect that I am having thoughts of the first–order with a certain content. In this sense I can know what I'm thinking about. But this state of knowledge, this relation to a second–order representation, itself must have a content within which the content of the first–order thought is represented. Whatever problem there is in grasping the content of the first–order use of LEAF will also attach to the content of the second–order use of LEAF. Hence, the idea that I do know what I'm thinking about is an illusion. It's LEAVES, not leaves, all the way down.

In fact I think there is something right about this response, but rather than solve the problem it actually serves to deepen it. The point is, there seems to be an experience I have of knowing my own thoughts, and on an externalist theory, at any rate, it turns out to be a kind of illusion. Put another way, on the externalist story, which the response above is part of, there is no model of how I could have the kind of knowledge I patently have. To deny I have it is no solution.

Now, on the Classical Empiricist view, it may have seemed that there is no problem of this sort. Following Fodor's guide to the classics, let's suppose that according to a Humean theory of concepts mental images are the vehicles of thought and resemblance is the relation that constitutes reference. Entertaining a thought about leaves is to token the LEAF image in the appropriate manner. We seem to have no problem with identifying the contents of our thoughts on two grounds: first, the images themselves, which constitute the immediate object of thought, are "inside", so grasping them seems to be no problem; second, we can know what the internal image is about by looking at the image itself since it's about whatever it resembles.

With regard to the first consideration, one might object that we have the very same ground for avoiding the problem with modern externalist theories as well, since the vehicle of thought, the mental symbol, is inside as well. If the internality of the image is a help, why not the internality of a symbol? Perhaps in the end there is no difference, but there certainly might seem to be one. The point is that with a mere symbol, there is nothing intrinsic to it that even looks like a content —its content is in some sense arbitrarily related to it. However, an image seems to carry a content intrinsically, which is the basis for determining its external reference by resemblance. Thus, the two grounds for avoiding the problem are intimately connected.

So, once we drop the idea that images are the primary vehicles of thought, and we replace it with the idea of a symbol which has no intrinsic relation to what it picks out, the problem of how we know what we're thinking about arises. It then makes sense to focus on the epistemology of concept possession, and, in turn, to reject theories that remove all content from within the subject. Surely I have something in mind which I can grasp —and which I can know that I grasp— when I think of a leaf that is more than merely the symbol LEAF. Well, if it isn't, by hypothesis, the image of a leaf, perhaps it's the definition of LEAF.

As I said above, I make no historical claims for this account of the epistemological turn. Perhaps no one of its adherents ever had this concern with self–knowledge in mind (though clearly some have worried about self–knowledge in relation to externalist theories of reference). But for me, anyway, it captures the feeling of dissatisfaction I have with the sort of atomistic account of concepts that seems to be the principal alternative to the pragmatist theories Fodor criticizes.

Two questions arise immediately. First, does any sort of conceptual role theory help with this problem? Second, is there any other sort of theory that could help? As I just mentioned above, it might seem that it would help explain how I know what I'm thinking about if entertaining a concept involved entertaining its definition. But there are two problems with this line of reasoning. First, it's just not plausible that in entertaining a concept I am literally entertaining its definition, at least not if entertaining is taken to be an occurrent, conscious state. Of course there may be rare cases where this is indeed what's going on, but they can't provide a model for the general case. But then the content of my thought about leaves, say, is determined not by what's happening at that moment, but by the dispositional properties of the state I'm in —for instance, that I'm disposed to infer that what I'm looking at is alive from the fact that it's a leaf, even though that it's alive is not a thought I'm cur-

rently entertaining. How can awareness of my thought's dispositional properties help anymore than awareness of its relations to external properties? For sure, I have better access to what I'm disposed to infer than to the biological essence of a leaf, but it certainly isn't so much better that it can ground the sort of immediate access to content allegedly involved in my grasp of my own thoughts.

The second problem is that dispositions to infer other thoughts, to entertain other concepts, can't help us to get out from behind the veil of symbols, which is the fundamental difficulty in the first place. We want to understand how I grasp a concept by tokening a symbol which has no intrinsic relation to that concept. All the conceptual role theories can provide is more tokenings of more symbols. How does any of this symbolic activity amount to thinking about leaves? If it's the causal relation of the entire system to the world that does the work, then why can't the causal relation between LEAF and leaves suffice? If causal relations don't suffice, for the reasons set out above, then many more symbols with many more causal relations aren't going to make it any better.

Earlier I mentioned that on the Hume–according–to–Fodor story it seemed unproblematic how I know the contents of my own thoughts since the images, which are intrinsically contentful, are themselves inside the mind. On at least one way of understanding the Fregean story —the Platonic version— it might also seem unproblematic. Mediating between mind–independent, external properties (which, as I understand it, is closer to what Frege had in mind by the term "concept" than what we mean by "concept" now) and our mental states stands the realm of senses. A sense is an abstract entity which constitutes a mode of presenting the objective, mind–independent world to the mind. When we think we grasp a sense, and thereby think about what the sense presents to us. Senses, like Humean images, are intrinsically contentful, and, though not inside the mind, are of such a nature as to be immediately graspable by a mind in much the way an internal image might be.

But neither of these stories can hold up, at least not for a naturalist. There is a problem with both the relata and the relation posited by both the image and sense stories. There don't seem to be images inside the head (at least not of the Humean variety), and, as for senses, they too, understood Platonically (as opposed, say, to descriptions), constitute an ontological extravagance from the point of view of a naturalistic metaphysics. But even more problematic is the relation required by these stories. After all, how does the mind apprehend an image, even if it is inside, or grasp a sense? The only sorts of relations intelligible within a naturalistic framework

are those constructed from causal mechanisms, and, as we've seen, causal relations do not give one the sort of content apprehension that solves the problem. But even if we throw caution to the wind and abandon naturalistic scruples, can we envision how an even un–natural relation might provide the sort of immediate apprehension of content we're after? I don't see how myself, unless we just stipulate the existence of a primitive "intellectual grasping" relation.

To sum up, I've argued that a problem about self–knowledge of content might be behind the epistemological turn that Fodor criticizes. It might seem that only by imputing more substantive content inside the mind can we account for the character of occurrent episodes of entertaining thoughts. However, I've also argued that so long as we restrict ourselves to relations sanctioned by a naturalistic framework —which ultimately are all reducible to physical causal mechanisms— adding more internal, narrow content won't help with this problem. No natural relation would seem to provide the sort of transparency, or immediacy inherent in our conception of what it is to grasp a concept.

I would like to make one final point before concluding. Some philosophers, most notably Searle, have argued that there is an inherent connection between consciousness and intentionality. Nothing I've said here is meant to suggest this. Certainly there can be unconscious thoughts, and I see no reason the contents of such thoughts can't be determined according to some form of information–theoretic semantics (or at least no reason stemming from their being unconscious). The empirical arguments for the existence of unconscious computational states are overwhelming, as far as I can see, and to my mind that should settle the matter.

However, the concern I've raised in this paper with the first–person experience of grasping a thought does connect the issues of consciousness and intentionality in at least two ways. First, it is when thoughts are conscious that the problem seems to arise, which suggests that consciousness in some way affects the character of the intentional relation itself. Second, the problem that consciousness poses in general for a theory of the mind has a lot to do with the same issue of immediacy, or transparency as the one we've been discussing. Qualia are generally representational —the quale I experience when looking at something red is a way of presenting red surfaces to the mind. If I understood how such a presentation were possible, I bet I'd understand how it is possible to grasp a thought in the way that gives rise to full–blooded self–knowledge of content. At present, however, both seem equally mysterious.

PHILOSOPHICAL ISSUES, 6
Content, 1995

Can Semantic Properties Be Non–Causal?

Pierre Jacob

As you must have gathered from Jerry's paper, he is not really enthusiastic about what he calls *Pragmatist* accounts of concepts. By his lights, a Pragmatist account claims that having a concept is "being able to *do* certain things rather than being able to *think* certain things" (p. 7). This Pragmatic view, he says, is typical of Current (or Contemporary) thinking about concepts. And it is deeply wrong.

Now, it is tempting to question Jerry's contrast between thinking something and doing something, on the grounds that thinking something can't be doing nothing. On second thought, however, and on behalf of Jerry's anti–Pragmatism, it might well be that a person's thinking a thought does not reduce to anything else the person may do. Perhaps it does not reduce to any non–thinking behavior in which the person may engage. Jerry's anti–Pragmatist stance would then merely amount to an innocuous aversion to Behaviorism. Nor would the identity between a person's thinking something and the electrical activity of her brain cells vindicate a Pragmatist theory of thinking, since the electrical activity of a person's brain cells is not something the *person* does. Rather, it is something which occurs within her.

However, I am not interested in pursuing this line of thought as such. For I take it, what Fodor objects to is a Pragmatist theory of *concepts*, not a Pragmatist theory of thinking. The issue raised by Fodor, I think, is not whether thinking something is or is not doing something. The question rather is whether a system can have a concept and not do anything with *it*. The issue is whether a concept can stand by itself in a creature's mind without a network of surrounding concepts, without entering a system of beliefs and propositional attitudes, without being part of a number of epistemic capacities, without entering inferences or inferential processes.

Granted, the issue is not whether thinking a thought is or is not doing something or other. There is, however, a slightly different question which, I think, is relevant to Fodor's project. The question is: What is the difference between thinking about Xs and being able to reason, to make inferences and to recognize Xs? In the process of exploring this slightly different question, I would like to suggest that there is a tension in Jerry's thinking: on the one hand, Jerry subscribes to an informational view of the content (or meaning) of primitive concepts. On the other hand, in this paper, he argues for keeping the semantic properties of symbols apart from their causal properties. This is the content of his claim that semantics is not part of psychology. But since on any informational account, the content of a symbol depends on its causal relations with properties in the environment, one may wonder whether Jerry can divorce the causal and the semantic properties of concepts as he wants to.

To start with, notice a crucial difference between Jerry's notion of a concept and the Fregean tradition. In the Fregean tradition, a concept is a sense (or mode of presentation of an object) —something abstract which is expressed by a symbol (e.g., a word of a natural language). Presumably, in the Fregean tradition, concepts do not have causal properties. But, on Jerry's view, they do because concepts *are* symbols in the language of thought.

With Fodor, let's distinguish primitive (undefinable) concepts and complex concepts built out of the former. Not only complex concepts, but also propositional attitudes are built out of primitive concepts. Primitive concepts are the simplest kind of things with both semantic and causal properties. Now, I will first try and characterize accurately Jerry's present position. I will ignore logical and mathematical concepts or concepts without an extension (I will assume that concepts without an extension can't be primitive).

Although concepts have both semantic and causal properties, only their semantic properties, not their causal properties, are relevant to the individuation of primitive concepts. Presumably, on Fodor's

present informationally–based view of concepts, a primitive concept derives its representational and misrepresentational powers from the Asymmetric Dependency Condition. Now, what of coextensional concepts? On the assumption that they derive their representational and misrepresentational powers from the Asymmetric Dependency Condition, aren't they bound to share all their semantic properties? Well, on Fodor's view, there is still room for distinguishing two coextensional concepts: coextensional concepts can still be distinguished by the syntactic properties of their respective vehicles, specifically, by the constituent structure of the distinct vehicles (the distinct mental symbol–types or vehicle–types) expressing the concepts. So the concept CREATURE WITH A HEART will differ from the concept CREATURE WITH A KIDNEY in virtue of the fact that the vehicle of the former includes the vehicle "heart" and what expresses the latter includes the vehicle "kidney". The concept WATER differs from the concept H_2O in that one can have the former, not the latter, without having the concepts HYDROGEN, OXYGEN and NUMBER 2. It follows, of course, that distinct coextensional concepts whose respective vehicles have different constituent structures are precluded by Jerry's view.

What is distinctive of Fodor's present informationally based view of the nature of primitive concepts is that, as he likes to put it, "semantics is not part of psychology". It's the business of semantics, not psychology, to individuate concepts. So the division between semantics and psychology merely reflects the difference between the semantic and the causal properties of concepts. The task of semantics is to understand what makes it possible for elementary and complex mental representations to have the power to represent things. In virtue of what does some physical object (or some state of a physical device) have the ability to represent (and misrepresent) things and states of affairs in the world? Where do mental representations derive their power to represent things from? How could a purely physical system represent and misrepresent things out there? This is the task of a naturalistic semantics. In a sense, naturalistic semantics —for the language of thought— has been elevated to the rank of metaphysics. Fodor wants to know how symbols in the language of thought hook up onto the world. And you don't need psychology to do this. (Nor do you need epistemology.) The task of psychology is to produce a theory of the underlying causal mental processes turning a thought into another thought or leading from thought to intentional behavior. Given that, on Fodor's view, the causal properties of mental representations are syntactic or formal (i.e., non–semantic), we have a nice division of labor: psychology is computational. It deals with

causal processes. Semantics has nothing to do with it. It deals with the representational properties of symbols.

Mirroring the distinction between semantics and psychology — the distinction between the semantic and the causal properties of concepts— is the distinction —central to Jerry's present paper— between concept–individuation and concept–possession. The two main theses of Jerry's present paper are the following two:

1. Every theorist faces the following option: either concept–individuation is derivative (or parasitic upon) concept–possession or vice–versa concept–possession is derivative (or parasitic upon) concept–individuation. For cats, says Jerry, it's clear, which should come first: first, you individuate cats; and then you say what having one of those consists in. For pains and jobs, he concedes, it goes the other way around: first, you say what it is to have one; and individuation follows.

2. For concepts, individuation comes first; possession comes next.

Classical theories derived concept–possession from concept–individuation. And they were right to do so. Jerry's prototype of Classical theorizing about concepts is Hume. On Hume's view, concepts derived their semantic properties from Resemblance. Causal properties of concepts depended on laws of Association. Now, Fodor thinks that Hume's account of both semantic and causal properties of concepts was wrong. On Fodor's view, semantic properties —as I said— are informational; and causal properties are computational. As he said somewhere else: Dretske and Turing bewteen them have solved the problems of intentionality. But Hume was right to give priority to concept–individuation over concept–possession.

On his view, Current criticisms of Classical theories have completely missed their target. What they should have done is reject Hume's particular account of the semantic and causal properties of concepts but keep the priority of concept–individuation over concept–possession. Instead, they have taken the other option: they have given priority to concept–possession over concept–individuation. And they are deeply wrong to have done so. In Jerry's words: "this subtle and largely inarticulate difference between Contemporary RTM and their Classical forbearers has... the most profound implications for our cognitive science... I suspect that it was a wrong turn —on balance, a catastrophe...".

What Jerry calls a Pragmatist view of concepts is a view which gives priority to concept–possession over concept–individuation. Presumably, the reason why Jerry calls such a view a Pragmatist

view is that a view which gives priority to concept–possession over concept–individuation is bound to individuate a concept by means of what a creature who possesses the concept is likely to do with it. By contrast, Jerry would say that a view which gives priority to concept–individuation over concept–possession would individuate a concept by saying what a concept is a concept of. A Pragmatist view, then, is a view which says that semantics *is* part of psychology. It might —in a Fregean style— define concept–possession by the capacity to have certain beliefs and other propositional attitudes. And so a concept would then be individuated in terms of the beliefs whose complex contents contain it as a constituent. From Jerry's perspective, this is unacceptable since it falls prey to Meaning Holism, which, in turn, is inconsistent with scientific psychology.

Now, I'd like to make four points. First, I'd like to ask Jerry for clarification of a textual matter. In the course of recontructing the misdirected criticisms of Current thinking to the Classical view, he lists three objections —methodological, metaphysical and epistemological. It's the third kind of objection which raises a puzzle in my mind. As I understand it, it goes something like this: What is wrong about the Classical picture (from Descartes to Kant) is that it postulates a system of internal representations —a Veil of ideas. A representation (something mental) can only make contact with other representations (other mental things). So the question arises: How can one regain contact with the world after one has lost it by postulating the Veil of ideas? And Fodor suggests that giving priority to concept–possession over concept–individuation was then thought of as a way out of the Veil of ideas. This is what I don't get: Assuming that the problem is to get out of the bottle of mental representations, then how could giving priority to concept–possession show the fly the way out of the bottle of mental representations? Why should a system of epistemic capacities make it any easier to contact the world again? It seems to me that giving priority to concept–possession does not get rid of the Veil of ideas; rather, it reinstates the Veil of ideas. Externalism would have seemed to be the cure; not epistemic capacities.

My second point has to do with Frege's Puzzle. Now, a Pragmatist view of concepts gives priority to concept–possession. As I mentioned earlier, making a primitive concept sensitive to the complex contents of beliefs of which it might be a constituent may itself be taken to be inspired by Fregean considerations. It's no suprise, therefore, if the priority of concept–possession (over concept–individuation) to which Jerry strongly objects —courtesy of his Semantic Atomism— offers a natural response to what is called Frege's Puzzle, namely the

problem of how a rational thinker S may accept it, or believe, that Fa, and not accept it, or believe, that Fb, when "a" and "b" are primitive and coreferential, without contradicting himself or herself. Fregean modes of presentation were introduced precisely in order to allow S to rationally believe that Fa (under the "a"–mode) while disbelieving that Fb (under the "b"–mode).

Now, it seems to me, on Jerry's informational line "Fa" and "Fb" have the same content. Of course, one might try to concoct some notion of narrow content such that for S, "Fa" and "Fb" have the same broad content and different narrow contents. But I take it, Jerry has now given up on this line. So Jerry must —as I have heard him do in lectures a year ago (in Paris where he delivered the Jean Nicod lectures)— account for Fregean cases in *non*–semantic terms, i.e., in purely causal, syntactic or computational terms. On Jerry's view, syntactic structures in the language of thought play precisely the role of Fregean modes of presentation. And I think it's a serious problem to give up on intentional explanation of what is going on in Fregean cases. The reason it's a serious problem is that giving up on an intentional story seems not compatible with the assumption that S —the thinker— is rational. Not to know that $a = b$ is a case of ignorance, not irrationality. And non–intentional explanations —lapses from intentional explanation— seem to fit irrational, not rational, behavior.

A third (psychological) problem for Jerry's sharp distinction between the semantic and the causal properties of concepts is that it may make it difficult for him to recognize the difference between the contents of genuine propositional attitudes and the contents of merely information–carrying subdoxastic states. I am not questioning the fact that information–carrying subdoxastic states may, as well as beliefs and desires, represent and misrepresent the world. And Jerry certainly has an account of the representational and misrepresentational powers of such information–carrying states, in terms of the Asymmetric Depedency Condition. The (psychological) problem rather has to do with the fact that, I take it, a natural constraint on the conceptual content of beliefs —as opposed to the non–conceptual content of subdoxastic states— is that the causal powers of the former should depend on their contents. In the production of intentional behavior —behavior done for reasons—, the contents of states having causal efficacy should be doing some job. As many people have pointed out, it is one thing for a state having semantic properties to be causally efficacious in virtue of its non–semantic properties. It's another to be causally efficacious in virtue of having the semantic properties that it has.

Now, Jerry might want to distinguish between subdoxastic information–carrying states and beliefs without adverting to any difference in their semantic properties. He might say: the difference resides entirely in the different causal or functional properties of the vehicles. The difference would be this: in virtue of their causal properties, beliefs belong to belief–box in which they have the computational potential to interact with any other beliefs (and with other propositional attitudes for that matter). Subdoxastic information–carrying states are *modular*: in virtue of their computational causal properties, they belong to a vastly smaller box with a vastly more restricted potential for interacting with other information–carrying states. He would, therefore, claim that the difference I am alluding to can be exhaustively accounted for in terms of the causal properties of representations —without appealing to their semantic properties.

But then on the one hand, he has to admit that concepts *are* partly individuated by their causal properties. Actually, when you think of it, it's not clear that Jerry is really entitled to his separation between semantics (or metaphysics) and psychology. After all, his own solution to the problem 'Where do mental representations derive their power to represent things from?' appeals to the Asymmetric Dependency Condition, which in turn is no other than a higher order relation between causal dependencies. Now, if mental symbols derive their semantic properties from a higher order relation between causal dependencies, then the divorce between semantics and psychology —or between the semantic and the causal properties of mental symbols— becomes problematic. In general, it seems to me, informational semantic theories must appeal to some causal properties of the symbols whose semantic properties they try to account for (or to some of the properties in virtue of which such symbols enter causal relations).

On the other hand, it seems to me, a conceptual representation of Xs (say, dogs) must differ from a perceptual representation of a particular X (a particular dog), not only in virtue of its causal properties, but also in virtue of its semantic properties, in virtue of its representational qualities. The perceptual representation will be informationally far richer than the conceptual representation. I do not want to get into the issue of innate concepts, which is a question about the origins —the ontogeny— of concepts. I just want to claim that my concept of a dog must be less informationally profuse than my percept of a particular dog. The former, not the latter, must apply to any number of dogs. Being more or less informationally profuse is a semantic property of a symbol.

Fourthly and lastly, I want to go back to the question I raised at the beginning: What is the difference between thinking about Xs and being able to reason about Xs. Consider again Jerry's contrast between giving priority to concept–possession and giving priority to concept–individuation. On the Contemporary view, on which priority is given to concept–possession, "having a concept is having certain capacities. To have the concept of X is to be able to *recognize* Xs and/or be able to reason about Xs in certain ways". But on the Classical view, on which priority is given to concept–individuation, "having a concept of X is just being able to have thoughts about Xs". As Jerry says of the latter: "the concept DOG... is the concept of dogs; it's the concept that you use to think about dogs with".

But now, on the one hand, pursuing the previous contrast between percept and concept (between a perceptual and a conceptual representation of a dog), there is a sense in which a concept of a dog —a genuinely cognitive structure— must have to do with *recognition* or *identification* of dogs, in a way in which having a perceptual sensory representation of a dog or experiencing a dog does not. After all, some creatures which experience dogs may well lack a dog–concept; and the difference must reside inter alia in what they may or may not do with their dog–representation —in particular, in whether they may or not categorize or sort particular dogs as members of the category of dogs.

On the other hand, the difference between using a concept of X to think about Xs (which, on Jerry's view, is OK) and being able to recognize Xs and reason about Xs (which is not OK) seems to me to be vanishingly small. Any account of using a concept of X to think about Xs will, I surmise, come very close indeed to recognizing Xs and reasoning about them. Again the bridge between semantic and causal properties of concepts seems narrow indeed, as does the contrast between concept–individuation and concept–possession.[1]

---

[1]I am grateful to Nenad Miscevic for conversation on the topic of this paper. I am grateful to Dan Sperber and especially Paul Horwich for very helpful criticisms.

PHILOSOPHICAL ISSUES, 6
Content, 1995

Reference from the First Person Perspective

Brian Loar

In this paper I wish to address two questions about *reference* that are among the most fundamental issues in the theory of meaning and intentionality. They are 1) what makes different external (e.g. causal) relations count as semantic relations, count as reference; and 2) whether reference is objectively indeterminate or inscrutable and whether a positive answer subverts our commonsense conceptions of semantic facts. It appears to me that the two matters are deeply connected, and I will propose a more or less simple idea that permits a unifying answer.

1 What makes a causal relation a semantic relation?

First issue. Suppose we accept that reference is determined by causal and other externally constituted relations. There are various ways in which an individual's thoughts and concepts, i.e. can refer to, be about, pick out, a given thing or property etc. And these ways appear to be correlated with what are intuitively psychologically distinct kinds of concept. Here are some examples.

Concept category	*External reference relation*
Visual demonstratives that tree (which one sees)	An optical relation
Visual–memory demonstratives that tree (which one remembers seeing)	A memory–cum–optical relation
Recognitional–type–demonstrative that kind of thing (see below)	A discriminative ability–cum–memory relation
Socially–deferential names Julius Caesar, Richard Wagner	A social–causal–historical relation to a particular
Socially–deferential kind–terms aluminum, arthritis	A social–cum–? relation to a kind or property

It is widely supposed that what makes these relations reference relations is their role in determining the *truth–conditions* of the thoughts in which the corresponding concepts occur. But it is possible to be unhappy with this. It is difficult to believe that there is a naturalistic explanation of non–deflationary, robust, truth–conditions that does not rely on reference relations that are (partially) independently constituted. It appears to me that the widespread view that the notion of truth–conditions is somehow prior to, or more basic than the notion of reference is not well motivated. Granted that, if there were no such things as thoughts and utterances with truth–conditions, the reference relation would have no instances. But it does not follow that the notion of reference is asymmetrically derivative from the notion of truth–conditions. As I will try to show, there is an answer to the question of what those relations have in common that is of quite a different sort, and that (it seems to me) throws light not just on the notion of reference but also on the semantic notion of truth–conditions.

2 The Objective Indeterminacy of Reference

The causal–external relations that vaguely come to mind do not uniquely determine reference. Suppose O is the relation of belonging to a causal chain that initiates the occurrence of a visual–demonstrative concept. e.g. "that tree". O does not single out a tree, by contrast with a retinal image, a stream of photons, a tree surface, a tree–trunk. This underdetermination is masked from uncritical common sense; for demonstratives have implicit *qualifying concepts*, e.g.

'tree' (or perhaps 'detached object'). The demonstrative thought conceives its referent as a tree, and the referent is the tree to which the concept bears the relation O, and not the tree–surface etc. Commonsensically, a visual demonstrative concept's referent is the object that falls under the qualifier and to which the concept stands in that optical relation.

But of course the demonstrative concept refers to a tree only given that 'tree' refers to the kind *tree* and not the kind *tree–surface*. And the question of inscrutability arises also about the determinacy of kind–reference. Consider Eve, who fashions an idea of trees in general from her pattern–recognitional groupings of objects in Eden. Whether her concept 'thing of that kind' refers to the kind *tree* or the kind *tree surface* depends on whether Eve counts detached objects or surfaces as 'things of that kind'. But this in turn depends on her singular demonstrative references. For what grounds her ascriptions of 'tree' or 'thing of that kind' to a certain group of objects are her judgments 'this is one, that is one too, and here's another...'. If we wish to assign reference to Eve's concept 'tree', we have to rely on the reference of her singular demonstratives.

We are then caught in a circle of interpretation. (This is not, essentially, an epistemological circle.) The determinacy of singular demonstrative reference depends on the determinacy of the reference of the kind–concepts that serve as qualifiers of the singular demonstratives, and vice versa. To take the idea a step further, notice that, if the reference of certain socially deferential names and kind–terms depends on the demonstrative references of individuals who establish usage, the reference of socially deferential names and terms is similarly indeterminate. And resolving the indeterminacy of reference for them means breaking out of the more basic circle of interpretation.

It is instructive to consider a certain stipulative resolution of reference indeterminacy for visual demonstrative concepts. (I make these demonstrative concepts central in this paper not only to have a central example, but also because, it seems to me, a theory of reference for demonstratives must be regarded as quite fundamental in the theory of reference.) The stipulation of a reference scheme, from among various other reference schemes that determine the desired truth conditions, might begin like this. Suppose that D is an externally determined relation that maps qualifying concepts —predicates such as 'tree', 'detached object', 'object–surface'— onto certain sets, e.g. the set of trees, of detached objects, of object surfaces. We then stipulate that D counts as *denotation* (in the reference scheme we are constructing.) Now the reference of a visual demonstrative depends not just on its qualifier but also on the unrestricted optical

relation O: the reference of a demonstrative concept with qualifier Q is determined by restricting O to objects in $D(Q)$. We might think of O as the reference–anchoring relation for visual demonstratives.

So given that stipulation of D as denotation, the "optical relation" mentioned at the outset is the restricted optical relation O^*:

> (where x is a concept–token and y an object) O^*xy iff Oxy and, if Q is x's qualifier, y belongs to $D(Q)$.

O^* determines reference for visual demonstrative concepts. The effect of this is equivalent to having stipulated that O^* —the restricted optical relation— counts as reference for visual demonstrative concepts.

There are various ways to think about indeterminacy, but the foregoing structure gives us a fairly straightforward way. The question can be put like this: what makes the relation O^* a *reference* relation —rather than some competing relation O^{**} that is defined from O together with a different function Ref′? Or, more or less equivalently, what makes the relation D rather than some other relation D' a *denotation* relation?

The stipulative answer says: these are the relations we count as reference or as denotation. The point of picking this relation rather than that relation (actually, this overall reference scheme rather than that one) is that we ought to fix on one, to account for the compositionality of our *independently* motivated assignments of truth–conditions. Or, at least, that is the stipulative account I have in mind.[1]

But there are reasonable worries about the point of such a stipulative solution. Here are two that come to mind. (i) Why do we find the causal relation O important enough that we have bothered to promote one of its restrictions —O^*— as (so to speak) the designated representative of the set of O–restrictions? This is in effect to ask why we want a notion of reference at all. Let me consider two answers. (ia) The first, and more familiar, is that substantive reference relations are needed to explain the compositional determination of truth–conditions (where the latter are themselves held not to be conventional or stipulated facts.) This has been supposed by philosophers who have held reference to be inscrutable, or to be (what is not so different) scrutable by convention.[2] But if reference is merely stipulated, it is obscure how it can explain something with a more

---

[1]This bears some resemblance to Hartry Field's conventionalist solution to referential inscrutability, in Field 1975.

[2]Field 1975, Davidson 1979.

secure standing, viz. truth–conditional compositionality. Perhaps it will be said that what does the explaining is collectively the set of all reference–schemes that generate those truth–conditions (where each restriction of O belongs to one such reference–scheme.) But this undercuts the idea that one of those reference schemes —and its constituent O^*— is to be awarded a special status. We are left with the rather disappointing idea that one of those schemes is designated to *represent* the explanatory role of the collective.

The more appealing idea is that some factor in the determination of reference is independent of its role in determining truth–conditions. Truth–conditional compositionality will then depend asymmetrically on reference in this strong sense: it is a fact that needs explaining only if constituted (in part) constituted by relations whose semantic status is independent of that constitution.[3]ut it is far from evident that a compositional explanation of language–understanding must be truth–conditional. Other explanations, that involve translation rather than truth–theory, appeal to rules —for mapping others' sentences onto our sentences— that are in their way compositional but not truth–theoretic. For an argument of this sort see Schiffer 198, chapter 7.

(ib) A second reason for wanting a notion of reference might be this: those causal–informational relations that we count as reference relations play a crucial role in explaining the *reliability* of beliefs and utterances (relative to a standard assignment of truth–conditions).[4] Yes. But while this may eliminate some truth–conditionally adequate reference schemes, it does not eliminate "gavagai"–type alternative reference schemes.[5] And as regards that restricted class of schemes, something similar to the point above, about compositionality, applies: explaining reliability in terms of informational–causal relations does not require stipulating a special role for one among the adequate (gavagai–type) schemes.

Indeed a more fundamental point threatens any such motivation

---

[3]Understanding the novel utterances of others is widely held to require compositional explanation; and it may be said that this requires assuming that some single reference scheme plays an implicit conceptual or computational role in that explanation. B

[4]For this point see Loar 1981. Truth–conditions are construed loose–grainedly, so that different overall reference schemes are consistent with a single assignment of truth conditions.

[5]In the pages mentioned in footnote 5, I acknowledge this point, although I go on to attempt to minimize the significance of the gavagai alternatives. In the present paper I take this more seriously, in part because of the point about the first person that follows.

for a stipulative determination of reference. Reliability is relative to a specific assignment of (loose–grained) truth–conditions. But it is not implausible that quite different such truth–condition assignments permit similar degrees of reliability. The question then of course is what —apart from an indepenent and non–arbitrary selection of reference schemes— makes a given assignment of truth–conditions privileged. In the face of a striking absence of candidates, we have a strong incentive to find such a non–arbitrary factor in reference–determination.

(ii) Even if we accepted that third person reference ascriptions are determinate only by selective stipulation, that would hardly fit first person ascriptions. When I judge that a current visual–demonstrative thought of my own is about a certain tree, other interpretations are not optional —there is no intuitive scope for arbitrary selection. And this suggests that the first person perspective on reference is special. When we consider the references of our own thoughts, it appears bizarre to suggest they are established as such by stipulation.

3 Disquotation as a Constraint on Reference

Certain deflationary accounts of reference give a special role to the first person.[6] On such a theory to say that

> "George Sand" names George Sand, or "Tree" denotes trees, or
> The next occurrence of "this hand" refers to this hand,

is to say no more than is captured by the disquotational structure of those ascriptions. And disquotation can be granted a special status in determining reference only for the terms of one's current language.[7] (It is conceptually open whether, on another person's use, "George Sand" names George Sand.)

There is something bizarre, as we have remarked, about finding indeterminacy in the first person. However disconcerting and disruptive of ordinary assumptions the idea of indeterminacy in the third person is, we at least find intuitive room for entertaining the possibility when it is forced on our attention. And this asymmetry between the two perspectives has of course a nice explanation on a disquotational–deflationary theory of reference and truth. On

---

[6] See Field 1994.

[7] As Field puts it, disquotational truth and reference is defined only for expressions that one understands.

a deflationary account, third–person ascriptions of reference of the form

> '"*N*" as she uses it refers to *t*'

are explained thus:

> "*N*" as she uses it is translated by "*t*" as I use it and "*t*" refers (disquotationally) to *t*.

There is room for indeterminacy in third–person ascriptions of reference if there is room for indeterminacy in term–term translation.

Deflationism may then count the intuitive asymmetry as support. It is not on our present agenda, however, to consider the merits of deflationism. We are working within the non–deflationist assumption that reference is determined by various contingent external relations of the kinds mentioned above. Now, the thesis that first–person disquotational ascriptions have a special status is distinct from the deflationary thesis that 'refers' is a pure disquotational notion. And it is quite compatible with the non–deflationist assumption that disquotational ascriptions of reference have a privileged status,[8] and that such ascriptions —or something akin— play a special and non–arbitrary role in determining which of various external relations qualify as reference. Still, elaborating this role is not as straightforward as one might casually suppose.

Here is a first try. Deflationary theories may be interpreted as saying that the set of term–object pairs that instantiate reference for my language can be laid out, for each concept category, as follows (to take proper names as examples):

> ⟨"George Sand", George Sand⟩, ⟨"Cary Grant", Cary Grant⟩, ⟨"Lisbon", Lisbon⟩, etc.

Suppose that, for the proper names in my language, some robust relation R holds of all and only such pairs, open–endedly. The proposal then is that R counts as a semantic relation: it determines *reference*; and likewise for each relation that satisfies the disquotational condition for each other concept–category. Suppose that, in our (pre–philosophical) counterfactual judgments about R and about reference, we can find no counterexample to the thesis that

> for any person P, if N is used as a proper name by P, then N bears R to x iff N refers to x in P's language/thoughts,

---

[8]For an extended discussion of puzzles about this idea, see Loar 1994.

and similarly for any relations that match the disquotational configuration for other concept–categories. We have then pointed to R by way of its holding of (as we might say) disquotational pairs, and then have verified that, given one's ordinary notion of reference as applied to the languages and thoughts of others, R coincides with reference for the proper names of others. (This of course allows that R holds for names used by others but not by me.) Similarly for other concept–categories.

Might we not then reasonably claim to have explained what makes a robust external relation a semantic relation —makes it *reference*— by essential appeal to the disquotational pattern? And would we not then have explained our intuitive conception of the determinacy of robust reference relations? The idea's attraction lies in this. The disquotationality of the home case seems an unchallengable constraint on reference, indeed constitutive in part of the very idea of reference. If the disquotational generation of term–object pairs suffices, given external contingencies, to pick out relations that fit intuitions about reference in third–person cases, we have found a single feature —home case disquotationality— that is necessary and sufficient for otherwise disparate external relations to count as semantic.[9] And this feature satisfies our further desideratum, a constraint on reference that is independent of the role of reference in determining truth–conditions.

A reference relation is a certain relation that applies to disquotational pairs. That is the idea; but does it makes sense? There is an elementary problem. No disquotational pattern is to be found among pairs of terms and *objects*. You do not produce the woman George Sand by peeling away quotes from "George Sand". No disquotational principle unifies any set of concept–object pairs, or explains what later pairs have in common with earlier pairs. I cannot then point to a relation between terms and objects by virtue of its holding of certain disquotational term–object pairs, it seems; for there are no such pairs.

The disquotational pattern is of course found among term–term or concept–concept pairs: ⟨"'George Sand'", 'George Sand'⟩, ⟨"'Lisbon'", 'Lisbon'⟩,.... Now this is quite an important fact, for while it does not directly constrain reference relations, there is no reason why it should not explain what makes a relation–concept a semantic concept. We will return to this. Our present question,

---

[9]There is the awkward possibility that other people have concept–categories different from ours. We would not then have explained why any given relation between terms of those categories and things/properties should count as reference.

however, concerns relations and not relation–concepts. And those term–term pairs cannot count as determining reference relations indirectly. Call a term–object pair $\langle$'N', $x\rangle$ "disquotational" just in case it is picked out by a pair of (one's own) terms $\langle$"'N'", 'N'$\rangle$. Suppose one then says that R is a reference relation if R holds of disquotational pairs. (Somehow open–endedly?) But that of course won't advance our project, for it appeals to a semantic relation ("picked out by") whose determinacy, and constitution as semantic, are what we seek to explain.[10]

Our problem in explaining how disquotation can be a constraint on reference is then this. A disquotational constraint on term–object pairs or term–object relations makes no sense. And while a disquotational constraint on term–term pairs and term–term relations does make sense, it cannot be counted as constraining term–object reference relations without presupposing semantic relations.

4 Disquotational term–object pairs from the subjective perspective. The interplay of subjective and objective perspectives

There is a way around the problem, but it involves a shift of perspective. It does appear to me that the idea of "disquotational" term–object pairs is not completely hopeless. But we have to look at the matter phenomenologically, with an eye to how things appear in the first person.

Visual demonstratives again. Produce, as realistically as you can, pairs of concepts in the following pattern,

(V) the next occurrence of 'that tree', that tree
the next occurrence of 'that hand', that hand
the next occurrence of 'that chair', that chair
...,

where the second member of each pair is a visual–demonstrative concept. Now when I say produce those pairs, I mean *exercise* those concepts. Use them to think with, and don't (merely) mention them. What is the upshot? From your perspective —that is, the perspective of the user of those concept pairs— pairs of visual demonstrative concepts and certain objects appear on the scene. (From the third–

---

[10]Compare: O^* is a reference relation, because O^* and not O^{**} is the one I pick out in saying 'refers'.

person or objective perspective —which one also can take towards oneself— pairs of concepts are being exercised.) How are those concepts and objects apparently linked? This is a delicate matter. I am tempted to say this: they are linked *in that characteristic way*, (which as it happens is consequent upon exercises of those concept–pairs above). If, from our first–person perspective, we count concepts and objects as linked in a "disquotational" way, we make *some* sort of sense. That idea is basic in the present account. Before returning to the question what sort of sense we might thereby make, let us be clear how this would serve our project.

We seek a constraint on concept–object relations that makes them count as semantic relations, as reference. The first step is that certain concept–object pairs may be counted as suitably disquotational from a phenomenological, subjective, or first–person perspective. So we lay on the table, for further investigation, as many such concept–object pairs as we can. Now here's a key point. Nothing prevents our shifting our perspective back to the objective perspective at this point, keeping those pairs in mind while considering them afresh from that third–person perspective. Does some well–carved external, objective, relation hold of them, in some projectible way? Presumably it is natural to think there is such a relation, the one we have been calling O^*, even though we do not know its scientific details. But we apprehend that relation as having a special status. It is as if O^* is latched onto by the exercise of those concepts; for —as the phenomenology has it— that conceptual exercise generates those concept–object pairs. It moreover appears to do so open–endedly, so that the projectibility of that concept–object "generation" matches the relation's open–ended application. That relation can then hardly be seen, from the phenomenological perspective, except as having an intimate connection with those concepts. That is why O^* counts as semantic, as a reference relation. A certain primitive idea of what a thought is *about*, it is natural to say, derives from concepts' appearing to bring objects onto the subjective scene, and the corresponding optical relation O^* is privileged by its intimate connection with this phenomenon.[11]

Back to the plausibility of construing the felt link between those concepts and objects phenomenologically —"related in that familiar

---

[11] This primitive aboutness relation appears to be, in its way, "disquotational". Should we not then count reference as diquotational, at bottom so to speak?

This would be too quick. There are good reasons to count reference as constituted by externally determined relations. Not the least of them is the phenomenon of counterfactual reference shift (as we might call it).

way". The suspicion is that the link is just the old visual–demonstrative reference relation, that causal relation, somewhat obscured by the vagueness of the phenomenological attitude. But the suspicion is, I think, not borne out. In your conception of those concept–object pairs, you use no semantic or intentional notions; you use the concept–concept pairs (quotation concept and visual–demonstrative concept) without semantic commentary. Having put the concept–object pairs on the table, you *may* then of course go on to think, self–consciously: I have used that concept–token 'that tree' to refer to that tree. But this does not appear to be the inevitable way in which one intuitively discerns a link between the concept and the tree. For it seems quite natural that when such a concept–object pair is thus displayed —without explicit semantic commentary— it will appear to have a characteristic configuration, a distinctive feel, as if the object is peculiarly *salient* in the company of that concept. This apparent link between concept and object is hardly magic. We might think of it as a straight *projection* of the intimate link between disquotation concept and visual–demonstrative concept onto the corresponding concept–object pair, the latter being brought onto the scene by the exercise of the former concepts. The appearance of such a common linkage in those concept–object pairs seems quite expectable. And the phenomenon is registered without the use of explicitly semantic concepts: the apprehension of the object's salience in the presence of the concept is far more primitive and elementary than characterizing their relation in causal–referential or intentional terms.

Describing the special concept–object link as "apparent" is apt, for it is an illusion: no objective relation —"x is salient in the presence of concept C"— between object and concept is created by the mere exercise of the two concepts. (But recognizing this does not destabilize the phenomenology; the illusion is robust.) What matters is that the name–object pairs we lay on the table are subjectively privileged, their status explainable as a projection of the exercise of disquotionally related concepts. This of course permits proceeding to the further stage of investigation, when we shift perspectives, and see that those concepts and objects are objectively related after all, independently of the phenomenology.

5 Appearances

In the foregoing phenomenological description we appealed to appearances. This raises the question whether we might be presuppos-

ing referential determinacy in the context "it appears that...". That that could be a problem may be seen by considering the following, different, proposal.

> It appears to me that, by the following occurrence of 'that tree', I pick out that tree (rather than that tree–surface etc.) It is of course an illusion that some privileged objective relation holds between that term and that tree and not between that term and that tree–surface. But still we may say that as the result of the appearance of that connection, the former term–object pair has a special status that the latter does not.

Now, interpreted as an objective statement about appearances, this of course does not work. The construction "it appears to me that... that tree..." is a *de re* propositional attitude ascription, and the question of inscrutability arises as much for such constructions as for ascriptions of reference. We cannot assume that that appearance–configuration objectively relates me and the tree and not also me and the tree surface.

But the phenomenological remarks do not make essential use of the construction "it appears that...", do not depend on (as it were) the objective ascription of appearances. We capture appearances also by conveying what a certain experience is like (e.g. by saying "it is as if..."), and the appearance language in which we convey this does not require objective construal. In phenomenological description anything goes, and literality is mostly a barrier.

It is not part of the proposal that those pairs are linked by a certain objective fact about how things appear. No special objective relation, of any sort, is being asserted to single out those pairs. When we shift from the phenomenological perspective, the pairs that we have placed on the table are not special, from the new objective perspective. Our objective investigation can be seen as having a significant semantic upshot only from the subjective perspective. And so, success in advancing this present proposal requires the reader's phenomenological cooperation. The apparent concept–object linkages are (I venture to suggest) something recognized in one's own experience, when one exercises the corresponding concept–concept pairs.[12]

---

[12]This presupposes that we can understand or conceive our conceptual and experiential situations independently of that–clauses; evidently I am committed to that. For an account of such conceptions see my "How to Conceive Mental Content", forthcoming.

6 What makes a relation a reference relation?

There are two ways to take what we have said so far, namely, as a proposal about what makes a relation like O^* a reference relation, and as a proposal about what makes a relation–concept a reference–concept. As a proposal of the first kind we have this: O^* is a reference relation because it holds of pairs linked *in that way*. Here we have to suppose that "in that way" applies across concept–categories. This means that pairs such as the following have an intuitive phenomenological similarity: ⟨the next occurrence of 'that tree', that tree⟩, ⟨'Plato', Plato⟩, ⟨'copper', copper⟩ and so on. Evidently this makes sense only when one exercises those concepts for oneself. Given that the similarity among those pairs is not objective —is merely subjective, then so is what makes a relation a reference relation. But we have a reasonable understanding of what that is.

7 What makes a relation–concept a reference–concept?

We have a conception of the interesting relation that holds between visual–demonstrative concepts and objects, and it has the form "the relation that holds of this pair —⟨the next occurrence of 'that tree', that tree⟩— and that pair —⟨the next occurrence of 'that hand', that hand⟩— and this other pair...". That conception incorporates concept–pairs that exhibit a "disquotational" pattern. Call them *mention–use* pairs of concepts. Here we have an appropriate combination of subjective and objective factors: the pairs are those presented as noted, and the relation we conceive as objectively holding of those pairs. And what makes this commonsense conception a reference–relation–conception is the role those disquotational concept–pairs play in that complex descriptive concept.

But this is somewhat disappointing. I appear to have a more general conception of the optical reference relation than such a description, which mentions specific demonstrative concepts and objects. My conception does not change from day to day according to the sightings I keep in mind. It is rather an open–ended conception, not tied to particular references. What I want to propose is that this concept is in its way demonstrative —"that relation", and that in a certain way the disquotational, or mention–use, configuration constitutes that demonstrative concept's defining perspective.

8 Recognitional concepts and perspectives

We take a digression. Some of our concepts are, as we might say, recognitional type–demonstratives, or more briefly *recognitional concepts*. Here is an example. In East Africa you see certain four–legged creatures up close. Not knowing their name, you yet quickly learn to identify others of "that kind", as you think of it (a kind of gazelle in fact). It is not that you then re–identify the kind by remembering the earlier individuals one by one. You may not keep the individuals in mind, having formed a generalized type–memory–demonstrative, as well as an inclination to identify further things as of the same kind. Suppose you also then spot some creatures wandering on the far slope across the valley, with distinctive gait and distant appearance. You form another demonstrative conception, of creatures of that kind. Then you raise the question whether this kind = that kind. Let us say that they are of the same kind, although for you it is an open question. You have two independent recognitional concepts with the same reference. They differ *perspectivally*, by which I mean, not in the external relations that hold between them and the two groups of creatures, but in some psychological or subjective perspectival aspect. So there are type–demonstrative recognitional concepts that are individuated at least in part perspectivally.[13]

9 Conceiving relations from the mention–use perspective

This notion of recognitional concepts can account (if we take an abstract enough view of them) for the above mentioned demonstrative conception of the visual–demonstrative relation, as well as for demonstrative conceptions of those relations that constitute reference for other concept–categories. We can account, moreover, for what makes those relation–conceptions *semantic* or *intentional* conceptions. Proposal:

> One's demonstrative conception of the visual–demonstrative relation is a higher–order recognitional concept, which purports to pick out a certain external relation from the mention–use perspective. ("Higher–order" because concepts are in the domain of the conceived relation.)

---

[13]For a more extended discussion of recognitional concepts, see Loar 1990.

My conception of the visual–demonstrative relation appeals to no particular array of mention–use pairs. It does not change over time, as it would if it involved reference to the particular mention–use pairs one might have in mind at the moment. Compare the above recognitional concepts of African animals; they need involve no memories of particular animals. Those concepts' *type–demonstrative* status depends on past and potential discriminations of animals of a certain kind; they point in a certain characteristic way to certain kinds. But they do not point descriptively via specific reference to particulars.

Similarly for one's demonstrative conception of the visual–demonstrative relation. From the perspective from which we exercise such conceptions of concept–object pairs in the mention–use configuration, the pairs themselves are phenomenologically presented as from a distinctive perspective. It is as if there is a rule that generates, open–endedly, the concept–object pairs themselves —even though we know reflectively that the rule governs not the pairs themselves but our conceptions of them. This merely subjective perspective on concept–object pairs is phenomenologically stable, and from it we are able to conceive *that* objective external relation (as we judge it), which holds of this pair and that pair etc. It is in this sense that we may count our conception of the visual–demonstrative relation as a recognitional concept that is formed from an essentially first–person perspective. I retain, from day to day, the same demonstrative conception of the visual demonstrative relation, even though the instances I have in mind change. This parallels the fact that I retain the same recognitional concept of a certain unnamed kind flower that I see around town, despite the change in its instances.

We want the relation thus demonstrated to be the one we referred to earlier as O^*, for otherwise we have not advanced. But this raises the question why we should regard that recognitional concept, formed from the mention–use perspective on visual–demonstrative concepts, as discriminating O^* rather than say O itself. (Recall that O holds not just between the visual demonstrative concept 'this tree' and a certain tree, but also between that concept and a certain tree–surface.) Granted that the pairs one picks out are all instances of the restricted relation O^*; they are also all instances of the broader relation O.

The wrong answer is this: if we have a conception of a concept–object pair that does not a disquotational or mention–use conception of that pair, then our recognitional concept picks out relation R only if R does not hold of that pair. That excludes too much: for there are many ways to conceive of concept–object pairs that are in

the extension of O^*. We are not restricted to conceiving of them disquotationally, in the mention–use configuration.

Back to the phenomenological concept–object link. Our recognitional concept 'that relation' is sensitive to pairs that present themselves as linked thus. So, if there is no way of conceiving of a concept–object pair that engenders that apparent linkage, then that pair is not in the extension of the relation discriminated as 'that relation'. Indeed, this implies that the relation picked out from the mention–use perspective must be a many–one relation, a function. Then we can see that the pair ⟨'that tree', that tree–surface⟩ is actively ruled out. The reason is fairly obvious. The the tree–surface is not identical with the tree. But our conception from the mention–use perspective 'that relation' is a conception of a function, and given that function has in its extension the pair ⟨'that tree', that tree⟩, the pair ⟨'that tree', that tree–surface⟩ is not in its extension. In other words: if the apparent rule picks out $\langle c, a\rangle$ and one judges that a is not identical with b, then the rule rejects $\langle c, b\rangle$.[14] A more direct route to the same conclusion is this: it is clear from the pattern of our discriminations that we conceive of the demonstrated relation as a function, and O^* is a function and O is not.

What then makes a relation–conception a semantic conception? The recognitional concept 'that relation' is a semantic–relation–concept just in case it is related as described to the mention–use perspective. I will consider below the significance of this subjective or first–person nature of semantic concepts.

10 Factoring O^*

As we saw earlier, we cannot simply declare that O^* determines reference for visual–demonstrative concepts and be done with it. We have shown why O^* and not some other relation determines reference for visual demonstrative concepts. But we have not explained the structure of O^*'s reference determination. And this matters.

A deflationist may take a hard line, asserting that all that matters in determining reference, for expressions simple and complex, is the disquotational or mention–use configuration, and that the standard

---

[14] Note well that this does not imply that, if the rule picks out $\langle c, a\rangle$ and fails to pick out $\langle c, b\rangle$ —i.e. thus conceived, then the rule actively rejects $\langle c, b\rangle$. For one may not know whether $a = b$.

assumption that semantic structure plays an explanatory role in the determination of reference is a mistake.[15]

On a non–deflationary view, though, it appears undeniable that the reference of visual demonstratives is determined by two factors: the reference of the qualifying concept, and the reference–anchoring relation O. But this is all right. Once we have motivated counting O^* as reference, a semantic factoring of O^* is straightforwardly motivated. For two simple facts dovetail. a) It is a fact from the third–person perspective that in any large array of positive and negative instances this pattern obtains: whenever x and y stand in the relation O^* they stand in the relation O, and the restriction to O^* fits exactly a certain straightforward mapping D of concept–types to sets of objects. b) From the first person perspective, all instances of the mapping D are validated in the mention–use configuration. We have, that is:

"'tree' denotes trees",
"physical object" denotes physical objects,
"surface" denotes surfaces,

and so throughout all qualifying concepts. Thus there is a convergence: D emerges from a straightforwardly factoring of O^*'s instances, and D also emerges from adopting the mention–use perspective on the referential qualifiers of visual demonstratives. This would appear to vindicate the non–deflationists' regarding the reference of visual demonstratives in two ways: as determined by O^* (picked out from the mention–use perspective), and as determined compositionally by the joint action of D and O, the reference of the qualifier and the visual–demonstrative relation. There is no conflict between the two perspectives on reference.

11 Reference for concepts of other categories

Let us consider the socially–deferential name–relation, and how it is conceived demonstratively. This is the relation that holds, for

[15]Regarding semantic structure, a deflationist could say this. Explaining the generative or open–ended nature of our capacities to produce sentences in thought and in communication does require some sort of compositional structure. But that structure explains conceptual roles, or dispositions of "use", and not reference or truth–conditions. (Cf. Schiffer on compositionality, in 1987, chapter 7.)

The reference of syntactically complex expressions does of course exhibit certain semantic structure. But that is just a matter of patterns in reference that are consequential upon the primary determination of reference, which on a deflationist theory is disquotational.

instance, of my proper name 'George Sand' and the woman George Sand. Speaking of my conception of that relation as a recognitional concept could be puzzling. The relevant concept–object pairs can be widely separated in time. And the relation is complex and difficult to specify, a compound of my relation to a certain group of speakers and their relation —by way of a causal–historical chain as they say— to George Sand. I cannot take all that in so to speak at a glance. But still there is a sense in which I have a recognitional concept of that relation, and it is rather modest. Once again, there is a mention–use configuration, in this case involving proper names in thought. And this generates an open–ended way of conceiving of concept–object pairs: ⟨"Cary Grant", Cary Grant⟩, ⟨"Lisbon", Lisbon⟩ and so on indefinitely. From the point of view from which one conceives those pairs in that way, one can think of *that* externally determined relation, with the open–ended concept–pair–generating disposition guiding the demonstrative. If some externally determined relation is in fact thereby singled out —as we might note from a combined first–person/third–person perspective— that relation is then the referent of "that relation". It is in this sense that we have a recognitional concept of the socially–deferential name relation.[16]

---

[16]Evidently we conceive of the socially deferential name relation in somewhat more analytical detail. It is the product of two contingent relations, one that holds between me and other people, and the other that holds between them and some person, place, building, object. (We can envisage a further empirical investigation into the details of these relations.) It could seem that, if I can analyze the socially deferential name relation into those two components, and if each of those components is conceived in a way that is not essentially "first–person", then my conception of that relation would, unlike my conception of the visual–memory relation, not be essentially from a first–person perspective. And that would contradict the account below of the unity of semantic concepts. But it is not obvious that we have such a third–person analysis. For the fact that one conceives of the socially–deferential name relation as the product of two such relations hardly implies that one can conceive those relations independently of one's conception of the composite relation. In fact we have a way of conceiving the composite relation —viz. from the open–ended mention–use perspective on socially deferential proper names. Perhaps the first of the component two relations— the one that holds between me and some group to whose usage I defer —can be seen as a sort of abstraction from the demonstratively conceived composite relation. And as for the second —the relation that holds between a population, a name and the name's referent— it is hardly obvious that we have a conception of this relation that does not depend on a conception of individuals' using names socially–deferentially. (The apparent circularity is intrinsic to our conception of social meaning; this is why social meaning appears not to be explicable in terms of individual meaning.) In any event, it is apparently an elementary fact that we have a first–person, mention–use, conception of the socially–deferential relation.

12 Objective vs subjective determinacy of reference

Finally we turn to the bearing this account has on the determinacy of reference and truth–conditions. It appears quite likely to me that there is no resolution of the indeterminacy problem in objective terms, and hence that the Quine view is in a certain straightforward way correct. But it seems to me also that the objective indeterminacy of semantic properties does not have the eliminative, subversive consequences it is usually seen to have. For, our ordinary conception of semantic relations are in part subjective; that is, it is essential to those conceptions that they conceive the relations they pick out from the mention–use configuration. From the subjective perspective, reference is determinate, while realized by external objective relations that are themselves, as Quine points out, not unique in their objective roles.

The import of these remarks will be clearer if we return to the comparison of our familiar O^* and another relation O^{**}. The latter maps "that tree" onto not a tree but a tree–surface. The correlation of the conceptual role of "that tree" with the class of trees as opposed to the class of tree–surfaces will from a third person perspective appear arbitrary, to whoever finds the basic Quinean inscrutability argument plausible. And then of course no objective requirements on reference selects between O^* and O^{**}. Evidently this is where the subjective perspective matters; for it is O^* and not the relation O^{**} that is discriminated from the mention–use perspective on visual demonstrative concepts. And of course we must understand "discriminated" itself subjectively.

Earlier we remarked that, from the mention–use perspective on proper names, it is *as if* there is a rule that generates an open–ended series of term–object pairs. And this could easily be taken to mean that the intuitive determinacy of reference is based on an illusion. It is a short step from this to an eliminative view of semantic relations, that is, to conceding to Quine the main point.

The issue is delicate. There is a straightforward sense in which, it appears to me, Quine ought to be conceded one of the main points, namely that there is no principled objective way of marking off O^* from O^{**}. That is eliminative of a *philosophical* supposition that reference is objectively determinate. But is it eliminative of the

---

Suppose we also have an independent third–person conception of that relation. Then we can say that the two conceptions converge in what they pick out.

intuitive conceptions that one brings to one's initial philosophical reflections about the nature of semantic facts? This is far from obvious. In reflecting on the import of what I have been suggesting, I find that my basic intuitions appear to be left in place. There is indeed an objective relation between my visual demonstrative concept "that tree" and trees that does not obtain between that concept and tree–surfaces —namely, the relation O^*. The question is whether the present account of what makes O^* a semantic relation is subversive or eliminative. I cannot see that the answer proposed in any obvious way overthrows my intuitive perspective on the nature of thought and the intentionality of thoughts. On the contrary. It seems rather intuitively clear that the first person perspective on the references of my own thoughts is somehow constitutive of my conception of reference, of intentionality, of "aboutness". The present proposal explains how this can be so even though reference is externally determined.[17]

REFERENCES

Davidson, Donald 1979. "The inscrutability of reference". *The Southwestern Journal of Philosophy* 10.

Field, Hartry 1975. "Conventionalism and instrumentalism in semantics". *Nous* 9, 375-406.

Field, Hartry 1994.

Loar, Brian 1981. *Mind and Meaning*, Cambridge University Press.

——— 1990. "Personal References", in *Information, Semantics and Epistemology*, ed. Enrique Villanueva, Blackwell, 1990.

——— 1994. "Self–interpretation and the Constitution of Reference", *Philosophical Perspectives* 8.

Schiffer, Stephen 1987. *Remnants of Meaning*, Bradford Books/MIT Press.

---

[17]Thanks are due to Paul Horwich and David Sosa, for their penetrating and helpful remarks on this paper at the Lisbon meetings. Thanks also for comments by Jaegwon Kim and Paul Boghossian, and for very useful extended comments by Barry Loewer, which got me to reformulate a number of points.

PHILOSOPHICAL ISSUES, 6
Content, 1995

Disquotation and Cause in the Theory of Reference

Paul Horwich

In the last twenty–five years or so discussion of *reference* has been focussed on the relative merits of three alternative models. First there is the *description* theory according to which x refers to y when x is associated with a description of y. (This is the Frege–Russell view, championed these days by Searle and Katz). Second there is the *causal* theory according to which x refers to y when there is a certain sort of causal chain relating x and y to one another. (This sort of idea has been promoted by Kripke, Putnam, Evans, early Field, Stampe, Dretske and Fodor; it deserves to be called the mainstream). And third, there is a relatively recent arrival on the scene, the *deflationary* theory (also known as 'minimalism' or 'disquotationalism') according to which x referring to y is roughly a matter of x being the *word* "N" (in quotes) and y being the *thing* N (out of quotes). (This point of view is advocated by the present author and in recent work by Field).

Where does Loar stand in relation to these alternatives? As I see it, what he is proposing is a predominantly causal theory, laced with a dash of disquotationalism. More specifically, his view is that we have a pre–theoretical conception of reference which is tied to the

disquotation schema, and that this conception points us towards the true nature of reference, which turns out to be a collection of causal relations. Let me try to improve this oversimplified characterization of Loar's view by explaining how he arrives at it.

He begins with the intuition that reference relations are objective, external (typically causal) relations betweem mental terms (i.e. concepts) and the things to which they refer. However his view is more liberal, in two respects, than the standard causal theory. For Loar does not maintain that *all* terms (not even all *names*) refer in this way —only terms that belong certain semantic categories (for example, socially–deferential names and recognitional concepts). Moreover he does not maintain that there is a *single* causal relation constituting reference: he supposes that terms falling in different semantic categories will exhibit different objective reference relations.

But this flexibility leaves him with two difficulties. First, for a given category of term, what singles out some particular causal relation as the reference relation for the terms in that category? For example, why should we identify reference with the causal relation between the word "tree" and *trees*, rather than the different one connecting the word "tree" and *tree–surfaces*? (This, he takes to be Quine's issue of the indeterminacy, or inscrutability, of reference). And second, what shared feature of the various causal reference relations —those that constitute reference for the various semantic categories— marks each of them as relations of *reference*? How can this bunch of different objective relations all count as reference?

Loar's simultaneous solution to these two problems is to invoke disquotation. It is in virtue of our possessing a general disquotational conception of reference, and prior to our having identified which causal relations are the reference relations, that we know, for example, that our word "tree" refers to *trees* and not to *tree–surfaces*. Thus our pre–theoretical disquotational conception of reference provides us, firstly, with constraints that allow us to determine, for a given semantic category, which causal relation is its reference relation; and secondly, with an explanation for our grouping together, under the same heading ("reference relations"), the various causal relations that are identified in this way.

Thus we see the ingenious combination of causal and disquotational elements in Loar's position: reference is a causal relation that is identified by means of a disquotational conception. But there remains, he thinks, a further difficulty. Having said that our conception of reference somehow engenders the knowledge that "tree" refers to trees, "London" to London, and so on, it remains to understand just how this is done. We need a characterization of our conception

of reference that will explain exactly how it could be that we are led by that conception to an appreciation of what refers to what, and thereby to the knowledge of what the reference relations are.

This is especially problematic insofar as we are tempted to think of our conception of reference as a *descriptive characterization of that relation.* For it proves to be impossible to find such a thing. It can't simply be

> Reference is the relation between disquotational pairs

since disquotation relates pairs of *expressions*, not expressions and things. (As Loar says, we don't arrive at an object by taking quotation marks away from a word). Nor can our characterization be

> Reference is the relation between the referents of disquotational pairs

which is circular. And nor can it be

> Reference is the relation between the word "London" and London, the word "tree" and trees, and so on

because that would make our conception implausibly dependent upon the particular words we happened to have in mind.

Loar's answer is that our concept of reference is *demonstrative* rather than *descriptive.* When a term/object pair $\langle x, y \rangle$ is conceptualized as an instance of 'mention–use', or disquotation, there is, he says, a particular *phenomenology*, a particular *feel*, which the pair induces. What is going on in my mind when I identify the pair ⟨"London", London⟩ as a case of reference is similar to what is going on when I so identify the pair ⟨"Aristotle", Aristotle⟩ —in both cases I have in mind a pair of terms that are related disquotationally. Thus Loar arrives at the view that reference is what he calls "a recognitional–type–demonstrative"— a concept of the form 'that property again', which is triggered by the peculiar phenomenology associated with disquotation. Once again, the pair of entities, x and y, may be identified in a pair of ways, C and D, that are related disquotationally; this has a special feel to it which triggers a certain concept, 'that relation again'; and this is our concept of reference.

So much for exposition. Now let me say briefly what seems to me to be correct, and what incorrect, in Loar's position. As a deflationist, I like the emphasis on disquotation. I agree that our conception of reference is intimately tied to the disquotation schema. Indeed I would say that our possession of this concept consists in nothing more or less than our disposition to instantiate that schema. I am

not convinced that Loar's *phenomenological* remarks add anything important to this idea; but nor do I see any particular harm in them. On the negative side, however, I doubt that reference has any sort of underlying nature. So I don't agree with Loar's view that our disquotational concept points us toward the true causal nature of reference. Let me end by making three points which seem to me to support a sceptical attitude towads the idea that reference is constituted by some underlying relation or relations.

First: our conception of reference presupposes no such a thing and provides us with no reason to expect it. Suppose our conception does consist in a disposition to accept instances of the disquotation schema. Then this is the fact about the term "refers" that provides it (whether in English, Spanish, or mentalese) with its meaning. And insofar as we are prepared to deploy a liberal notion of 'property' whereby all meaningful predicates express properties, then we have every right to suppose that reference is a genuine relation. However, the question of whether or not this relation is constituted by some causal relation (or by any other objective, external relation) is an entirely separate issue. And the disquotational account of our concept —even given Loar's characterization of it as a recognitional–type–demonstrative— gives no reason to expect that there are any such constituting relations.

Second: deploying our disquotational conception, we can identify an unlimited number of term/object pairs, $\langle x, y \rangle$, such that x refers to y. In this way, as Loar points out, we are able to put ourselves in a position to discover a causal relation, R, such that, for all terms within a given semantic category,

x refers to y *if and only if* xRy

And on the basis of such a discovery, he says, we would be justified in concluding that reference (for that category) is identical to, or constituted by, the relation R. However, I think this final step is mistaken. We indeed *might* (though there is no reason to think we *will*) find some causal relation that is *co–extensive* with reference. But without a great deal of further argument it could not be concluded that reference is *constituted* by R. Consider another example. The basis for our supposing that the property of 'being water' is constituted by the property of 'being composed of H_2O molecules' is not merely that every sample of water is made of H_2O. It is, in addition, that there exists a certain *explanatory* relationship between these properties: namely that the characteristics in virtue of which a certain sample is recognized as water are explained by the fact that it is composed of H_2O molecules. In general, our grounds for supposing

that superficial property S is constituted by underlying property U is that the possession of U explains the characteristics symptomatic of the possession of S. Thus a case for concluding that reference is constituted by R would have to include a specification of other properties that are correlated with and indicators of reference, plus an argument to the effect that the possession of those characteristic symptoms of reference is best explained by the holding of relation R. Therefore the burden of argument that must be met before we can suppose that R underlies reference is more severe that Loar suggests. Not only do we have no reason to believe we will ever discover a relation R that is so much as co–extensive with reference (for a given semantic category); but even that unlikely discovery would be insufficient for the conclusion that reference is constituted by R.

Thirdly: it isn't just a matter of there being no reason to believe that such a case can be made: there is positive reason to believe that it *cannot* be made —there is positive reason to think that the characteristics of reference *could not* be explained in terms of an underlying causal relation. Therefeore, given the previous point, we have a positive reason to think that reference has no underlying nature. This reason (which I'm afraid I can only sketch here) is based on a plausible assumption about the *function* of our concept of reference: namely, that it is a device of semantic ascent —a device enabling us to formulate generalizations of a very special kind that would otherwise call for substitutional quantification. Here is an example. Suppose someone, speaking a language we don't fully understand, attributes a property —say, *redness*— to some object or other; but we don't know which is the thing to which redness is being attributed because we don't understand the singular term, "#", that is being used. How can we report what is said? With no concept of reference, we would have to put the matter as follows:

> If "#" means "a", then a is red, and
>
> if "#" means "b", then b is red, and
>
> if ...

where "a", "b",... are the singular terms of (extended) English. But this is an unstatable infinite list; there is an item for each thing "#" might mean. If, however, we have a term, "refers", satisfying the schematic disquotation principle

> If x means "$*$", then the referent of x is identical to $*$

then that infinite list can be captured in a single statement. For, given the instances of that schema,

If "#" means "a", then the referent of "#" = a, and

if "#" means "b", then the referent of "#" = b, and

if ...

the original list is equivalent to

If "#" means "a", then the referent of "#" is red, and

if "#" means "b", then the referent of "#" is red, and

if ...

whose content (given that one of the antecedents is true) is

The referent of "#" is red

Thus the notion of reference, insofar as it satisfies the disquotational principle, enables to capture certain generalizations that cannot be captured merely by using the usual devices (that is, "all", "every" or the universal objectual quantifier). If this is right, then facts articulated with the concept of reference (including those that correlate reference with other properties) cannot be deduced from ordinary non–semantic generalizations and, therefore, cannot be explained by them. Thus there could be no explanatory gain in supposing that reference has a non–semantic underlying nature. So the conditions for supposing that something constitutes the reference relation cannot be met.

In summary, I see Loar as an advocate of the causal theory who has found that certain problems can be handled by treating it with a dose of disquotationalism. But this could be one of those cases in which the cure is more dangerous than the disease. For the disquotational account of our conception of reference, not only solves the indeterminacy problems, but threatens to leave nothing for a causal theory to explain.

PHILOSOPHICAL ISSUES, 6
Content, 1995

Reference from a Perspective *versus* Reference

David Sosa

Loar is concerned with two questions about reference. *First*: what makes different external (*e.g.* causal) relations count as semantic relations, count as reference? *Second*: if reference is objectively indeterminate or inscrutable, does this subvert our commonsense conception of semantic facts?

The first is a question about the different relations (which he takes to be different kinds of causal relation) that constitute reference. What picks those relations out, from among all the possible causal relations, as the relations that constitute *reference*? What gives them their semantic character? According to Loar, one popular response, that their semantic force derives from their role in determining truth–conditions, does not answer the question. For it would seem that truth–conditions themselves can be understood only in terms of the reference relation. Loar doubts that the notion of reference is asymmetrically derivative from the notion of truth–conditions. He will attempt to offer a different explanation of the semantic character of the various reference relations.

With respect to the second question, we are offered the elementary example of the causal relations in which the concept *that tree* might stand. These relations would not distinguish a particular tree by

contrast with its surface. How is it then that the tree, as opposed to the tree surface, is the referent of the concept *that tree*?

Loar considers a stipulative resolution of referential indeterminacy (or inscrutability). Take visual demonstratives. These demonstratives are said to have implicit qualifying concepts. However, that fact alone is insufficient to make their reference determinate. The reference of the qualifying concepts is itself indeterminate. But we can imagine a relation, D, that "maps qualifying concepts [or qualifying predicates —Loar is concerned equally and indiscriminately with the reference of thought and language] —predicates such as 'tree,' 'detached object,' 'object surface'— onto certain sets, *e.g.* the set of trees, of detached objects, of object surfaces. We then stipulate that D counts as *denotation*..." (p. 55).[1] Now we restrict the relevant optical relation that is involved in visual demonstrative reference by requiring that any referent (of a visual demonstrative) be in the set of objects associated (by D) with the implicit qualifying concept (of that visual demonstrative). The stipulative response relies on the following formula (in which 'O' expresses the relation of belonging to a causal chain that initiates the occurrence of a visual–demonstrative concept, 'x' is a variable for concept–tokens, and 'y' is a variable for objects).

O^*xy iff Oxy and, if Q is x's qualifier, y belongs to $D(Q)$

Loar says, "O^* determines reference for visual demonstrative concepts. The effect of this is equivalent to having stipulated that O^* —the restricted optical relation— counts as reference for visual demonstrative concepts" (p. 56).

This resolution provides only what we can have by presupposition to begin with: some relation is *taken* (by stipulation) to be the reference relation. The problem remains and proceeds in spite of the attempted resolution. If D is a set of ordered pairs of concepts and sets of objects, then how do we stipulate that D, rather than its close competitor D^* (which pairs the concept *tree* with the set of tree surfaces), is the reference relation? If we could refer to D and predicate of it that it shall be called the 'reference' relation, then we could perform the requisite stipulation. But how we might do such a thing, what it would be to refer to D in particular, is precisely what is in question. We can no more stipulate that D is to count as denotation than we can (successfully) stipulate that 'tree' shall refer to trees rather than to tree surfaces.

---

[1] Parenthetical references here are to Brian Loar's "Reference form the First Person Perspective", this journal, this issue.

Furthermore, even if we could have performed the required stipulation, another problem would remain. D would be stipulated to "count" as denotation —but would it really *be* denotation? Is our stipulation *correct*? Stipulation is, in the first instance, a *terminological* matter —if the underlying phenomenon is ill–understood, no terminological legislation will settle the deeper issue. Some might argue that given the stipulation, the question whether D really is the denotation relation makes no sense. But that is grossly to overestimate the metaphysical power of stipulation. Suppose we believe there are some independent metaphysical facts about reference, for example, that certain relations can constitute that relation in some cases and others cannot. We cannot now avoid the question, "in virtue of what do those relations amount to denotation", by stipulating that we will only use "denotation" for the ordered pairs in D.

Of course, we might change our mind about the existence of independent metaphysical facts about reference. We might judge our earlier questions to be confused or illegitimate and re–interpret them as questions best understood in terms of the stipulatively so–called reference relation. But notice that an analog of *that* move is available to us independently of any stipulation. That is, confronted with the question of why the relation we intuitively suppose to be the reference relation (by contrast to those competitors we suppose not to be reference relations) *is* the reference relation, we could respond that the question is illegitimate. That relation *is* the reference relation simply because we intuitively take it to be —because it is so–called, so to speak. This is to deny a realist presupposition of the question.

Loar does treat the stipulative resolution as subject to serious objection. But his remarks do not reveal the essence of the response's inadequacy. He thinks that consideration of this alleged resolution will highlight a significant distinction between first–person and third–person *perspectives* on reference: "Even if we accepted that third–person reference ascriptions are determinate only by selective stipulation, that would hardly fit first person ascriptions. ... When we consider the references of our own thoughts, it appears bizarre to suggest they are established as such by stipulation" (p. 58). I've already mentioned that because stipulation itself requires reference, the idea of a stipulative resolution to referential indeterminacy may be circular. Still, it is the alleged differential security of reference from the first–person perspective (over the third–person perspective) that leads Loar to consider whether a deflationary account of reference might be especially attractive.

I, for one, find it equally bizarre to suggest that *anyone's* references are established by stipulation (as that anyone else's —even mine—

should be so established). That is, when I compare first–person and third–person reference–ascription, I find them on a par with respect to their dependence on stipulation. Is it just undue caution on my part to be as insecure about the determinacy of the reference of my own thoughts, in the face of general worries about indeterminacy, as I am about the determinacy of any other arbitrary thought? I am, of course, certain that my thought that, say, Gell–Mann was brilliant is determinately about Gell–Mann. It wouldn't *be* the thought that *Gell–Mann* was brilliant unless it were determinately about him. But I am in that sense exactly as confident that *S*'s thought, for any *S*, that Gell–Mann was brilliant is determinately about Gell–Mann —and for exactly the same reason. To put the point briefly and in jargon, if thought ascription —be it first–person or third— is *de re*, then the truth value of the ascription depends precisely on the reference of the thought (and of the singular term used in the ascription).

In another sense, however, one might be concerned about whether a given thought (of one's own) is about Gell–Mann or rather about Feynman. One might even be concerned that a given thought might not be determinately about either Gell–Mann or Feynman. I do not arrogantly suppose that I am somehow a much more *determinate* thinker than others. Why should the reference of my own thoughts be secure in some special or systematic way that is not shared by the thoughts of others? In just the way I worry about how it might be that (and indeed, whether) *any* thought is about a rabbit rather than an undetached rabbit part, so I worry about how it might be that (and indeed, whether) any thought of mine is about a rabbit rather than an undetached rabbit part. To put this second observation briefly and in jargon, if thought ascription —be it first–person or third— is *de dicto*, then the question of which *de re* ascription is true (given the *de dicto* ascription) is hostage precisely to the issue of referential determinacy.

We can and do almost always suspend these philosophical concerns about the determinacy of meaning. This is a contingent fact about our psychology. It may even be true that this suspension of philosophical doubt is more natural or harder to avoid in self–ascription than in the third–person case. But if Loar means to be making a philosophical point in distinguishing sharply between the security of reference from the first–person perspective and its indeterminacy from the third–person perspective, then I do not agree.

How might disquotational ascriptions of reference play a special role in determining which of various external relations qualify as reference? A first try is to suppose that a reference relation is some

relation (R) that applies to disquotational pairs. But this try is unsuccessful. Loar rightly notes that the disquotational pattern holds, when it does, among pairs of *terms*, and not among pairs of terms and objects (that are not themselves terms). There will be no distinguishing a relation as reference in virtue of its holding of 'disquotational term–object pairs': there are no such pairs.

As a second try, we might define 'disquotational term–object pair' as any pair that is picked out by a pair of terms that really do exhibit the disquotational pattern. Again, Loar rightly rejects this effort. The point of using the disquotation theory is to *explain* how one relation rather than another might reasonably be distinguished as the reference relation and to explain how all the disparate relations that constitute reference for disparate concept–types might be unified as instances of the general kind *reference*. If a semantic notion such as *picked out* (is this any different from *referred to*?) plays an essential role in the explanation, by distinguishing certain term–object pairs as 'disquotational term–object pairs,' then the explanation begs the question.

At this point, Loar changes his own perspective on the problems before us. He now urges us "to look at the matter phenomenologically, with an eye to how things appear in the first person" (p. 63). The literal content of this metaphor escapes me. However, if Loar is urging us to keep both the phenomenology of the matter and the matter itself in mind, then I have done so since the enquiry began (why leave the phenomenology out?). This shift of perspective, according to Loar, will provide a way around our problem and yield a unified solution to the two questions with which he is concerned. How might it do so?

Loar asks for our phenomenological cooperation: we are to produce a series of concept pairs where the second member of each pair is a visual–demonstrative concept. The production of the pairs is meant to involve the *exercise* or those pairs. Now we are told: "What is the upshot? From your perspective —that is, the perspective of the user of those concept pairs— pairs of visual demonstrative concepts and certain objects appear on the scene" (p. 63).

We must look carefully at Loar's use of the 'from your perspective' operator here. It is an intensional operator: from ⟨from you perspective, P⟩, nothing follows about the truth value of P. So, in particular, from the fact that having exercised a pair of concepts, a certain concept–object pair has 'appeared on the scene' *from my perspective*, it does not follow that any particular concept–object pair has actually appeared on the scene. Nevertheless, Loar asks how these concepts and objects are apparently related. *Which* concepts

and objects? —we might well ask. Remember, something other than the appearance of a concept and object has occurred. What has occurred is the, so to speak, *from–my–perspective* appearance of a concept and object. And this does not necessarily produce a concept and object pair for me to consider the apparent relation between.

"This is a delicate matter", (p. 62) Loar says. He is tempted to say that the concept and object that have appeared on the scene are linked "*in that characteristic way*". And this, he claims, is sufficient for us to be making *some* sort of sense when we now count concepts and objects as linked in a disquotational way, *from the first person perspective.* Insofar as I can understand this, it seems inadequate: no matter what is true from the first–person perspective, nothing *objective* about any concept–object pairs can be assumed.

It may be that a certain set of concept–object pairs have in fact been referred to (we needn't deny that reference occurs). But, if the issue is what makes one relation rather than another the reference relation, then there is an issue about *which* concepts and objects have appeared on the scene. If one relation constitutes the reference relation, then one concept–object pair has appeared on the scene. If another relation is constitutive, then a different pair has appeared. Of course, it may *seem* that a particular pair has been singled out. But if we seek to establish a link between what has actually come onto the scene (performing this linkage from the first–person perspective), we must be at a loss. Let us interpret Loar then as asking that we establish a phenomenological link between the concept and object that *seem*, *from the first person perspective*, to have appeared on the scene.

Unfortunately, the text does not univocally encourage this interpretation. The next paragraph moves from, "certain concept–object pairs may be counted as suitably disquotational from a phenomenological, subjective, or first–person perspective", to, "we lay on the table, for further investigation, as many such concept–object pairs as we can" (p. 62). The problem is that in an important sense we cannot lay even a single such pair on the table. If reference might be indeterminate, our phenomenological exercise has no determinate objective implications. The only sense in which we might lay concept–object pairs "on the table" is one which will not permit further investigation (for we have no way of *referring* to them —of singling out *which* such pairs we've laid down).

What Loar asserts next begs the question: "Nothing prevents our shifting our perspective back to the objective perspective at this point, keeping *those* pairs in mind while considering *them* afresh

from that third–person perspective. Does some well–carved external, objective, relation hold of *them* in some projectible way?" (p. 62, emphasis mine). But we cannot consider *those* pairs from the third–person perspective. Or, to be precise, any such consideration we do will be fortuitous: we cannot assume that we would be considering any particular concept–object pairs from the third person perspective. To think we can is to assume that there is a way to refer to one concept–object pair rather than another.

Loar later seems to address this worry. He claims that "[t]he construction 'it appears to me that... that tree...' is a *de re* propositional attitude ascription, and the question of inscrutability arises as much for such constructions as for ascriptions of reference" (p. 64). And he denies that his remarks depend on (as it were) the objective ascription of appearances. Granted, an objective (*de re*, extensional) construal of his phenomenological remarks would be unfair. On the other hand, and this is the point I have been stressing, without such a construal, there is no way to shift perspectives and consider any particular concept–object pairs; no objective facts have been established. It is admitted that "[w]hen we shift from the phenomenological perspective, the pairs that we have placed on the table are not special, from the objective perspective" (p. 64). But it has not been established even that we have determinately placed any particular pairs on the table. If the investigation is meant to concern one set of concept–object pairs rather than another, then whether we can perform the objective investigation is what is at issue.

Consider the following. "Perhaps reference is indeterminate. But now exercise the concept *that rabbit*. You thereby appear, from your first person perspective, to have put a rabbit on the table for further investigation. Nothing now prevents you from switching perspectives and considering that rabbit from the objective perspective". Are not Loar's remarks ultimately analogous to these?

Loar believes that by exercising certain concept pairs, we make certain concept–object pairs salient. In a sense this is right. We make it *appear* that certain concept–object pairs are salient. But although salience is mostly a matter of appearances, it does not follow that what we appear to make salient, we make salient. This, perhaps, is a critical misstep. We simply cannot assume that in the course of Loar's thought experiment we actually apprehend the salience of a particular object in the presence of a particular concept. We appear to achieve that comprehension. But the question remains: Do we achieve it?

Elsewhere, Loar seems to appreciate this idea. He says that "[d]escribing the special concept–object link as 'apparent' is apt, for it

is an illusion: no objective relation —'x is salient in the presence of concept C'— between object and concept is created by the mere exercise of the two concepts" (p. 63). But ultimately, the same illicit assumption shows through. "[T]hat the name–object pairs we lay on the table are subjectively privileged... permits proceeding to the further stage of investigation, when we shift perspectives, and see that those concepts and objects are objectively related after all, independently of the phenomenology" (p. 63). No; in shifting perspectives we gain no direct, objective access to any particular concept–object pairs. Any claims are claims about concept–object pairs *we know not which*. In an important sense we cannot perform an investigation from the third person perspective. Such investigaton requires *reference* to the objects of investigation. We would need to be able to say something like, "there is concept c_1 and object o_1. How are *they* related? By R?...". But we cannot, from the third person perspective *name* the concept–object pairs that we have "laid out on the table" (in Loar's perhaps misleading metaphor), for we cannot determinately refer to *them*.

It may be right that "[i]n phenomenological description anything goes, and literality is mostly a barrier" (p. 64). But understanding Loar's metaphors, even non–literally, presents a considerable barrier of its own. He claims that "[o]ur objective perspective can be seen as having a significant semantic upshot only from the subjective perspective" (p. 64). Does this mean that our objective investigation has no *objective* result? But if we know this in advance, why perform it at all? Indeed, isn't an investigation that has 'an upshot' only from a subjective perspective really only a subjective investigation. Why is the investigation we perform when we 'shift perspectives' (have we really done this after all?) an *objective* investigation? "No special objective relation, of any sort, is being asserted to single out those pairs", we are told (p. 64).

All this is no advance on our original question. We know from the outset that the phenomenology of determinate reference is robust. Intuitively, when I use the concept *rabbit*, I refer to rabbits and not to undetached rabbit–parts. Of course, as I argued earlier, when anyone else uses the concept *rabbit*, she refers to rabbits too. I do not see the first–person/third person asymmetry. And I do not see how the phenomenological points put before us help to advance any significant philosophical project.

The question at the beginning was whether the indeterminacy or inscrutability of reference has a subversive effect on our commonsense conception of semantic facts. It is no response to this question to point out that, according to our commonsense conception, seman-

tic facts are determinate. Can we maintain this conception in the face of philosophical doubts about determinacy? On one reading, this is a question for psychology. We seem to be able to maintain all kinds of conceptions in the teeth of overwhelming data that the conception is flawed. So perhaps this is another case in point. But if the question is whether it can be a part of the commonsense conception of reference that reference is indeterminate, then I do not see how anything Loar has said would make us reconsider the very tempting negative answer.

There is a further issue. So far I have been worrying about the foundation of Loar's project. What he says about the phenomenology of referential determinacy does not seem to advance our understanding of referential determinacy itself. But even if it did, it is unclear what role the disquotational schema is playing. If it were possible simply by exercising a pair of concepts, to (pulling a *rabbit* out of the hat) lay a particular concept–object pair on the table such that we could then investigate that concept–object pair from the third–person perspective, then I do not see what is so special about the concept pairs that we are asked to exercise. According to Loar, those concept pairs exhibit a "disquotational pattern". He calls them "*mention–use*" pairs of concepts. But I'm not clear about what is involved in being a mention–use pair.

One obvious explanation is not available in the present context: a concept mentions another iff it refers to it. In Loar's examples, the visual demonstrative concept is mentioned by the higher–order concept because it is the referent of that higher–order concept. But these pairs are meant to figure in an explanation of how it is that our conception of determinate reference can be maintained in the face of familiar philosophical worries. To appeal to that same notion would therefore involve us in an explanatory circle. Furthermore, only pairs of *terms*, it seems to me, can exhibit a disquotational structure (since only terms are apt for quotation in the first place). To be fair, Loar puts "disquotational" in quotations (!) in the text; he is clear that his is an extension of the basic notion. And he means "mention–use" pair to be suggestive. But where he is not clear is in exactly how the basic notion of a disquotational pair is to be extended to that of a mention–use pair. There's more than one way to mention a concept with another.

For example, why couldn't the following pairs exhibit the mention–use structure?

> the next occurrence of a canine visual demonstrative concept,
> that dog

the next occurrence of a botanic visual demonstrative concept, that tree

the next occurrence of an anatomic visual demonstrative concept, that hand

If these pairs do not exhibit the mention–use structure, then I wonder why not. Is there, after all, a way to put quotation marks around a concept? On the other hand, if these pairs *do* exhibit a mention–use structure, then what work, exactly, is the disquotational schema doing in the overall project?

Loar thinks that he has found a way to privilege the reference relation over its nearby competitors. We have a higher–order recognitional concept which *purports* to pick out a certain external relation (reference) from the mention–use perspective. The relation demonstrated by the higher order concept will be one that holds of the concept–object pair ⟨*that tree*, that tree⟩ and not of ⟨*that tree*, that tree–surface⟩. But why will it? Our recognitional concept of the reference relation is said to be sensitive to pairs that present themselves as linked *in that characteristic way*. This may be so. But the main problem is whether corresponding to the appearance that certain concept–object pairs *present* themselves as linked in that characteristic way, there really are pairs that really *are* linked in the 'referential' way.

There was never a doubt that *apparently*, some concept–object pairs exhibit the reference relation. It may be that some concept–object pair (or other) is presenting itself as the ⟨*that tree*, that tree⟩ pair. If it were so, that concept–object pair (whichever it is) would be presenting itself in a distinctive way —we might even call it a 'mention–use' way. The problem is that the concept–object pair in question might fail, in fact, to be the ⟨*that tree*, that tree⟩ concept–object pair. It might be some other pair instead, masquerading, to our limited referential capacity, as the ⟨*that tree*, that tree⟩ pair. The distinctive appearance does not guarantee even that some concept–object pair is appearing. And if in certain cases some concept–object pair is appearing in a distinctive way, that pair may still fail to be referred to by the concepts that constitute the appearance.

Toward the end Loar makes a number of comments that make explicit the conflation I have been alleging is implicit in most of the paper. He is comparing objective and subjective determinacy of reference. He admits that it appears to him "quite likely that there is no resolution of the indeterminacy problem in objective terms, and hence that the Quine view is in a certain straightforward way correct" (p. 71). But he denies that this result has the eliminative,

subversive consequences it is usually seen to have. Loar has never specified exactly which consequences he has in mind here. If it is a condition on reference that it be determinate, and if the Quine view is that nothing that could be reference is determinate, then one subversive consequence of the Quine view is that reference does not exist. If Loar is trying to resist this consequence by pointing out that we have a robust illusion of referential determinacy in spite of appreciating the indeterminacy problem, then what he has shown is at most that 'reference–from–a–perspective' does not entail referential determinacy (perhaps it does entail 'referential–determinacy–from–a–perspective'). And our conception of reference is not our conception of 'reference–from–a–perspective' (not even from our own perspective).

Loar claims that the referential relation is discriminated from its competitors from the subjective perspective. But we must, he notes, understand "discriminated" itself subjectively (p. 71). It is *as if* there is a rule that generates an open–ended series of term–object pairs. Does this '*as if*' harbor an illusion?

"The issue is delicate", (p. 71) Loar says. He concludes that the Quine position is, indeed, eliminative of a *philosophical* supposition that reference is objectively determinate. But he believes that he has given an account of why a certain relation is the reference relation.

I believe we have been given at most an account of why any apparent experience of the reference relation should count as such an experience. From a perspective it appears that some relation is the reference relation. The experience counts as apparently of the reference relation because it is *as* of a pair that exhibits a certain mention–use pattern. Actually being an instance of the reference relation, however, is *actually* exhibiting the pattern.

The indeterminacy problem is the problem that there may be no fact of the matter determining that some pairs exhibit the pattern and others do not —it may never be determinately true that any pair actually exhibits the pattern. Loar's phenomenological investigation cannot be assumed to be an investigation of the phenomenology of reference; it might be an investigation of the phenomenology of the relation that holds, *inter alia*, betweeen the concept *rabbit* and rabbit surfaces. It is consistent with the problem of indeterminacy that *apparently*, some pairs fit a pattern and others do not. What cannot be assumed is that this appearance in any way limits what might be objectively true about reference itself. And the possibility of referential indeterminacy deeply challenges our basic intuitions about the nature of semantic facts.

Philosophical Issues, 6
Content, 1995

Fregean Reference Defended

Ernest Sosa

What is involved in acquiring a russellian proposition $\langle x, \phi \rangle$ as content of an attitude: what does it take for one to acquire such an attitude *de re*? How do we gain access to x itself so as to be able to have $\langle x, \phi \rangle$ as content of our thought?

1

"What makes my idea of him an idea of *him*?" So queried Wittgenstein, and his query is part of a family. You can refer to someone through an idea but also through a thought. So: "What makes my *thought* about him a thought about *him*?" Such questions may be given a broadly "fregean" answer as follows (where α spans both absolute individuators d that pick out their referents independently of context and perspectival individuators i that pick out their referents only with the aid of a context of use):

(FT) A subject S has at time t a thought (belief, intention, etc.) *about* x (*of* x) if S thinks (believes, intends, etc.) *de dicto* a proposition that predicates some property ϕ with respect to

some individuating concept (or individuator) α of x for S at that time.[1]

Here then is an answer to Wittgenstein: My thought about him may be about *him* in virtue of having for its content a proposition that predicates something with respect to some individuator which, in the context, is satisfied (uniquely) by *him*.

2

The *prima facie* plausibility of our fregean answer FT is highlighted by a simple argument:

P1. If there is such a thing as the F then the proposition that the F is G is about the F and attributes being G to the F.

P2. If one believes proposition P, and P is about x and attributes being G to x, then one's belief is about x and attributes being G to x.

C. Therefore, FT.

This simple argument shows that FT is right for *some* natural and familiar senses of 'about' and 'of' (and with respect to associated constructions such as 'attributing a property to a thing'). Yet many have rejected FT and its associated aboutness.

According to a competing doctrine of *de re* attitudes you do not attain genuine reference merely by having in your thought some individuating concept that picks out an object. To think about, to refer in thought to, the tallest spy, it is not enough simply to have the thought that the tallest spy is a spy, even supposing there is such a spy. This competing doctrine may be sketched as follows:

N A genuine relation of reference must be constituted by some special relation binding the thinker with the object of reference, probably some causal psychological relation like perception or memory.

---

[1] FT permits individuating concepts to be "perspectival" or "indexical", since they need only be satisfied not absolutely but in a "perspective", one that supplies the indices or parameters required for such a concept to be satisfied. Note also that α is an individuator *of* x only if x exists.

A compromise position —adopted, for example, by Stephen Boër and William Lycan, accepts FT for one minimal sense of aboutness,[2] but requires N for other senses. Higher grades of aboutness require referents that "...are determined *solely* by the (appropriately shaped) causal chains that ground them in their referents".[3] In any case, to the credit of Boër and Lycan, they do emphatically recognize a basic, fregean, "latitudinarian" sense of aboutness involved in FT. Many others have been less perceptive.

3

N has widely been thought to be favored, as against FT, by the likes of Donnellan's famous case. A partygoer, viewing a man in a corner holding a glass of clear liquid with an olive, says "The martini drinker in the corner is getting tipsy". Behind a column in that corner a little girl in fact downs a dry martini, while the man drinks only water. According to fregeanism, what *S* explicitly thinks and says is then strictly about the little girl, since, after all, the proposition thought and affirmed is about her and not about the man. It is the girl, not the man, who is drinking a martini in the corner. Nevertheless, isn't it obvious that *S* really has the man in mind, is really thinking or talking about the man, and really referring to him?

Suppose, again, *S* holds up a picture of someone who looks just like Stalin but is actually, unknown to *S*, an actor made up to look like Stalin. If *S* thinks, or says, "The man in the picture was evil incarnate", whom is he referring to or thinking about: the unknown actor or Stalin himself? According to fregeanism, what *S* explicitly thinks or says is strictly about the actor, since, after all, the proposition thought or affirmed is about the actor and not about Stalin. It is the actor, not Stalin, who is pictured (even if Stalin is in some sense "represented" by the actor with that make–up). Nevertheless, is there not some sense in which *S* really has Stalin in mind, is really

---

[2] "Latitudinarian" aboutness, as they call it, using a label taken from Chisholm ("Knowledge and Belief: 'De Dicto' and 'De Re'", *Philosophical Studies*, 29 (1976): 1-20) and applied to a view defended in my "Propositional Attitudes De Dicto and De Re", *The Journal of Philosophy*, 67 (1970): 883-896. That paper in effect adopts FT and defends fregeanism by explaining away (through appeal to pragmatic implicatures) intuitions often taken to support N. The present paper supplements that effort through other means. (See S. Boër and W. Lycan, *Knowing Who* (Cambridge, MA: MIT Press, 1986).)

[3] *Knowing Who*, p. 128.

thinking or talking about Stalin, and really referring to him? What sense can this be?

Suppose further that S next holds up the picture and thinks or says "The man in the picture, as you can see, had a full head of hair". Of course there is still a straightforward sense in which S is thinking or talking about the actor pictured, the sense captured by the fregean account. But isn't it about as plausible that there is some sense in which, again, S really has Stalin in mind, is really thinking or talking about Stalin, etc.? Once again, what sense can this be?

Compare, finally, a political rally where someone introduces the Presidential candidate by saying "Ladies and gentlemen, with us today is the next President of the United States", where against all odds it is the opposition candidate who is about to be elected. Whom does the speaker have in mind, whom is he talking about, whom is he referring to? Is it his own party's candidate, next to him on that platform, or is it the opposition candidate? The answer seems to me clear, yet it is equally clear that our fregean account is unable to accommodate it. For the individuating concept [the next President] does *not* for our speaker pick out the woman next to him on that platform.

We are in search of a better account of cases like that of the Presidential candidate, that of the Stalinesque actor, and that of the little girl with the martini. Here now are some technical concepts that will aid the formulation of a proposed solution.

D_1. Ω is a *referential conception* for S at t IFF Ω is a maximal set of individuating concepts (absolute or perspectival) such that at t every pair of such concepts are connected by a chain of individuators, adjoining links of which are always believed by S to be codesignative.[4]

D_2. Individuator β *epistemically derives from* individuator α for S at t iff S at t believes (justifiably) [there is such a thing as β] on the basis of believing (justifiably) [there is such a thing as α] and [whatever is a is β]; but not conversely.

D_3. $\epsilon(\Omega)$ is for S at t the *epistemic basic source* —or, alternatively, *epistemic basis*— of referential conception Ω for S at t IFF

---

[4]More formally: "...such that at t every pair of such concepts are "connected" in the following sense: for every such pair —α, β— there is a sequence of individuating concepts —$\alpha_1, \ldots, \alpha_n$— such that S believes (at least implicitly) the propositions $[\alpha = \alpha_1]$ and $[\alpha_n = \beta]$, and for every i $(1 \leq i < n)$, S believes also the proposition $[\alpha_i = \alpha_{i+1}]$".

$\epsilon(\Omega)$ is a minimal subset of Ω such that (a) each individuator a in $\epsilon(\Omega)$ epistemically derives, for S at t, at most from other members of $\epsilon(\Omega)$, and (b) every individuator β in Ω that is *not* a member of $\epsilon(\Omega)$ epistemically derives, for S at t, from members of $\epsilon(\Omega)$. (Members of such a minimal subset then qualify as "epistemically basic" individuators.)

D_4. Referential conception Ω is for S at t *collectively about* x IFF the great preponderance of the members of its epistemic basis $\epsilon(\Omega)$ are individually about x, relative to S at t.

D_5. Individuator α is for S at t *associatively about* x IFF there is an Ω such that (a) Ω is a referential conception for S at t, and (α is a member of Ω); and (b) Ω is for S at t collectively about x.

Here are some facts about associative aboutness, all relative to an arbitrary subject S and time t. First of all, an individuator could possibly be individually about x and associatively about y even if x is not numerically identical to y. Second, just as no individuator could be individually about more than one thing (and indeed for that very reason), so no individuator could be associatively about more than one thing. Third, in a great variety of ordinary cases one refers associatively to a concrete contingent entity other than oneself, in virtue of being causally related to it. In such cases the basic source of one's referential conception Ω that is collectively about x contains individuators like [the man I see drinking a clear liquid] or [the person I saw drinking just now] or [the one I seem to see drinking]. Since there is a hidden causal commitment in the claim that one perceives or seems to perceive someone or something, any such referential conception Ω will have a basic source $\epsilon(\Omega)$ substantially composed of such causally committed individuators. This is how I would propose to capture the intuitions that underlie causal accounts of reference (in thought).

The concepts of a referential conception and of associative aboutness enable us to deal uniformly with the various problematic cases before us. In each case there is a thought T (of the form [α is ϕ]) such that T is about x in a straightforward, fregean sense, since it is x that satisfies the individuating concept α, while yet in some natural sense T is not really "about" x: it is not x that S really "has in mind" and is "referring to". What is that further sense? I suggest the concept of associative aboutness just defined. In terms of that definition, it may be seen: (a) that the thought "The martini–drinker in the corner is getting tipsy", in the Donnellan example, is associatively about the man drinking the clear liquid (and not about

the true martini–drinker in the corner, the little girl out of sight); (b) that the thoughts "The man in the picture was evil" and "The man in the picture, as you can see, had a full head of hair" are associatively about Stalin (and not about the innocent actor made up to look like Stalin and photographed); and (c) that the thought "With us today is the next President of the US" is associatively about the candidate on the platform (and not about the opposition candidate about to win against all odds).

Given that we have adopted here only the modest project of defending the fregean sufficiency claim FT, it might be wondered why we do not stop with the simple argument of section II above. Answer: Because that would not be persuasive to anyone impressed by examples like those we have discussed. So our strategy is rather this:

a. To use that simple argument in support of the idea that FT is right for *some* natural and straightforward senses of 'about' and 'of', etc.

b. To grant that there is at least one other set of senses of 'about', 'of', etc., which go beyond that superficial, natural sense; to grant the motivation for this provided by examples like those discussed.

c. To argue that these further senses can be captured in terms of the superficial, natural sense, by imposing some further restrictions, none of which appeals to any basic thought/world relation other than that of unique instantiation or unique satisfaction of an individuator by an object (relative to a context).

Even with the notion of associative aboutness no appeal is made to any causal relation, between term and object or concept and object, in order to explicate these richer conceptions of aboutness or reference. Causation does play an important role in associative aboutness, again, but it does so only in a derivative, "by the way" fashion. The role played by causation derives from the way causation is bound to enter in the constitution of the epistemically basic individuators that make up the basic source of a referential conception. The individuators in the basic source of a referential conception determine what that conception is collectively about, and this in turn determines what all the individuators in the conception are associatively about. No wonder causation plays an important role in determining what a speaker's or a thinker's individuators are associatively about. But this is a mere by–product of what is involved in the very

concepts of aboutness as defined here, a mere by–product that derives from (a) the definition of associative aboutness, together with (b) the important role that causation must play in many of the individuators that would function as epistemically basic individuators for a speaker or thinker at a time.

4

Our fregean approach is meant to cover propositional attitudes in general and intention in particular. Thus:

(FI) S at time t intends to Rx if there is an individuator α of x for S at t such that at t S intends *de dicto* a proposition that predicates the relation of R–ing with respect to himself and α in that order.

If S intends *de dicto* a proposition that predicates R–ing with respect to himself (as himself) and [the F], in that order, then not only the property of R–ing, but also the individuating concepts [Myself] and [the F], both enter into his conception of how, according to his intention, things are meant to turn out. Thus if I intend *de dicto* (the proposition) that I feed the cat before me, then not only the property of feeding but also the individuating concepts [Myself] and [the cat before me] both enter into my conception of how things are to turn out.

According to a further objection, now, such fregeanism is refuted by the "shell game" problem:

> Suppose there was earlier a pepper mill (x) to the left of S, and that S saw it and (mis)took it to be a full salt–shaker. Later the pepper mill was removed and replaced with a full salt–shaker (y), without S's realizing it and without S's ever perceiving or knowing of y. S now reaches to the left —still without looking, without seeing the object there, relying only on his memory— in order to pick up the object he believes to be there (the pepper mill, presumably, the one that he erroneously believed and believes to be a full salt–shaker). S does pick up the object that is now there: as it happens, fortuitously, a full salt–shaker, but not the object S was reaching for, the object S believed and believes to be there.

It has been objected that in this example the conditions laid down by fregeanism (FI above) for S to intend to Ry (the salt–shaker now

to S's left) are all satisfied.[5] For, we are told, [the salt–shaker to the left] is an individuating concept a such that:

(i) α is an individuator of y for S at t; and

(ii) S, in C, intends *de dicto* a proposition that predicates reaching with respect to himself and α in that order.

Yet, according to the objection, the agent S really intends to reach the pepper mill rather than the salt–shaker, since it is the pepper mill that he saw and remembers and believes to be still there (even though he misperceived it and erroneously believed and believes it to be a full salt–shaker).

This example is aimed against any fregean conception of reference in thought, with the goal of promoting a narrower conception such as N above. It is supposed that through a combination of memory and perception the agent S is genuinely related in thought to the pepper mill and not to the "impostor" salt–shaker, which happens to fulfill by accident the abstract conceptual content of S's intention.

However, we now have available to us a sense in which the subject is "really" reaching for the pepper mill, one acceptable to fregean intuitions. For, relative to that subject and time, and to the situation of the example, the subject's thought is associatively about the pepper mill and not the salt shaker. An observer who wishes to explain the subject's reaching as he does might then offer the following explanation, as she points to the pepper mill: "The agent is now reaching as he does because he intends to reach thus if by so reaching he can reach *this* (the pepper mill) and he thinks that by so reaching he *can* reach this". In offering that explanation, moreover, the observer would implicitly assume about the belief *and* the intention combined for that explanation, that in them both the agent thinks in the same way, via the same individuator a, that he reaches *this* (i.e., the pepper mill).

5

I have defended a form of fregeanism according to which de re attitudes about entities always derive from de dicto attitudes with fregean or quasi–fregean propositions as contents. In simple cases the de re attitude derives in the straightforward way of FT: here the

---

[5] Cf. T. McKay's "Actions and *De Re* Beliefs", *Canadian Journal of Philosophy*, 14 (1984): 631-635.

content of the de dicto attitude requires only an individuator (fregean and absolute or quasi–fregean and perspectival) which denotes the object of the de re attitude relative to the context of the thinker at the time (or absolutely, and thus, trivially, also relative to that context). More complex cases are also possible, however, as is well brought out by the extensive literature on the so–called "problem of exportation". However, I have suggested that such more complex cases can be handled without departing from our fregeanism. Many of them can be handled, for example, through the concept of "associative aboutness" (definition D_5). The strategy is to introduce a relation of aboutness that still works via fregean or quasi–fregean individuators but also makes use of a richer set of devices such as those of a referential conception and its basic source, and, eventually, that of associative aboutness itself. The main point here is this: no such device appeals to any basic thought/world connection other than unique instantiation or unique satisfaction of an individuator by an object (relative to a context) —no special or essential causal mechanism of reference is required in general.

PHILOSOPHICAL ISSUES, 6
Content, 1995

On Sosa's "Fregean Reference Defended"

William G. Lycan

Professor Sosa has done two particularly interesting things: First, he has applied Donnellan's famous "near miss" type of example–and–argument[1] to thoughts as opposed to public utterances. Second, he has offered us, very neatly and without fanfare, a new psychosemantics, one designed to accommodate Donnellanian intuitions. (A "psychosemantics" —the term is Jerry Fodor's— is a theory of mental referring, i.e., as Sosa says, an answer to the Wittgensteinian question, "What makes my thought of him a thought of him?", or as I would put it more tendentiously: in virtue of what is an individual person or thing the particular referent or representatum of a particular mental representation occurring in someone's head?)

The first achievement is important because the most obvious *linguistic* treatment of Donnellan's puzzle cases is not obviously available for the case of pure thought. That now fairly standard treatment is Kripke's:[2] to distinguish Gricean utterer's or speaker–reference from public–semantic reference and maintain that while the latter goes by literal descriptive content as determined by the public con-

---

[1] "Reference and Definite Descriptions", *Philosophical Review*, Vol. 75 (1966), pp. 281-304.

[2] Saul Kripke, "Speaker's Reference and Semantic Reference", in P. French, T.E. Uehling and H. Wettstein (eds.), *Midwest Studies in Philosophy*, Vol. II: Studies in the Philosophy of Language (Minneapolis: University of Minnesota Press, 1977).

ventions of the public language, the former can differ because it goes by whom or what the speaker has more directly in mind. But in the case of pure thought, it is harder to motivate a distinction between the person or thing determined by literal descriptive content in the thinker's language–of–thought and the person or thing the speaker "has in mind".

For myself, I suspect we will ultimately need such a distinction anyway, but this is hardly the occasion; I turn to Sosa's ingenious psychosemantics. In essence, if you recall, it is the claim that a thought/representation designates (or is "really" about, "really... pick[s] out") an individual O just in case it contains an individuator α that is "associatively about" O, which (being interpreted) means that α is a member of at least one of the subject's Referential Conceptions Ω, such that the individual O satisfies "the great preponderance of" the descriptive concepts that together comprise Ω's Basic Source (whether or not O satisfies any of the other, probably more numerous descriptive concepts in Ω).

What is distinctive, and I think laudable, about this analysis is that it preserves traditional descriptivist insights (those of the *other* Russell, purveyor of the description theory of ordinary linguistic referring, better known to most philosophers than the "singular proposition" Russell to whom Sosa alludes), while accommodating some key objections to descriptivism and in particular explaining away the urgent plausibility of Kripke's–Donnellan's–Putnam's dominating "causal–historical" intuitions about reference generally and causal psychosemantics in particular. (Why is this revanchist action laudable? Because the opposing causal–historical views, especially those directed upon the case of pure thought, have run into very nasty problems, of circularity and the like, and the relation of *a thing's having the individuating property codified in a description* is far less vexed than is that of *a thing's being causally–historically connected "in the right way" to someone's present thought/representation.*)

Still, I would like to find a counterexample to Sosa's psychosemantics —since we have learned from his colleague Roderick Chisholm that if an analysis or explication cannot be counterexampled, it is not yet precise enough to be taken seriously or even well understood.

An initial range of *putative* counterexamples can be carried over from Donnellan's and Kripke's attacks on Description theories of linguistic referring.[3] That is because Sosa's theory is in one way quite demanding: It requires that, in order for a subject S to have any

---

[3]Keith Donnellan, "Proper Names and Identifying Descriptions", and Saul

thought that is about O, even associatively, S must deploy at least one individuator that O actually and nontrivially fits. Donnellan and Kripke think they have refuted any such requirement, by describing cases whose protagonists lack uniquely identifying information but in which reference is secured anyway: (1) Donnellan's example[4] of the child who says (or thinks) "Tom is a nice man", though the child encountered Tom only once, while half–asleep in the middle of the night, and has virtually no memory of the encounter. (2) Kripke's example[5] of "Cicero": All some people know about Cicero is that he was a Roman orator who is now (by English speakers) called "Cicero", which may not suffice to individuate. Or finally: (3) My middle–aged memory is going fast (and as has been shown by a fair bit of recent psychological research, event–memory *at its best* is pathetically unreliable). Sadly, I now have only the vaguest memories of certain individuals who meant a great deal to me twenty–five or thirty years ago; I have no identifying descriptions of them, but only causal chains of memories extending back into time.

Sosa may simply not share Donnellan's and Kripke's intuitive judgments about such cases; not everyone does. But here is another class of examples that may faze him.

In these examples, as in the Donnellanian "near miss" cases that Sosa has accommodated in his paper, the identifying description that we saliently voice to ourselves is not in fact satisfied by the "real" referent of our token. But it will not be so obvious, in these, that the incorrect salient description rests epistemically on a Basic–Source or core body of information that is preponderantly correct. I have in mind cases in which our information is all false *and* we got it by hearsay over a large spatial or temporal distance. (4) Kripke's example[6] of the Biblical Moses will do, if we assume *arguendo* that nothing predicated of Moses in scripture is historically true. Or (5) David Kaplan's example[7] of Robin Hood, as allegedly investigated by historians of England a decade or two ago, is another such: There was a real person in whom the modern Robin Hood stories are grounded, but he had virtually none of the properties ascribed to him in those stories. Or consider (6) a body of lore that we have acquired through television or through our less reputable newspapers,

---

Kripke, "Naming and Necessity", both in D. Davidson and G. Harman (eds.), *Semantics of Natural Language* (Dordrecht: D. Reidel, 1972).

[4] *Op. cit.*, p. 364.

[5] *Op. cit.*, pp. 291-3.

[6] *Op. cit.*, pp. 277, 282.

[7] Offered in a lecture to the National Endowment for the Humanities Summer Institute in the Philosophy of Language, University of California at Irvine, 1971.

about some celebrity or other public figure. The body of lore is a pack of lies, completely false, but it is told of a real person to whom our only access is the television and newspaper reports themselves.

On Sosa's model, our various unsatisfied individuators epistemically derive from the other individuators comprising a Basic Source, and those others are mostly satisfied. But what are those more accurate and reliable individuators, in cases of myth, false legend and media fantasy? Since our epistemic window on the slandered hero of such a tale is the story itself and nothing but, the only individuators we have that are actually satisfied by that person are concepts like "the person described in this myth", "the person now referred to as 'Robin Hood'", "the person these stories are about", etc. And these concepts all have semantical concepts as constituents. Indeed, more specifically, they all presuppose a relation of *referring*.

There is no immediate problem, since the referring in question is public–linguistic rather than mental. But I see pitfalls down the road. For our theory of linguistic reference must itself sustain cases of referring despite massive misinformation, and ways of doing that are limited.

First, there is no perfect analogue of Sosa's method, for linguistic reference, since his model requires evidential basing–relations obtaining idiosyncratically within individual subjects. Waiving that, we could still go for a small hard core of secure descriptions, of the metalinguistic sort I have mentioned; but a kind of grounding problem would arise, if not a vicious circularity: Our test for tracing the referent of a public singular term would involve finding "whoever is the referent of" that term occurring in such–and–such a story, i.e., tracing the referent of that term. (The latter might be the next earlier step in a recursion, rather than a flat circularity, but what, then, would be the recursion's base?)

There are two other possibilities. We might bring mental reference back into our descriptive contents, in some way. But offhand I do not see how that would help, and it also would threaten the whole project with circularity.

Or we might just give up on the descriptivist theory of linguistic referring, in favor of a causal–historical account or whatever. But then it seems simplest to do the same for mental aboutness, carrying the new rival account inward. (If, in particular, the causal–historical theory works for language, it surely works for mind, though the converse may well be false.)

I am sure there are tenable ways around this general problem, but I have run out of space, and it is Sosa's job to find them, anyway.

PHILOSOPHICAL ISSUES, 6
Content, 1995

Doubts about Fregean Reference

Manuel García–Carpintero

Sosa's paper attempts to provide the beginning of a solution to some of the problems raised by so–called *de re* attitudes which have devilled philosophers for the past thirty years. The problems Sosa confronts might be divided into two separate classes: (i) problems regarding the semantics of *de re* attitude reports, and (ii) problems regarding the nature of the reported attitudes themselves. My commentary is in two parts. In the first I stress the fact —acknowledged by Sosa himself— that the Fregean motivation for the picture of *de re* thoughts he gives cannot be semantical, i.e., to offer a more simple account of the truth–conditions of *de re* ascriptions. I shall argue that Sosa's Fregean views are in that respect unmotivated, for, as he himself acknowledges, we need to take into consideration in some way or other the points made by philosophers opposing the Fregean paradigm, to give a proper account of the semantic facts. In the following section I shall deal with the deeper questions belonging to the philosophy of mind proper. I indicate where the true motivation for Sosa's views should be looked for, and I give some reasons for doubting that the program to which the account is intended to contribute might succeed. What I shall argue in this part of my

comment is, firstly, that we have been left without information that is crucial to a proper appraisal of Sosa's program; and, second and more ambitiously, that if properly completed the account could not be developed in the ways Sosa apparently envisages.

1

(I shall follow Sosa in taking into consideration just attitudes concerning basic propositions, i.e., propositions whose truth–conditions involve just simple state of affairs: the application of a n–adic property to a sequence of n objects.) Sosa's basic proposal to deal with *de re* ascriptions is contained in his principle *FT*. This is the simplest possible Fregean idea already suggested by Quine in his classic "Quantifiers and Propositional Attitudes" (Quine 1966): all it takes for a subject to have a relational thought about x at a time is for him to have a *de dicto* thought involving a descriptive individuating concept δ such that, as a matter of fact, δ uniquely individuates x. The innovation is that the descriptive concept is allowed to be "perspectival" —to handle the much discussed data on indexicality in attitude–ascription. (The data will perhaps demand that all *de dicto* thoughts allowing correct *de re* ascriptions be taken as in fact *de se*, as explained in Lewis 1983a.) This simplest possible Fregean account is defended on the basis of an argument which I shall not explicitly discuss here. Let me just say that the very same objections I shall presently raise to its conclusion would immediately affect its first premise. 'Proposition' and 'aboutness' are too much theoretically–charged terms for pure intuitions to support claims involving them without further theoretical considerations.

Let us distinguish *de re* thoughts and attitudes in general from *de re* ascriptions of attitudes. Sosa's usage implies that a *de re* attitude is for him one whose content is a Russellian proposition (cf. the beginning of section II in his paper, for instance). On the other hand, nothing in the paper allows an easy characterization of *de re* ascriptions. A *de re* ascription is, of course, one made to attribute a *de re* attitude. There are well–known attempts to distinguish *de re* from *de dicto* ascriptions according to two semantic criteria: the truth–conditions of *de re* ascriptions should not be modified when at least some singular terms under the scope of the attitudinal verb are substituted by any coreferential singular term (they allow substitutivity of coreferential terms, for some terms inside the clause giving the content of the attitude); and *de re* ascriptions should allow unrestricted existential generalization, also relative to some terms inside

the clause giving the content of the attitude. Notwithstanding the classic Quinean argument (in broad outline: if a term is in an utterance to refer to something —and *some* term under the scope of the attitudinal verb in a *de re* ascription *must* be there to refer to something— we should be able to replace it by any coreferential term without affecting the utterance's truth conditions, still less its truth–value), the first criterion raises important doubts: after all, a term could well be doing, as it were, double duty—like 'Giorgione' in Quine's famous example, "Giorgione was so–called because of his size." Not wishing to get tangled here in thorny issues, let us stick to the simple and not very useful characterization: a *de re* ascription is one made to attribute a *de re* attitude. Now, a consideration which might be possibly advanced in favor of Sosa's theory of *de re* attitudes is that it allows an extremely simple characterization of the truth–conditions of *de re* ascriptions by existentially quantifying on *de dicto* attitudes: for a *de re* ascription ascribes a *de re* attitude, and an attitude *de x* exists (according to Sosa's FT) when the subject has some or other *de dicto* attitude involving an individuating concept of x; therefore, a *de re* ascription is true when the ascribed subject has some or other *de dicto* attitude about the *res* in question.

Now, as it has been frequently noted (see for instance Richard 1990, p. 68), such an assignment of truth–conditions to attitude–ascriptions is equivalent to the "Russellian" one advanced by the most radical defenders of what is called *Direct Reference* theory; and it has the same fundamental problem. The problem, broadly put, lies in the opacity that very robust intuitions force on intuitively *de re* attitude reports —it is mildly ironic that this problem was the original motivation for the primitive Fregean theory of attitude–ascriptions. If the simplest semantic account of *de re* attitude reports which flows from Sosa's FT–principle were correct, then the substitution of coreferential terms should not affect truth–value; however, we have firm intuitions indicating that it does. There are contexts, easy to set up, such that an utterance of 'John believes that Cicero was a senator' —while still intending to attribute a *de re* attitude— provides information not only on what *res* the attitude is about, but also on the character of the subject's conception of it. The intuitions are to be explained by the fact that when we ascribe *de re* thoughts, the singular terms we use to indicate what *res* those thoughts are about do not contribute just that *res* to the individuation of what we mean; i.e., they do not provide just the partial information that the ascribed thought includes some or other "individuating concept" which as a matter of fact is about that *res*. The singular terms we use in many attitude–ascriptions contribute to what we mean some addi-

tional more specific information on the nature of the "individuating concepts."[1]

A correct account of the semantics of *de re* ascriptions, thus, cannot be of any simple sort. It cannot be just that in *de re* ascriptions the subject's individuating conceptions are simply generalized away; for, as we see, in many *de re* ascriptions a particular subset of the subject's individuating conceptions is highlighted —so that not all of his conceptions which as a matter of fact are about the same thing count for the same. In section III of "Fregean Reference Defended," Sosa introduces the notion of "associative aboutness" to account for intuitions regarding what *res* is involved in some cases that he considers (the Donnellan case, the case of the Stalinesque picture and the case of the unsuccessful presidential candidate), which are at odds with his simpler theory. Sosa makes his theory coincident with our intuitions regarding those examples by making it more complex. In fact, if we analyzed the "opacity" examples indicating that in many *de re* ascriptions a particular subset of the subject's individuating conceptions is pointed out, we would see that the principles required to draw the needed distinction between "individuating conceptions" —introducing some principled discrimination between them— and to begin to establish a correct semantics for *de re* attributions would not be far away from the ones Sosa offers here. For in those examples of *de re* attributions in which not only the *res*, but also some of the subject's conceptions of it seem to be pointed to, our intuitions typically indicate that the conceptions involve those same acquaintance relations between the subject and the *res* (through perception, memory, or linguistic practices) that would probably constitute Sosa's "basic sources" in most cases.

I said earlier that the intuitions producing the counterexamples to any simple Fregean account of *de re* ascriptions are to be explained by the fact that the singular terms we use to indicate what *res* the ascribed thoughts are about might contribute specific information on the nature of the "individuating conceptions" constituting the ascribed thought to what is meant. It is important to notice that

---

[1]It is my view that, in some cases, the singular terms we use in attitude reports *just* contribute the information on the nature of the concepts to the characterization of the attitude (and not at all the object referred to by it), as when we say "John believed that you would be Chinese" in a context in which it is clear that we are attributing to John just a *general* attitude about the winner of the 100 meters race in the first Olympic Games disputed in the twenty–first century, whoever he or she was, and the winner happens in fact to be my addressee. But Sosa would not accept this view, and it is anyway controversial, so I am presupposing it is not correct in the main text.

this is so no matter what our theoretical account of the facts is. We could account for it in broadly pragmatic terms, interpreting such contribution as a matter of what is non–literally meant —as Sosa suggested in his earlier paper 'Propositional Attitudes *De Dicto* and *De Re*' (Sosa 1970), or as philosophers like J. Barwise and J. Perry, M. Richard, N. Salmon, S. Soames, and some others have proposed more recently; or we could instead rely on aspects of what is literally conveyed that are partially determined by semantics and partially determined by contextual facts–as philosophers like M. Crimmings, S. Schiffer and later selves of M. Richard and J. Perry have proposed.[2] No matter what our choice is, the point remains that very robust intuitions clearly show that something like Sosa's primitive account based on his FT principle cannot be correct. FT does not give us a sufficient condition for correct ascriptions of *de re* thoughts.

My first conclusion is then that the expectation of providing a very simple semantic account of *de re* ascriptions which could be raised by Sosa's suggestion cannot be fulfilled. By the time we have completed a correct account of all facts regarding the semantics of *de re* attributions, we shall have acknowledged the very same points on the crucial contribution to them of broadly causal aspects that the pristine Fregean theory purported to avoid. We shall have acknowledged that *de re* thoughts involve a particular subset of the much more general class of "individuating concepts": conceptions involving characteristics perceived, remembered, or preserved through a socially constituted chain of communication.

2

As I said, Sosa himself partly acknowledges this fact, and therefore the main motivation for his account is to be found elsewhere. His remarks make it clear that this motivation does not lie in semantic issues. Many philosophers have found relational properties offending, on several different (and sometimes opposing) grounds.

---

[2]Actually, I do not think that when the dust settles the difference between those accounts will be important enoughfor them to be counted as two. The friends of the Gricean, pragmatic account have to acknowledge that the non–literal meaning is in this case considerably more general and systematic than in paradigm cases of Gricean implicatures. The friends of the semantic "hidden–indexical" account have to acknowledge for their part that the contribution is in this case much less guided by specific linguistic conventions than in paradigm cases of semantically determined hidden–indexical contributions —like the spatial contribution in 'it is raining'.

To the extent that we take intentional properties seriously, those grounds give a bonus to a "narrow" account of them. One of the reasons most frequently mentioned for rejecting accounts of intentional properties which make causal relations partially constitutive of their nature mentions alleged difficulties of these accounts with intuitions regarding self–knowledge of the attitudes. Another which is more commonly pointed out nowadays is that relational theories would make intentional properties unexplanatory, or causally inefficacious. So says David Lewis, in a paper in which he puts forward views which I find close to the ones advanced here by Sosa: "[...] attributions of beliefs enter into a systematic common–sense psychology, and [...] for that purpose beliefs had better be in the head. [...] but [...] beliefs *de re*, in general, are not. Beliefs *de re* are not really beliefs. They are states of affairs that obtain in virtue of the relations of the subject's beliefs to the *res* in question". (Lewis (1983a), 151-2.)

After complicating his account to take care of the kind of problems we discussed before, Sosa suggests that it is this sort of consideration that plays a crucial part in his thinking. He insists that in his full account "no appeal is made to any causal relation, between term and object or concept and object," for although "[c]ausation does play an important role in associative aboutness," [this is the kind of more demanding aboutness needed to address the issues raised in the previous section], "it does so only in a derivative, 'by the way' fashion. The role played by causation derives from the way causation is bound to enter in the constitution of the epistemically basic individuators that make up the basic source of a referential conception." ("Fregean Reference Defended," end of section III.) David Lewis, in the paper where from where I quoted before, similarly remarks "I think of a thing as that which I am causally acquainted with in such–and–such way, perhaps perceptually or perhaps through a channel of acquaintance that involves the naming of the thing and my picking up of the name. I refer to that thing in my thought, and derivatively in language, because it is the thing that fits this causal and egocentric description extracted from my theory of the world and of my place in the world". (Lewis 1983b, 371.) Similar views are defended in Schiffer 1978.

The idea here is that the mechanism involved in the possession of *de re* thoughts in the more discriminating case when some conceptions and not others are involved is still the same broadly Fregean mechanism of the primitive account: it just involves having *de dicto* thoughts in which the object is presented through some description (perhaps one involving a mental indexical), which is as a matter of

fact uniquely satisfied by the object. It is just that the description implicitly or explicitly involves in some cases the concept of causation. This is how the theory is supposed to be properly "narrow" and to avoid the alleged problems raised by relational properties.

This seems to be promising but, unfortunately, I do not think any of the proposals so far advanced for narrow individuation of contents really works. Robert Stalnaker, for instance, has made compelling criticisms to suggestions in that direction by Daniel Dennett, Jerry Fodor and Brian Loar in a series of papers published a few years ago (Stalnaker 1989, 1990 and 1991). I shall indicate my misgivings about Sosa's related proposals in the rest of my commentary.

Let us assume that Sosa's proposals make the constituent of thoughts which contribute objects to their truth–conditions conveniently non–relational. (This is an assumption I grant only for strategic reasons; in the last resort, the assumption would fail on grounds which might be gathered from the ones I am about to discuss. I grant the point just because I want to avoid the usual unfruitful blandishing of intuitions, by choosing a more theoretical path of argument.) The obvious immediate worry concerns then the general concepts: how are their contents individuated in their turn? Are they properly "narrow"? These questions are pressing; the philosopher who really thinks that there is no other hope for the full–fledged causal–explanatory "reality" of mental properties than what is provided by non–relational individuation cannot afford to escape giving us some clues as to their answer.

Let me try to indicate how pressing these questions are and why the philosopher of Sosa's persuasion should make at least a start in the direction of offering some answer to them. Commenting on Putnam's infamous "model–theoretic" argument against realism, David Lewis suggests that the argument, instead of being a *reductio* of realism, is in fact a *reductio* of Putnam's premises; particularly, of the (futuristic) global descriptive theory of reference which, Putnam contends, is the only one available. (Lewis 1983b, 370-1; see also Lewis 1984 for a more detailed exposition.) This allows a dilemma to be posed for Sosa. Either he also has in mind a descriptive theory for the content of general concepts, including the concept of causation and all related concepts, subjunctive and nomic talk in general, or he does not. The first horn of the dilemma is not very appealing, if the anti–realist conclusion that Putnam and Lewis think follows from it does indeed follow from it. Moreover, lacking the "futuristic" aspects of Putnam's antirealism (i.e., the point that the descriptions in Putnam's theory are supposed to be provided by a theory satisfying strong epistemic requirements), the resulting variety of antirealism

would be deeply counterintuitive. What entities our thoughts are about cannot be just determined by descriptions of them furnished by our actual state of knowledge: otherwise, what seems to us to be correct would be always correct.

Consider now the second horn of the dilemma. Suppose that Sosa, like Lewis, does not find palatable the antirealist consequences of such an across–the–board descriptive theory of reference, and let us assume that he grants, with Lewis, that some general concepts somehow refer to *natural* properties, or *universals*. In most theories, the concept of *natural property* goes hand in hand with that of *objective law*, particularly *causal law*: natural properties are properties mentioned in true causal laws. Therefore, in this second alternative, both the interpretation of the causal concepts which are —according to Sosa's full account of singular aboutness— part of those of perception, memory and so on and the interpretation of some more general concepts would involve reference to natural properties. This poses a problem. In everybody's theory of natural properties, they are objective entities which admit of being thought in different ways as much as individual objects do. This suggests that we could generate the very same problems for property aboutness that Sosa's Fregean account of singular aboutness is intended to solve regarding object aboutness; that is to say, mental states individuated by natural properties would be relationally individuated as much as mental states individuated by objects are. The question then arises of how the sort of Fregean account Sosa relies on can handle general concepts. The first thing that comes to mind (it is not natural properties, but independently individuated descriptions of them that give the content of general concepts) takes us back to the first horn of the dilemma.

I think this is enough to show that there indeed are deeply–rooted doubts on how an account of *content* could really be *narrow*. Externalist theories of content are indeed intuitively at their most implausible when individual objects are concerned; moreover, the Strawsonian intuition that to have thoughts about a particular object requires having individuating conceptions of it of a peculiar sort (the concept of some acquaintance relation with it) is intuitively very well–entrenched. By combining these ideas, we come to the Fregean illusion that we will be able to design a properly individualistic descriptive theory of thought–content. We do this by conveniently forgetting about the issue of how to provide a coherent account of property–concepts. I think that the awareness that a coherently Fregean general theory of content is not in the offing is illuminating. It points towards (what I take to be) the fact that an easy solution —at least, one compatible with Intentional Realism— to the prob-

lems Sosa deals with in "Contents Fit for Explanation" is not to be had, and that the only way of tackling them requires us to revise our anti–externalistic intuitions.

In summary, I have cast doubt at the beginning of my contribution on the ability of Sosa's Fregean view, by itself, to give a proper account of the semantics of *de re* attitude–ascriptions. I have also expressed some misgivings about the possibility of extending Sosa's views to general empirical concepts in a way both compatible with Sosa's philosophical program and with the facts, and more in general on the viability of theories of narrow content.

REFERENCES

Lewis, D. (1983a): "Attitudes *De Dicto* and *De Se*", in D. Lewis, *Philosophical Papers*, vol. 1, Oxford: Oxford University Press.

Lewis, D. (1983b) "New Work for a Theory of Universals", *Australasian Journal of Philosophy*, 61, pp. 343-377.

Lewis, D. (1984) "Putnam's Paradox", *Australasian Journal of Philosophy*, LXII (1984), pp. 221-36.

Quine, W.V.O. (1966): "Quantifiers and Propositional Attitudes", in W.V.O. Quine *The Ways of Paradox*, New York: Random House.

Richard, M. (1990): *Propositional Attitudes*, Cambridge: Cambridge University Press.

Schiffer, E. (1978): "The Basis of Reference", *Erkenntnis*, 13, 171-206.

Sosa, Ernest (1970): "Propositional Attitudes *De Dicto* and *De Re*", *Journal of Philosophy*, LXVII, pp. 883-896.

Stalnaker, R. (1989): "On What is in the Head", in J. Tomberlin (ed.), *Philosophical Perspectives, 3: Philosophy of Mind and Action Theory*, Atascadero, California: Ridgeview Pub. Co., pp. 287-316.

Stalnaker, R. (1990): "Narrow Content", in C.A. Anderson and J. Owens (eds.), *Propositional Attitudes. The Role of Content in Logic, Language and Mind*, Stanford: CSLI, 1990, pp. 131-146.

Stalnaker, R. (1991): "How to Do Semantics for the Language of Thought", in B. Loewer and G. Rey (eds.), *Meaning and Mind. Fodor and his Critics*, Oxford: Basil Blackwell, pp. 229-238.

PHILOSOPHICAL ISSUES, 6
Content, 1995

More on Fregean Reference

Ernest Sosa

The comments of William Lycan, Manuel García–Carpintero, and others, in the SOFIA session and in subsequent discussion, have raised excellent questions that deserve full answers, and I would like to make a start on that in what follows.

1

Two main problems need to be faced, one familiar from the literature, and one that is new. The familiar problem arises from examples of false legend, myth, and media fantasy. Recall, for example, Kripke's supposition that the Biblical Moses never did the various things attributed to him in the Bible. That is to say, the story is definitely grounded in someone real, but the scribes get it nearly all wrong: in fact the man in question was not even named Moses (or its equivalent). Here it seems that our beliefs are not even associatively about that man, since our relevant referential conception is so thoroughly detached from him.

Compare a case[1] in which some flesh and blood English peasant

[1] One attributed to David Kaplan.

really grounds the Robin Hood legend, but, again, someone who lacks nearly all the attributes of the legendary Robin Hood. Here again our referential conception would seem to fit our man too loosely to explain how it is that he is really our referent.

Suppose, finally,[2] a body of lore that we have acquired through television or through tabloids, about some celebrity or other public figure. "The body of lore is a pack of lies, completely false, but it is told of a real person to whom our only access is the television or tabloid reports themselves".

It will be useful in dealing with such cases to introduce the following concept:

> $\sigma\alpha$, the source correlate of α, is the concept: [the one that my source refers to (my sources refer to) as α].

The concept of reference in play here is, let us suppose, just ordinary, everyday reference or aboutness; accordingly, it will often amount to associative aboutness, if our earlier reasoning is acceptable.[3]

In the case of Moses, therefore, the more familiar members of our referential conception Ω —namely, [Moses], [the deliverer of the Israelites], [the receiver of the tablets], etc.— will help induce an epistemic basis $\epsilon(\Omega)$ containing, among others, the individuators: σ[Moses], σ[the deliverer of the Israelites], σ[the receiver of the tablets], and so on.

Accordingly, our account bids fair to deliver the desired result. Our thoughts *will* still be about that Israelite who "grounds" the Biblical narrative, even if that narrative is nearly all wrong. For the individuators in the *epistemic basis* of our referential conception will be about that grounding Israelite: accordingly, our referential conception will be collectively about him, despite the almost total falsity of the Biblical narrative.

A similar response can be developed for the Robin Hood case (insofar as that can be turned into a persuasive case where our thoughts, as we go through the story, can be said to be about the grounding peasant). And, finally, a similar response seems also available for the case where we are thoroughly misled by the disreputable media.

---

[2] As was suggested by Lycan.

[3] See "Fregean Reference Defended" in this issue of *Philosophical Issues*.

2

We have considered a first sort of case that seemed initially problematic, but less so on closer inspection. Here now is a case that becomes increasingly threatening with reflection. In its simplest and most general lines:

> A child is put to bed and thinks later that he dreamed of someone with a friendly face, when actually it was someone real, a long absent aunt who was brought into the room for a peek at the "sleeping" child. In a haze, past heavy eyelids, the child did catch a clear enough glimpse of his aunt's face, which he had never seen. Even if the aunt went unrecognized, therefore, and even if the child knows nothing of her, even that she exists, when he later speaks or thinks of "the kindly lady in my dream", it is really his aunt that his speech or thought is about.[4]

Yet our account as it stands seems unable to accommodate this case. The child seems not even associatively to refer to the aunt, since the child appears to lack any referential conception with an epistemic basis preponderantly about the aunt. Having apparently no knowledge (or belief) of her existence, he seemingly lacks any individuator a of the aunt, any individuator a that uniquely specifies the aunt for him. Accordingly, no individuator would seem to appear in any epistemic basis that, from the perspective of the child, specifies the aunt.

That *seems* accordingly a true counterexample to our account. But is it really? The child does have a referential conception Ω of the aunt, after all, even if it is rather thin. Ω here would include [the woman in my conscious experience last night] or the like, and other such individuators. Here is what I think: The child does refer to his aunt, since, as it turns out, his aunt *is* the woman in his conscious experience (even though he also turns out wrong in thinking his conscious experience to have been just a dream). This is the line of response that I find most promising, though of course it still needs to be developed.

The concept of associative aboutness gives us therefore a way to capture a familiar and often invoked conception of mental reference and it does so without exceeding in any significant way the resources available to a fregean approach —one that fundamentally understands reference in terms of specification through a concept uniquely

---

[4]My thanks for this fascinating case to Jerrold Katz.

satisfied (perhaps relative to the context of use) by the referent thus specified. What we now are better able to see is only that the full understanding of "reference" in terms of such specification may have to be more indirect and involved than might once have appeared. Reference turns out to have more varieties than was once suspected, but is also more amenable to fregean explication than has seemed to many.

Our simple answer depends, however, on a move that will strike some as philosophically dubious: intentional objects, they will say, such as "a woman in one's conscious experience", *cannot* be strictly identical to flesh and blood people. Strictly, then, the woman in the child's experience can't have been his flesh and blood aunt. Our answer accordingly requires considerable development and support. For a start, I can sketch my preferred solution as follows. We must allow experiences as well as beliefs in the epistemic basis of a referential conception and we must admit a *second* form of epistemic derivation: derivation of a sort whereby [the kindly lady in my dream], as this figures in the child's later beliefs, can derive from [the woman that looks thus and so], as this figures in the earlier experience —despite the time gap involved and despite the subject's failure to believe that whatever satisfies the latter satisfies the former. There is surely some sense in which the former still epistemically derives from the latter. Here is a sketch of how that sense might be rendered explicit and how it might be used to expand our definition of epistemic derivation:

D'_2. Individuator β *epistemically derives* from individuator α for S at t iff *either* (a) S at t believes (justifiably) [there is such a thing as β] on the basis of believing (justifiably) [there is such a thing as α] and [whatever is α is β], but not vice versa; *or* (b) S at t believes (justifiably) [there is such a thing as β] on the basis of experiencing at t' ($t' \leq t$) [there is such a thing as α] or [here before me now is α], or the like, where α constitutes at time t' the "realization" of β, in S's experience of that time.[5]

3

García–Carpintero has three main problems for me:

---

[5]Where α constitutes the "realization" of β in the way [the lady that appears thus and so], as that helps constitute S's earlier conscious experience, "realizes" the later individual concept invoked by S when he later thinks of "the kindly lady in my dream".

A. First, my fregean account of *de re* thought, FT, cannot be supported by means of the simple argument I suggest, since the crucial terms of that argument (such as 'proposition' and 'aboutness') are too "theoretically charged... for pure intuitions to support claims involving them without further theoretical considerations" (p. 105).[6]

The argument involved in problem A crucially employs the notion of a term's being "theoretically charged". García–Carpintero's argument (GCA) might perhaps be put like this:

GCA
1. Term T is theoretically charged.
2. If 1, then (3) any claim P involving T cannot be supported by pure intuitions without further theoretical considerations.

Therefore, 3.

However, argument GCA itself uses at least four terms that *seem* "theoretically charged": (i) 'theoretically charged'; (ii) 'pure intuitions'; (iii) 'intuitions'; and (iv) 'theoretical considerations'.

Secondly, the further, theoretical, considerations involved in buttressing a claim P would often be further philosophical arguments, presumably, when P is itself a philosophically interesting claim. But if P involves essentially the term T, then some at least of these further arguments must also involve the term T in *their* premisses, if those premisses are to be appropriately relevant to the claim P. Take now a claim P' involved in such a further argument on behalf of P. Can't we now apply argument GCA to P' as well, seeing that it itself contains the, *ex hypothesi*, "theoretically charged" T?

Fortunately, it is not really clear that a claim with a theoretically charged term cannot derive some support from pure intuition even in the absence of further theoretical considerations. The following claims all *seem* counterexamples (at least until we hear more about the meaning of 'theoretically charged'):

a. If S *knows* that snow is white, then it is *true* that snow is white.

b. Snow is white iff it is *true* that snow is white.

c. $(\forall x, y)$ (If *molecule* x is *larger* than *molecule* y, then $x \neq y$).

---

[6]Parenthetical references here are to Manuel García–Carpintero's "Doubts on Fregean Reference", this journal, this issue.

This list could apparently be extended extensively. The support provided by pure intuition may not make such claims incontrovertible; nevertheless, it is surely strong support that can properly and responsibly be adduced even in the absence of further supporting theoretical considerations, if one can thereby buttress a conclusion that itself has some interesting theoretical use. This is what I propose with regard to the conclusion FT of the simple argument offered early in my paper.

It might be replied that GCA does not adequately capture the intended objection to the appeal to intuition. For example, perhaps the intended argument is restricted to philosophical terms that do not have an established use in the language prior to philosophical reflection. Let us briefly examine this possibility. The simple argument of mine under attack is an argument for the following conclusion:

(FT) A subject S has at time t a thought (belief, intention, etc.) *about* x (*of* x) if S thinks (believes, intends, etc.) *de dicto* a proposition that predicates some property ϕ with respect to some individuating concept (or individuator) α of x for S at that time.[7]

My simple argument is as follows:

P_1. If there is such a thing as the F then the proposition that the F is G is about the F and attributes being G to the F.

P_2. If one believes proposition P, and P is about x and attributes being G to x, then one's belief is about x and attributes being G to x.

C. Therefore, FT.

This argument shows, I contend, that FT is right for *some* natural and familiar senses of 'about' and 'of' (and with respect to associated constructions such as 'attributing a property to a thing'). Yet many have rejected FT and its associated aboutness. The new objections due to García–Carpintero put in question the worth of the intuitive support for P_1 and P_2. The crucial terms involved —'proposition' and 'about'— are supposed (i) to be "theoretically charged", and (ii) to have no established use in the language prior to philosophical

---

[7]FT permits individuating concepts to be "perspectival" or "indexical", since they need only be satisfied not absolutely but in a "perspective", one that supplies the indices or parameters required for such a concept to be satisfied. Note also that α is an individuator *of* x only if x exists.

reflection. And this is thought to reduce significantly or to remove altogether any probative force in favor of P_1 or P_2 that may be thought to derive from its intuitive plausibility.

Even with the specified restriction, I still cannot find persuasive the argument against the appeal to intuition in behalf of P_1 and P_2. I do not find much significant difference in relevant status among the following terms: 'knowledge', 'truth', 'identity', 'proposition', 'about'. All of them are much disputed in philosophy, in various respects and in various contexts. So much for their being terms of philosophical theory. On the other hand, all of them are found in our dictionaries, with at least one established use apiece.[8] So much, presumably, for their having an established use in the language "prior to philosophical reflection". Accordingly, it is not easy to see how to deny P_1 and P_2 any support from intuitive plausibility (for *some* established senses of the terms involved) while still allowing such support to the likes of (a)-(c). Therefore, either we withdraw the support of intuition to all five propositions —P_1, P_2, (a)-(c)— or we allow that support to them all. My own preference should be clear.[9]

What is more, if someone objects, on ontological or other grounds, to the use of any concept of proposition in premises like P_1 or P_2, one could in fact replace my simple argument with one that uses sentences in place of propositions.

Let us turn now to the second of the alleged problems for my view:

> B The "...expectation of providing a very simple semantic account of *de re* ascriptions which could be raised by Sosa's account cannot be fulfilled" (p. 108).

The operative argument, boiled down to its essentials, appears to be this:

1. There are contexts, easy to set up, such that an utterance of 'John believes that Cicero was a senator' —while still intending to attribute a *de re* attitude— provides information not only on what *res* the attitude is about, but also on the character of the subject's conception of it. The intuitions are to be explained by the fact that when we ascribe *de re* thoughts, the

---

[8]The terms 'proposition' and 'about' do seem to have established uses "prior to philosophical reflection", as is presumably shown by their presence in every dictionary of the English language.

[9]García–Carpintero may perhaps wish to place further restrictions on the sort of term against which his GCA is meant to be applied. But these restrictions are yet to be specified, and we must then see whether or not they are met by the relevant terms in P_1 and P_2.

> singular terms we use to indicate what *res* those thoughts are about do not contribute just that *res* to the individuation of what we mean.... The singular terms we use in many attitude–ascriptions contribute to what we mean some additional more specific information on the nature of the "individuating concepts" (p. 106).

But how can this really be a problem for my paper, which takes no position on the semantics of attitude ascriptions or on what "information is provided" by such ascriptions? What is relevant in my paper concerns only conditions within which thoughts, affirmations, and individuating concepts are "about" or "of" individuals, or "refer" to such. (See its p. 109 for an outline of its main strategy.) Nowhere do I suggest anything about the truth conditions for an attitude ascription such as 'John believes that Cicero was a senator'. About such ascriptions I would in fact say the following, all of which coheres well with my view of fregean reference, and also with much of what García–Carpintero has to say. Here is how I view such ascriptions:

1. In saying "John believes that Cicero was a senator" one might be saying any of at least three quite different things whose truth conditions are, respectively, as follows:

   a. That John believes the proposition that predicates the property ⟨x is a senator⟩, i.e., the property of being a senator, with respect to the individuating concept ⟨Cicero⟩; where one takes no position on whether or not there is anything that instantiates or satisfies that individuating concept. (For those with an aversion to propositions, we could offer the following alternative as the relevant truth condition: *that John believes–true the sentence 'Cicero was a senator'*; where one takes no position on whether the name 'Cicero' refers. And similar, more linguistic, alternatives might be concocted for b and c below.)

   b. That John believes some proposition which predicates being a senator with respect to some individuating concept satisfied by Cicero (where *any* such individuating concept will do).

   c. That John believes some proposition which predicates being a senator with respect to the individuating concept ⟨Cicero⟩, where in fact Cicero satisfies this individuating concept.

2. The attitude ascription "John believes that Cicero was a senator" can be rephrased in three different ways, corresponding to 1(a)–1(c) above, as follows:

a. "John believes the proposition that Cicero was a senator (though I do not say that there really is such a being as Cicero)".

b. "John believes *of* Cicero that he was a senator (though he does not necessarily believe it *of* Cicero *as* Cicero)".

c. "John believes *of* Cicero *as* Cicero that he was a senator (i.e., he believes the proposition that Cicero was a senator, a proposition that is in fact about Cicero)".

We turn, finally, to the third of the main problems raised by García–Carpintero:

C. The sort of fregean account defended in my paper is said to lead "... to the Fregean illusion that we will be able to design a properly individualistic theory of thought–content [in general.... The fact is, however,] that an easy solution... to the problems Sosa deals with... is not to be had, and that the only way of tackling them requires us to revise our anti–externalist intuitions" (p. 111).

Here the reasoning seems to be as follows. We would like an account of individual reference, the individual reference or content of individuating concepts or singular terms, but that is not all. There is also the problem of how we get properties to be part of the (predicational) content of our thoughts or affirmations. And even if individual, singular reference can be explained adequately along fregean lines in the way suggested, the wider project of defending narrow content seems indefensible or anyhow has not been defended adequately. To this objection I sketch now a reply.

First, my paper adopts no narrow content approach *in general.* Indeed, the fregean proposal defended in the paper concerning singular reference is coherently combinable with externalism and with a causal account of *predicative* content. The issue of narrow content in general is much discussed and very complex. But, again, (a) it is not one that I take up in my paper, and (b) what I do defend about exportation and singular reference does not commit one to any particular view on narrow versus wide content in general.

Note, however, that our thoughts can involve properties in two very different ways: (a) as referents, and (b) as predicables. Nor can it be that *whenever* they enter as predicables, they do so as referents: that way lies vicious regress. Thus when I say "This is red", pointing to an apple's surface, and thus predicate redness of that surface, I need not be *referring* to redness and saying of the apple's surface that it

has that property. For, if so, we would next face the similar question that arises when I say that x *has* F–ness: Am I then *referring* to exemplification and saying that the ordered couple $\langle x, F\text{–ness}\rangle$ *has* or *exemplifies* it? If so, we are on the way to our vicious regress. But, if not, one might wonder how *exemplification* can be involved in our thought when we say or think that the surface exemplifies redness, if not by being *referred* to. Answer: By being *predicated*, apparently. And in that case there is no obvious reason why redness itself cannot also be involved thus in our thought, by being predicated (of the surface) without *ipso facto* being also necessarily referred to.

In another thought, however, redness might be referred to without being predicated. Thus if I say "the color of the fruit in my bag is my favorite color" I would seem to be referring to redness without *ipso facto* predicating it.

In my paper I focussed only on reference, and said little about predication. But the fregean view defended is meant to apply not only to concreta such as martini–drinkers, but also to abstracta such as redness. This provides a further advantage for that view, since it does not *require* for reference to abstracta that we enter into causal relations with them, but only that we be able to specify them, either individually or associatively. And even associative reference requires only an epistemic basis preponderantly about the object of associative reference. And, finally, although associative reference to concreta will most often require causal relations of perception in the relevant epistemic basis (at some remove, via memory or testimony, if not immediately), associative reference to abstracta does not obviously require any such causal connections). How we manage to refer to abstracta in the absence of such causal intercourse is a fascinating question as well. But I hope it will not be thought that just leaving that an open question is a substantial *problem* for my fregean view in particular, one that can give the basis for a serious objection to it.

PHILOSOPHICAL ISSUES, 6
Content, 1995

Mental Causation: What? Me Worry?

Jaegwon Kim

The problem of mental causation, broadly, is that of explaining how it is possible for mental events to enter into causal relations with other events, whether mental or physical. Somewhat more narrowly, it is the problem of explaining how it is possible for mental events to causally affect the (physical) behavior of physical systems. Although the problem was coeval with the mind–body problem (Descartes, arguably, invented both) and was thought to be a special difficulty of Cartesian substance dualism, it has recently re–emerged, with a vengeance, within the largely materialist/physicalist framework which most of us accept today. The problem has been under intense debate during the past dozen years or so; it has been one of the focal points of controversy in recent philosophy of mind and psychology. The outcome of the debate is often thought to have implications for some issues of fundamental importance, such as the reality of the mental, the nature of psychological explanation both in everyday life and in systematic psychology, the status of psychology as a science and its relation to the physical–biological sciences, and our conception of ourselves as deliberators and agents. To some of us, the problem of mental causation, along with the problem of the

phenomenal and subjective character of the mental, constitutes the heart of the mind–body problem, the problem of understanding the place of mentally in an essentially material world.

One reaction on the part of some philosophers to this development, which appears to be gaining some momentum, is to try to dissipate the problem by arguing that there in fact is no such "problem" at all, or at any rate to downplay the philosophical significance of the problem. Thus, it has been argued that worries about mental causation arise out of our misplaced philosophical priorities; that our overindulgence in unmotivated metaphysical principles is the source of the problem; that an improper understanding of the causal relation as an extensional relation is at the root of the misconceived worries about the causal efficacy of the mental; that if there is a problem about mental causation, the same problem arises for all the special sciences, such as biology and chemistry, in their relation to more basic, lower–level sciences; that we should look to explanations and explanatory practices, not to metaphysics, to gain a proper understanding of mental causation; and so on.

This paper has two aims. One is to look into some of these deflationary arguments in some detail. In particular, I will discuss the following two questions: (1) Is it really the case that we can dissipate the problem by relying on our "explanatory practices" involving mental events and ignoring, or de–emphasizing, our metaphysics? (2) Is it true that the problem of mental causation, as it is usually formulated, generalizes to all special sciences and if so to what extent? Or are there special problems and difficulties that are unique to the mind–body case? I will argue that there is little philosophical illumination to be gained from looking merely toward psychological explanation for guidance on mental causation. The reason, briefly, is that as long as we take a reasonably realist stance on the nature of explanation (in particular, causal explanation), the metaphysical problem of mental causation will not go away. In fact, the possibility and reality of mental causation is arguably a precondition of the possibility of (causal) psychological explanation. Concerning (2), the situation turns out to be more complex than it is often thought to be, and I hope it will be seen that there is no quick and simple way to generalize the problem of mental causation to other special sciences, and that there are things to be learned from taking a fresh look at the familiar assumptions we make about hierarchically organized "levels" of entities and their properties when we think about physics and its relationship to the special sciences.

My second aim is to try to see, along the way, just how the problem of mental causation arises: what are the *assumptions*, metaphysical

or otherwise, that generate the problem? It will be seen, unsurprisingly, that there are alternative sets of assumptions that make mental causation prima facie problematic. Need for an account of the possibility of mental causation can arise from different sources, and, moreover, what counts as a reasonable "solution" will likely depend on the specific pressure that has generated the problem. Various metaphysical commitments, individually or in combination, threaten mental causation, or at least appear to do so with sufficient force and credibility, to create a prima facie problem for us. The following are some of the more familiar players on the scene, although not all of them will figure in the discussion to follow: the causal closure of the physical, mind–body supervenience, causal/explanatory exclusion, mind–body anomalism, content externalism, the normativity of the mental, and antireductionism.

1

In an interesting recent paper, Tyler Burge addresses what he takes to be unmotivated worries about epiphenomenalism. He writes:

> But what interests me more is the very existence of the worries. I think that they are symptomatic of a mistaken set of philosophical priorities. Materialist metaphysics has been given more weight than it deserves. Reflection on explanatory practice has been given too little. The metaphysical grounds that support the worries are vastly less strong than the more ordinary grounds we already have for rejecting them.[1]

What, on Burge's view, are these "ordinary grounds" we have for rejecting the "metaphysical grounds" that generate worries about mental causation? Burge's answer: the success and "probity" of mentalistic explanations. He urges:

> I think it more natural and fruitful to begin by assuming, defeasibly perhaps but firmly, that attributions of intentional mental events are central to psychological explanation both in ordinary life and in various parts of psychology. We may also assume that intentional mental events are often causes and that psychological explanation is often a form of causal explanation. Given these assumptions, the 'worry' about epiphenomenalism seems very remote. For if intentional mental events... enter

---

[1]Burge, "Mind–Body Causation and Explanatory Practice", in *Mental Causation*, ed. John Heil and Alfred Mele (Oxford: Clarendon Press, 1993), p. 97. See also Burge's shorter discussion in his "Philosophy of Language and Mind: 1950-1990", *Philosophical Review*, 101 (1992): 3-51; see pp. 36-39.

into causal relations and are cited (in terms of those aspects) in explanations, then there seems to be every reason to conclude that those aspects are causally efficacious. None of the metaphysical considerations advanced in current discussion seem to me remotely strong enough to threaten this conclusion.[2]

Lynne Rudder Baker is another recent writer who has urged a view much like Burge's: what we need to do to neutralize the problem of mental causation is to reverse our priorities as between metaphysics and explanation. She writes:

> My suggestion is to take as our philosophical starting–point, not a metaphysical doctrine about the nature of causation or of reality, but a range of explanations that have been found worthy of acceptance.... If we reverse the priority of explanation and causation that is favoured by the metaphysician, the problem of mental causation just melts away. We begin with the question: Does what we think ever affect what we do?... With the reversal of priority of cause and explanation, the metaphysical version of the question just does not arise, and the original question has an easy answer.[3]

What is this "easy answer"? To see that Jill's thought that she left her keys on the counter and her wanting them back caused her to return to the bookstore, all we need to do, Baker says, is to appreciate the following "explanatory fact": "If she hadn't thought that she had left her keys, then, other things being equal, she wouldn't have returned to the bookstore; and given that she did think that she had left her keys, then, other things being equal, her returning was inevitable".[4] Burge appears to have something pretty similar in mind when says: "...one can specify various ways in which mental causes 'make a difference' which do not conflict with physical explanations. The differences they make are specified by psychological causal explanations, and by counterfactuals associated with these explanations".[5]

Let us briefly look at Baker's suggestion. Her idea is to explain causation in terms of explanation (so that "causation becomes an explanatory concept"[6]) and then explain explanation in terms of

---

[2]Burge, p. 118.

[3]Baker, "Metaphysics and Mental Causation", in *Mental Causation*, ed. Heil and Mele, pp. 92-93.

[4]Baker, p. 93.

[5]Burge, p. 115.

[6]Baker, p. 93.

appropriate counterfactuals and the "inevitability" of the outcome given the putative cause. We can agree with Baker, and Burge, that our confidence in the reality of mental causation is grounded, substantially if not wholly,[7] in our acceptance of such explanations and counterfactuals, and that our explanatory practice involving intentional states and actions must be respected in any discussion of mental causation. However, it is not difficult to see that Baker's proposal will not make the need of further metaphysical clarification go away. To see this consider the epiphenomenalist who claims that some neural state, N, was the cause of both Jill's thought that she left her keys at the bookstore counter and of her returning to the bookstore. The epiphenomenalist may well be prepared to accept both clauses of Baker's explanatory/counterfactual gloss on the causal claim, for he might reason as follows: "If Jill's thought had not occurred, then Jill would not have been in N, and given that Jill's thought did occur, N must occur, and this made Jill's return inevitable". It seems to me that there is no incoherence in this epiphenomenalist account of the situation. I am not saying that an epiphenomenalist account of this form will work in every such situation; the point is that it is not ruled out by Baker's proposals.

In more general terms, the counterfactual test is a poor test to assess causal directionality: given that c caused e, the counterfactual "if e had not occurred, c would not have occurred either" —can often be defended, on almost any model of counterfactuals.[8] Also, when e_1 and e_2 are collateral effects of a single common cause c, the counterfactual "If e_1 had not occurred, e_2 would not have" and its converse can both be true, and there are likely to be cases of this kind where it is true to say that "given e_1, e_2 was inevitable". Moreover, the concept of "inevitability" used in Baker's proposal is badly in need of further clarification; the idiom "given c, e is inevitable" sounds very much like just another way of saying "c (causally) necessitates e" —a concept that has been the nemesis of the Humeans. Unless Baker can provide us with a nonquestion–begging explanation of the notion of inevitability involved, her project of replacing causation with explanation must be judged incomplete at best. The problem of mental causation won't just "melt away".

---

[7]For surely our belief in mental causation must arise also from our view of ourselves as agents —that is, the belief that our desires, beliefs, and intentions can move our limbs and thereby manage to rearrange things around us.

[8]David Lewis is an exception; but to eliminate such "back–tracking counterfactuals" Lewis has to bring out heavy–duty metaphysical armor. See Lewis, "Counterfactual Dependence and Time's Arrow" and "Causation", both reprinted in his *Philosophical Papers* II (New York and Oxford: Oxford University Press, 1986).

2

There is a respect in which Baker and Burge are right. As Burge says, our confidence in the truth of familiar intentional explanations does, and should, vastly exceed our commitment to any recondite metaphysical principles. In this sense, the epiphenomenalist "worries" are overstated. But I doubt that very many of us who have "worried" about mental causation are actually concerned about the possibility that epiphenomenalism might turn out to be true, that our thoughts and desires might turn out to have no causal powers to move our limbs. Our worries are not evidential or epistemological worries. In this sense, Burge is right when he says that there is an air of make–believe about the epiphenomenalist threats (he likens epiphenomenalism with epistemological skepticism). But what all this shows is that the problem of mental causation is primarily a theoretical metaphysical problem. It is the problem of showing how mental causation is possible, not *whether* it is possible. In raising the how–question, we are assuming, "defeasibly but firmly" as Burge says, that the whether–question has already been affirmatively answered.[9]

But metaphysical questions don't pop up in a vacuum. The how–question of mental causation arises because there are certain other commitments, whether metaphysical or of other sorts, which demand our respect but which make mental causation appear prima facie problematic. The issue is not metaphysics versus explanatory practice, as Burge would have it, or metaphysics versus epistemology, as Baker would have it. Nor is the issue one of choosing between metaphysics and mental causation: most of us have already chosen mental causation, although defeasibly —as philosophers we should regard pretty much everything ultimately negotiable. The issue is *how to make our metaphysics consistent with mental causation*, and the choice that we need to make is, at least usually, between various *metaphysical alternatives*, not between some metaphysical principle on the one hand and some cherished epistemological practice or principle on the other. I say, "usually", because, as we will see below, epistemology and metaphysics do have something to do with each other, and choices we make in one can require adjustments and accommodations in the other.

Would the metaphysical problem of mental causation politely take its leave if we did less metaphysics, as Burge and Baker urge, and

---

[9]So I disagree with Burge when he says, "Epiphenomenalism is often taken as a serious metaphysical option", p. 102-103. But see Peter Bieri, "Trying out Epiphenomenalism", *Erkenntnis*, 36 (1992): 283-309.

focused our attention on psychological explanation? Burge says that "our understanding of mental causation derives primarily from our understanding of mentalistic explanation".[10] But what is our understanding of mentalistic explanation? Burge doesn't address this question directly, but we can see roughly what he has in mind: such explanations are often causal explanations. In the second quoted passage above he says that "we may assume that intentional mental states are often causes and that psychological explanation is often a form of causal explanation". As you will recall, there used to be a debate, no less intense than the current debate on mental causation, in the 1950s and '60s, whether or not belief–desire explanation of action is causal explanation (that is, whether "reasons are causes"), and the noncausalists had the upper hand for many years, until Davidsonian causalism became the new orthodoxy.[11] If, as Burge says, we "may assume" that belief- desire explanation is a form of causal explanation, we owe this license substantially to Davidson (at least, I know I do). And what carried the day for causalism was Davidson's philosophical argument, not the pervasiveness in our life of the practice of rationalizing action in terms of belief and desire. There was no disagreement on the latter point; in fact, it was the presupposition of the entire debate.

Moreover, Davidson's own analysis of rationalizing explanation as causal explanation involved not an insubstantial amount of metaphysics —e.g., an analysis of singular causal statements and a view about events and their descriptions. A full statement of his argument is likely to implicate the whole metaphysical package of "anomalous monism", including the highly controversial doctrine of "the anomalism of the mental". Much of the current debate on mental causation has stemmed from a widely shared dissatisfaction with Davidson's account —the worry that Davidson's theory does not provide mental kinds and properties with an appropriate causal role in behavior production.[12] Surely, this brief history should suffice to persuade

---

[10]Burge, p. 103.

[11]See Donald Davidson, "Actions, Reasons, and Causes", *Journal of Philosophy*, 60 (1963), reprinted in his *Essays on Actions and Events* (Oxford: Cleared Press, 1980). See G.H. von Wright, *Explanation and Understanding* (Ithaca: Cornell University Press, 1971) for what may have been the last major statement of the noncausalist position.

[12]To cite a few of the papers in which this issue has been raised: Frederick Stoutland, "Oblique Causation and Reasons for Action", *Synthese*, 43 (1980): 351-67; Ted Honderich, "The Argument for Anomalous Monism", *Analysis*, 42 (1982): 59-64; Ernest Sosa, "Mind–Body Interaction and Supervenient Causation", *Midwest Studies in Philosophy*, 9 (1984): 271-81; Jaegwon Kim, "Self–Un-

us that we cannot easily insulate explanatory practices involving intentional states from metaphysical problems. The question whether rationalizations are a species of causal explanation itself involves substantive metaphysical issues.

Even after we have answered this question —affirmatively, let us assume— metaphysics still won't go away. The only way in which I believe we understand the idea of a causal explanation involves the idea that the event invoked in a causal explanation is in reality a cause of the phenomenon to be explained. That is, if c (or a description or representation of c) causally explains e, c must be a cause of e. If my desire for a drink of water causally explains my body's movement toward the kitchen, the desire must be a cause of the bodily movement. I take this to be an untendentious and uncontroversial point.

Suppose then that my desire for water causes a certain motion of my body. This is a case of mental–to–physical causation. So far so good. But metaphysical problems begin to emerge, in several ways. First, suppose we trace the causal chain back from my bodily motion —to simplify, the movement of my left foot as I take my first step toward the kitchen. I assume we have a pretty good neurophysiological story about how such a limb motion occurs, a story involving transmission of neural signals, contraction of a group of muscles, etc.; let us suppose that the story ends with some neural event in my central nervous system, presumably the activation of a group of neurons somewhere. There seems every reason to think that such a neurophysiological causal explanation also exists; at least, we cannot rule out such a possibility. What then is the relationship between this explanation and the intentional explanation in terms of my belief and desire? One invokes a neural state, N, as a cause of my foot movement; the other invokes my desire for a drink of water, as a cause of the very same event. How are these two causes related to each other?

When we are faced with two purported causes, or causal explanations, of a single event, the following alternative accounts of the situation are initially available: (a) each is a sufficient cause and the effect is causally overdetermined; (b) they are each necessary and jointly help make up a sufficient cause (that is, each is only a "partial cause"); (c) one is part of the other; (d) the causes are in

---

derstanding and Rationalizing Explanations", *Philosophia Naturalis*, 82 (1984): 309-20; Louise Antony, "Anomalous Monism and the Problem of Explanatory Force", *Philosophical Review*, 98 (1989): 153-87. Davidson defends his views in "Thinking Causes", in *Mental Causation*; see responses by Kim, Sosa, and Brian McLaughlin in the same volume.

fact one and the same; (e) one (presumably the mental cause in the present case) is in some appropriate sense reducible to the other; (f) one (again the mental cause) is a derivative cause; its causal status is dependent, or supervenient, on the neural cause, N. Perhaps, there are others, but it is clear that for our present case, most of them, including (a), (b), and (c), are nonstarters. The general point I want to stress, though, is this: the presence of two causal stories, each claiming to offer a full cause for a given event, creates an unstable situation requiring us to find an account of how the two purported causes are related to each other. This is the problem of "causal/explanatory exclusion".[13]

Burge, as I take it, would reply that intentional explanations and physiological explanations need not, and do not, compete with each other. He says:

> It would be perverse to think that the mentalistic explanation excludes or interferes with non–intentional explanation of the physical movement. I think that these ideas seem perverse not because we know that the mental events are material. They seem perverse because we know that the two causal explanations are explaining the same physical effect as the outcome of two very different patterns of events. The explanations of these patterns answer two very different types of inquiry. Neither type of explanation makes essential, specific assumptions about the other. ... The perversity of thinking that mental causes must fill gaps in physical chains of events probably has its source in traditional dualism, or in libertarian worries about free will.[14]

No, "the perversity" has nothing to do with dualism or free will; it has only to do with two causal claims, each purporting to provide a sufficient cause of a single effect. The interesting fact about explanations that Burge seems to miss is that two *explanations can be rival explanations even though their explanantia are mutually consistent and true, if they purport to explain (in particular, causally explain) an single explanandum.* It makes no difference that the two explanations arise from different areas of inquiry, involving distinct vocabularies, or that they are responses to two different epistemic or pragmatic concerns. Thus, a car accident is explained by an engineer as having been caused by the incorrect camber of the highway curve, and by a police officer as having been caused by the impaired

---

[13]See for details my "Mechanism, Purpose, and Explanatory Exclusion", *Philosophical Perspectives*, 3 (1989): 77-108; reprinted in my *Supervenience and Mind* (Cambridge: Cambridge University Press, 1993).

[14]Burge, p. 116.

driving of a drunken driver. But in such a case we naturally think of the two causes as partial causes; they together help make up a full and sufficient cause of the accident. As long as each claims to provide a full cause of the event to be explained, an epistemic tension is created and we are entitled to ask, indeed need to ask, how the two purported causes are related to each other. In fact, it is because "neither... explanation makes essential, specific assumptions about the other" that we need to know how the two explanations are related, how the two causal stories about a single phenomenon mesh with each other. Are the two stories at bottom one story couched in different languages? Do the two stories supplement one another, each being only partial? And so on. Metaphysics is the domain where different languages, theories, explanations, and conceptual systems encounter one another and have their ontological relationships clarified. If you believe that there is no such common domain, well, that's metaphysics, too.

The problem of causal/explanatory exclusion arises if there are cases of psychological explanations of physical behavior in which we are prepared to believe that the physical effect has, or must have, a physical causal explanation as well. And it seems to me that we do not need to subscribe to a general doctrine of the causal closure of the physical domain to believe that there indeed are such cases. To appreciate the exclusion problem we do not require much heavy–duty metaphysics —overarching doctrines about mental anomalism, "strict laws" in causal relations, a physical/mechanical conception of causality, token physicalism, and the rest. It arises from the very notion of causal explanation and what strikes me as perfectly intuitive and ordinary understanding of the causal relation. If this is right, turning away from metaphysics to embrace epistemology, or away from causation to embrace explanation, will not dissipate the need for an account of mental causation. There is a short and straight route from mentalistic explanation to mental causation, and from the possibility of dual explanations of a physical occurrence to the vexing problem of mental causation.

3

Baker, Burge, and others are of course right in pointing out that the problem of mental causation is primarily a problem that arises from metaphysical assumptions. But it seems to me that some of the assumptions that generate the problem are familiar ones which are widely accepted. Consider, for example, the thesis of mind–

body dependence. Both Burge and Baker explicitly acknowledge some form of mind–body supervenience or dependence. Burge says: "There are surely some systematic, even necessary, relations between mental events and underlying physical processes. We have good reason to believe that mental processes depend on underlying physical processes".[15] Although Baker thinks that the thesis she dubs "SS" (for strong mind–body supervenience) is an "idle speculation" (since we will never know, she says, all the microphysical bases on which mental properties supervene), she is willing to let it be and attack the idea of causation instead.

Suppose, then, that my desire for a drink of water supervenes, or depends, on a neural state, N. Again, it is not difficult to see that N can turn out to be a potential competitor of my desire as a cause of the movement of my foot. The issue here depends, at least partly, on what "dependence" or "supervenience" is supposed to mean. On the standard account (which Baker accepts), the base property is a sufficient condition (at least, with nomic force) for the supervenient property. If you think of the mental–neural relation in terms of the "realization" relation, the neural realizer of a mental state will again constitute a nomically sufficient condition for it. If we assume the standard nomological conception of causality, N threatens to preempt my desire as a cause of the foot movement. For my desire, as a cause of the movement, is nomologically sufficient for it, and this implies that N, too, is nomologically sufficient for it. So why isn't N a candidate, a better candidate, as the cause of the foot movement? (It will not do to posit a causal chain with the desire as an intermediate link between N and the foot movement; for one thing, the desire and N are perfectly simultaneous.)

Burge does not give a full characterization of the dependency relation, or the "underlying" relation, that he supposes to hold between the mental and physical. However, he is willing to go as far as this: "I want to leave open what 'underlie' is to mean here. I will assume, however, at least that mental states and events would not occur if some 'underlying' physical states and events did not occur".[16] So, on Burge's notion of "underlying" or "dependence", my desire for water would not have occurred if its neural underlier, N, had not occurred. We also have a counterfactual corresponding to the initial

---

[15]Burge, p. 116.

[16]Burge, p. 98. The quotation occurs in a paragraph in which Burge is expounding the picture he wants to criticize. However, it seems fair to take him to accept the view that physical states and processes "underlie" (in this sense) all mental events and states (especially, given the earlier quotation from his p. 116).

causal claim: if my desire had not occurred, my foot would not have moved. Given these two counterfactuals, a further counterfactual would seem to hold: If N had not occurred, my foot would not have moved. (I am not assuming here that counterfactuals are in general transitive; I am only saying that in the situation as envisaged, this counterfactual is plausibly true.) On the counterfactual conception of causation,[17] therefore, it would appear that N is also entitled to be a cause of my foot movement. So again we seem to have a case of overabundance of causes. Here, too, it doesn't make much sense to think of a causal chain from N to the foot movement *via* my desire.

Matters are no simpler when we consider mental–to–mental causation. Suppose a sharp pain in my chest causes an anxiety attack ("I might be having a heart attack"). Each of these events has a neural base on which it depends or supervenes, say N_1 and N_2, respectively. How might we understand the situation? Consider the effect, the sudden anxiety: how did it come about? There apparently are two answers: (1) it came about because it was caused by the chest pain, and (2) it came about because its subvenient base (or neural realizer), N_2, occurred. I believe that a tension, or instability, will be created if we accept these two answers without further clarification of their relationship. In view of (2), the causal claim (1) is difficult to understand: the anxiety attack would have occurred as long as its neural base, N_2, was there; it shouldn't matter what conditions had preceded it, unless they had something to do with the presence of N_2. It makes sense to suppose that if an event or state supervenes on, or is realized by, some underlying base, the only way to bring it about is to bring about its underlying base (or if it has multiple possible underlying bases, to cause one of these to be actualized). This means that for the chest pain to cause the anxiety attack, it must cause the latter's neural base to occur. This is a case of mental–to–physical causation, and the earlier considerations on such cases apply: the original causal relation between two mental events appears on the verge of collapsing into a causal relation between their respective neural bases.[18]

I don't see how these apparent consequences of the familiar assumption of mind–body supervenience or dependence, or the doctrine that all mental states must have neural/physical "realizers", can be ignored. Some have looked to mind–body supervenience as

---

[17] See Lewis, "Causation".

[18] Some of these issues are discussed in greater detail in my "The Nonreductivist's Troubles with Mental Causation", in *Mental Causation*, ed. Heil and Mele; reprinted in *Supervenience and Mind*.

the savior of mental causation (as I once did[19]); for example, Jerry Fodor says "if mind/body supervenience goes, the intelligibility of mental causation goes with it".[20] It is not an implausible idea that if we are to understand mental causation we must somehow integrate the mental into the physical system, and supervenience may, at first blush, look like just the relation we need to effect such an integration. The irony is that, instead of making mental causation intelligible, mind–body supervenience itself generates more puzzles —unless, that is, it is strengthened into some form of reductive identity.

4

And then there is of course the doctrine of the causal closure of the physical domain. There are various inequivalent ways of explaining this idea, but the following should capture the core idea. Let us say that a domain, D, of events (states, etc.) is causally closed just in case if x is in D and x causes y, or y causes x, then y is also in D. This means that no causal chain crosses the boundaries of D; or, to put it another way, if you pick any event in D and trace its causal ancestry or posterity, you will never get outside D. No assumption is made as to whether events in D have causes or effects. On the proposed definition, any set of causeless and effectless events, if such could be coherently conceived, will be causally closed. More realistically, a domain may be causally closed even if not every event in the domain has a cause or causal explanation; the constraint is only that if an event has a cause, or causal explanation, its cause must be found in that domain.[21]

Talk of cause and causal explanation inevitably leads to consideration of theories, theories over the given domain. Reference to theories is needed also when the issue of closure is put in terms of prediction as it sometimes is: in order to predict an event in the domain, does information concerning other events in the domain suffice, or is it necessary to go outside it? Also, to speak of the "physical" domain,

---

[19]In "Supervenience and Nomological Incommensurables", *American Philosophical Quarterly*, 15 (1978): 149-56; "Epiphenomenal and Supervenient Causation", *Midwest Studies in Philosophy*, 9 (1984): 257-70 (reprinted in *Supervenience and Mind*).

[20]*Psychosemantics* (Cambridge: MIT Press, 1987), p. 42.

[21]See Brian McLaughlin's entry on "anomalous monism" forthcoming in the *Routledge Encyclopedia of Philosophy*. Here and elsewhere in this section I am indebted to McLaughlin.

one needs at least a rough characterization of the physical, and again one looks to theories —that is, physical theories— for guidance. We can say that a theory, T, is "complete" or "comprehensive" (or, to use Quine's apt expression, gives "full coverage") over a domain, D, in case T provides for each event in the domain a description under which it can be explained and predicted, to the extent that it can be, on the basis of the laws and principle of T and without invoking events outside D. Let us say that T is a *causal theory* in case explanations of individual events it generates are causal explanations —that is, explanations that explain by citing causes. This presumably requires that at least some of T's laws are, in some appropriate sense, causal laws. Let us assume that c is a cause of e if and only if there in principle exists a law–based causal explanation of e which cites c as its cause. It appears to follow then that a domain is causally closed just in case there in principle is a complete causal theory over the domain (this claim will be modified below). These are rough and ready characterizations, and will need refinements. But let us move on.

Supposing then that the physical domain is causally closed, what follows about the possibility of mental causation? In the course of criticizing an argument purporting to lead to an epiphenomenalist conclusion, Burge writes:

> The sense in which only physical properties determine the causal powers of a physical event is just that within the patterns of causation described in the physical sciences and common–sense physicalistic discourse, physical properties suffice to provide a basis for the existence and understanding of the causal powers of physical events; and no other properties enter in. Both the chains of causation and the patterns of explanation are in no need of supplementation from outside the realm of physical properties or physical discourse.[22]

This, I think, is close enough to our sense of the causal closure of the physical domain. Burge goes on to say:

> If physical events have mental properties, one is not entitled to the view that only physical properties (properties specified in the physical sciences or in ordinary physicalistic discourse) determine all the causal powers of a physical event (as opposed to merely all the causal powers associated with physicalistic explanations of the physical event), unless one can show that mentalistic explanation is either non–causal or fails to describe patterns of causal properties. For the causal powers of a physical

[22] Burge, p. 100.

event that is mental might include possible effects that are specified in mentalistic explanation.[23]

Let c be a physical event, in the sense that it is an event with some physical property (we are here assuming a conception of events usually associated with Davidson,[24] as concrete structureless particulars which have, or instantiate, various properties), and let P be c's total physical property. Does P represent c's total causal powers, its total causal potential? Burge says no —not if c also has some mental property, M; for, as Burge says, event c, with mental property M, might have "possible effects that are specified in mentalistic explanation".

Let us assume, with Burge, that mental properties are outside the closed physical system (we will recur to this point below); we may also assume that *mentalistic* explanation he has in mind cites mental properties as causal factors (for why else call the explanation "mentalistic")?. In the present case, then, mentalistic explanations involving M must cite M as a causal factor; that is, the explanatory claim must be that c, *in virtue of having* M, brought about the effects to be explained. Burge doesn't say whether the "possible effects" he has in mind are mental or physical; we can consider various cases. So suppose that c, in virtue of M, causes an effect, e, which is "purely" physical in the sense that it has no nonphysical properties. Does this violate the causal closure of the physical domain? This points to a need to refine the notion earlier explained of causal closure, by relativizing both causation and explanation to properties of events —that is, properties of events in virtue of which they enter into causal/explanatory relations. For evidently the present possibility does not violate physical causal closure as defined, but most will agree, I believe, that it should. For cases of this kind, if real, would show that a complete physical theory of the physical domain is not possible, since an explanation of why e occurred must advert to a property, M, not within the purview of physical theory.[25] In this sense, physical theory would not even be able to give full coverage of the physical domain; there could, in principle, be no complete physical theory of the physical domain. This at bottom is the Cartesian picture of the world: for full coverage of the physical domain your

---

[23]Burge, pp. 100-101.

[24]See "The Logical Form of Action Sentences", "Causal Relations", and "The Individuation of Events", reprinted in Davidson's *Essays on Actions and Events.*

[25]But note that this does not in itself show that a "complete theory" of the physical domain is not possible; it only shows that any such theory must include nonphysical elements.

theory must refer to nonphysical causal agents. I believe that few of us would find this picture at all comfortable.

Let us consider the remaining two possibilities: (i) c, in virtue of its mental property M, causes an effect, e, with both mental property M^* and physical property P^*; (ii) the effect e is a purely mental event, with no physical property whatever. It is clear that case (i) leads to consequences similar to the case already considered: physical theory could not be a complete theory of the physical domain, since to explain, or predict, why an event with physical property P^* occurs, we must advert to a mental property. Case (ii) apparently leads to substance dualism; for *in* what, or *to* what, could such pure mental events occur? It difficult to think that such events could occur to physical systems.

What these considerations show is that if the physical domain is causally closed, with mental properties excluded from it, there is little room to accommodate mental–to–physical causation, a species of mental causation that must be countenanced if rationalizations are viable causal explanations of physical behavior. The only space left for the mental to play an explanatory/causal role appears to be the following: suppose that c causes e, where c and e are events with both physical and mental properties; it might be that c's mental property M is fully causally responsible for e's mental property M^*, while c's physical property P is causally responsible for e's physical property P^*.[26] This appears consistent with the causal closure of the physical domain with physical theory giving full coverage of it. However, this is not likely to be welcomed as giving full scope to the causal efficacy of mental properties in behavior production as it will not suffice to support a causal interpretation of rationalizing explanations. Moreover, if one believes in a supervenience thesis, to the effect that an event's mental properties supervene on its physical properties, the causal powers of mental property M are threatened, as we have seen, with preemption by physical property P.

5

In his discussion of the closure of the physical domain, Burge raises an interesting point worth discussing. He says:

> The claim that physical events can be caused only by virtue of physical properties of other physical events has problems entirely analogous to

---

[26]This reminds one of the strategy of some content externalists who argue for the causal/explanatory relevance of wide–content states by "widening" their explananda, i.e., by viewing them as actions under intentional descriptions.

those [covered in the Burge quotation in the preceding section]. The existence of a closed system reflects a pattern of causal relations and of causal explanation that needs no supplementation from the outside. There are no gaps. It does not follow from this that such a system excludes or overrides causal relations or causal explanation in terms of properties from outside the system. Indeed, if it did follow, as often been pointed out, there would be no room for causal efficacy in the special sciences, even in natural sciences like chemistry and physiology. For there is no gap (other than perhaps quantum gaps) in the causal relations explained in terms of the properties of physics. But few are tempted by the idea that physical events cannot be caused in virtue of physiological properties of physical events.[27]

Part of the concern raised here has already been addressed. The point on which I want to focus here is Burge's claim that if the causal closure of the physical domain were to exclude mental–to–physical causation (that is, the causal efficacy of mental properties in relation to physical properties), the same considerations would show that all special–science properties, e.g., chemical, biological, and physiological properties, are causally inefficacious with respect to their lower–level properties. To put another way: just as the causal closure of the fundamental physical domain does not exclude the causal efficacy of properties in the physical special sciences, like chemistry and biology, the causal closure of the physical domain (taken as a whole, to include biology, chemistry, etc.) does not exclude the causal efficacy of mental properties in favor of properties in the physical domain. This is a point often made, as Burge himself observes. For example, Baker writes:

> Moreover, I want to show that the metaphysical assumptions with which we began inevitably lead to scepticism not only about the efficacy of contentful thought, but about macro–causation generally. But if we lack warrant for claiming that macro–properties are generally causally relevant, and if we take explanations to mention causes, then most, if not all, of the putative explanations that are routinely offered and accepted in science and everyday life are not explanatory at all.[28]

Burge and Baker are not alone. This strategy is fairly common: try to defuse worries about mental causation by pointing out that mental properties are in the same boat as all other special science properties. Robert Van Gulick puts the point forcefully:

---

[27] Burge, p. 102.
[28] Baker, p. 77.

> . . . reserving causal status for strictly physical properties. . . would make not only intentional properties epiphenomenal, it would also make the properties of chemistry, biology, neurophysiology and every theory outside microphysics epiphenomenal. . . If the only sense in which intentional properties are epiphenomenal is a sense in which chemical and geological properties are also epiphenomenal, need we have any real concern about their status: they seem to be in the best of company and no one seems worried about the causal status of chemical properties.[29]

Perhaps no one is "worried" about the causal efficacy of chemical properties or biological properties, but then not many people are really "worried" about mental causation either. What some of us are worried about is finding an intelligible *account* of mental causation. This is a different worry, and, I dare say, a philosophically legitimate one. Do we have an account of causal efficacy for chemical or biological properties in relation to fundamental physical properties? Perhaps everyone believes that we can find one without too much trouble. Speaking for myself, I don't have one and I am not sure if anyone else does.[30] (If you are inclined to retort, "Who *needs* an account of why chemical properties have causal efficacy?", I have to confess I have nothing to say.) But can we be sure that an account that works, say, for chemical properties will work just as well in the mental–physical case?

What forms the background of the issues being raised here is the standard hierarchical picture of things of this world, and of their properties, vertically arranged micro to macro, from the elementary particles of microphysics to atoms and molecules, and their aggregates, and then upward to cells and organisms, and so on —a picture that has led to the familiar talk of "levels"— "levels of organization", "levels of description", "levels of analysis", "levels of explanation", and the like. I believe that the emergentists early in this century were the first to articulate such a model, but it has by now become an entrenched commonplace idea that continues to shape our thinking about physics and the special sciences, and about entities and properties posited in basic physics in their

---

[29] "Three Bad Arguments for Intentional Property Epiphenomenalism", *Erkenntnis*, p. 325. Hilary Kornblith made the same point these authors are making in his unpublished commentary on my paper "Mental Causation and Two Conceptions of Mental Properties", presented at the Atlanta meeting of the American Philosophical Association in 1993.

[30] Some years ago I offered a general account of macrocausation based on supervenience, in "Epiphenomenal and Supervenient Causation". As I have argued above, however, supervenience *sans* reduction doesn't seem adequate for the job.

relation to the structures and systems studied in the "higher–level" sciences.[31]

Our thinking about mentality and psychology, too, is strongly influenced by this model: psychology is a special science located at one of these levels, toward the higher end, in this multi–layered system, and mentality is a distinctive set of properties that make their first appearance at this level. William Lycan is very explicit about this:

> Very generally put, my objection is that "software"/"hardware" talk encourages the idea of a bipartite Nature, divided into two levels, roughly the physicochemical and the (supervenient) "functional" or higher–organizational —as against reality, which is multiple *hierarchy* of levels of nature, each level marked by nexus of nomic generalizations and supervenient on all those levels below it on the continuum. See Nature as hierarchically organized in this way, and the "function"/"structure" distinction goes *relative*: something is a role as opposed to occupant, a functional state as opposed to a realizer, or vice versa, only *modulo* a designated level of nature.[32]

For those who share this picture, there might seem nothing special about mental causation: if there is a difficulty in the mental/neural case, the same difficulty should arise at every level in relation to its underlying levels. Since there appears to be no particular problem at these lower levels, there can't be a problem at the mental level either. Let us call this "the generalization thesis".

Before we begin our examination of this thesis, let me first point out one reason for thinking that not all problems of mental causation can be thought to have analogues at other "levels". Some see difficulties with the causal status of the mental in the widely accepted claim that contentful intentional states are essentially relational and extrinsic, whereas we expect proximate causes of behavior to be intrinsic and local.[33] Now, it might be argued that this, on a deeper look, turns out not be a genuine difficulty, and that content externalism is entirely compatible with a full causal status of intentional

---

[31] A useful statement of the hierarchical model is found in Paul Oppenheim and Hilary Putnam, "Unity of Science as a Working Hypothesis", *Minnesota Studies in the Philosophy of Science*, vol. 2 (Minneapolis: University of Minnesota Press, 1958).

[32] William G. Lycan, *Consciousness* (Cambridge: MIT Press, 1987), p. 38.

[33] See, e.g., Pierre Jacob, "Externalism and Mental Causation", *Proceedings of the Aristotelian Society*, 92, Part 2 (1992): 203-219; Christopher Peacocke, "Externalist Explanation", *Proceedings of the Aristotelian Society*, 93 (1993): 203-30.

states. That may be, but I am here only making the simple point that if the relational character of content should pose a prima facie problem for which we need a response of some kind, we are not likely to get any guidance from looking at, say, the chemical/physical case. In any event, I will set aside for another occasion the implications of content externalism for mental causation.

In an earlier section we discussed how a problem about mental causation can arise from the doctrine of mind–body dependence or supervenience. Discussion there did not require an elaborate characterization of the dependence or supervenience relation involved. Reasoning that leads to the generalization thesis is based on the assumption that the mental/neural relationship is, *in all relevant respects*, the same kind of relationship that characterizes, say, the chemical/microphysical, or biological/physicochemical, case. My earlier argument concerning supervenience, if correct, shows that where there is supervenience there is a problem about the causal efficacy of the supervenient properties in relation to their base properties. However, the apparent generality of this argument may fail, since a given case of supervenience may involve special further factors which help resolve, or otherwise neutralize, the problem, and these factors may not be present in other cases of supervenience. In any case, the idea that there is some single hierarchical structure of properties, generated by the same ordering relation up and down, needs closer examination. I think it is time to take a fresh look at this venerable, and platitudinous, idea of hierarchy of "levels".

Let us begin by focusing on the idea that mental properties are "realized" by physical/neural properties, or what is usually taken to be an equivalent idea that mental properties are the "roles" of which the physical/neural properties are the "occupants"). This idea is central to functionalism —in fact, to much of current thinking on the mind–body problem. It is often assumed that this realization relation is what generates the hierarchical ordering of levels. (This assumption is evident in the Lycan quotation above.) On functionalism, a mental property is construed as a "second–order" property consisting in having some first–order property meeting a certain *causal specification*, H.[34] Any first–order property satisfying H is said to be a "realizer" of the mental property (or an "occupant" of the "role" defined by H). To generalize the idea, let us assume we have a stock

---

[34] I believe Hilary Putnam was first to use the locution of "second–order property" in connection with "functional states" in a sense very close to the present sense; see his "On Properties", in *Essays in Honor of Carl G. Hempel*, ed. Nicholas Rescher et al. (Dordrecht: Reidel, 1969), p. 244.

of first–order properties, $F_1, F_2, \ldots$ (let us refer to them collectively as $\boldsymbol{F}$). These need not be taken to be "first–order" in any absolute sense; in the case of mental, these F's can be taken to include neural/physiological properties, behavioral properties, and the like. As I take it, the general idea of a "second–order property" is this: G is a second–order property defined over $\boldsymbol{F}$ just in case it is the property of having (instantiating, exemplifying) some property in $\boldsymbol{F}$ which satisfies a certain specification H.[35] Thus, if $\boldsymbol{F}$ is a class of colors, "H" might be "being a primary color", in which case the second–order property defined by H is the property of having some primary color or other. Thus, second–order properties are generated by existential quantification over the given domain of first–order properties.

Cases of H that are of primary interest to the functionalist conception of mentality are those in which H gives a *causal/nomological* specification, of the form "having some property in $\boldsymbol{F}$ (typically) caused by F_i and (typically) causing F_j", where F_i and F_j are specific properties in $\boldsymbol{F}$. We may call second–order properties generated by causal specifications "functional properties" or "causal roles". Thus, pain, on the functionalist account, is a second–order functional property consisting in having some (first–order) property typically caused by tissue damage and typically causing wincings, cryings, and other "pain behaviors".[36] Any property in $\boldsymbol{F}$ that meets this causal requirement is a "realizer" of pain; the so–called doctrine of "multiple realizability" of the mental, when applied to pain, is the claim that there are many diverse and heterogeneous properties in $\boldsymbol{F}$ that meet pain's causal specification. We expect to find different neural realizers of pain in species with different neural anatomies, and we expect species to have developed different evolutionary solutions to the problem of acquiring a reliable tissue–damage detector. It is often pointed out that the pain realizer may differ from person to person, and even for the same person at different stages of neural maturation and development. Just to have another example on hand:[37] consider the property of dormitivity, having the power to put people to sleep. Dormitivity, then, is the property of having a

---

[35] We can also allow second–order properties of the form "having *every* property in $\boldsymbol{F}$ which meets specification H"; however, existential quantification over first–order properties suffices for our present purposes.

[36] A functionalist might add "and a sense of distress and desire to be rid of it". To keep matters reasonably simple I will here ignore the causation of further mental states.

[37] This example comes from Block, "Can the Mind Change the World?", in *Meaning and Method: Essays in Honor of Hilary Putnam*, ed. George Boolos (Cambridge: Cambridge University Press, 1990).

certain property that causes people to fall asleep; it is therefore a functional property which has diverse first–order chemical realizers, secobarbital in Seconal, diazepam in Valium, and so on.

One problem of mental causation raised by Block concerns the causal status of such functionally defined second–order properties —whether dormitivity, as a second–order property distinct from its first–order chemical realizers, can be accorded causal powers to put people to sleep.[38] It is plausible that admitting dormitivity as a cause results in an undesirable overabundance of causes: if you take a Seconal pill and falls asleep, there seem to be two distinct causes of it, the pill's dormitivity and its chemical ingredient, secobarbital. Given that the dose of secobarbital puts you to sleep, how can there be any *further* causal work dormitivity can contribute?[39] Block's own conclusion is pessimistic: we cannot escape the epiphenomenalist thesis that any second–functional property defined as having some property causing K is epiphenomenal with respect to K. This would mean that we could no longer invoke pain to explain wincings and cryings and other pain behaviors. I have argued elsewhere[40] that Block's epiphenomenalism, given his assumptions, must be generalized: the causal powers of any second–order property (functional or otherwise) are in danger of preemption by its first–order realizers. But these details will not matter to us here; our question is: Does Block's problem generalize?

But does Block's problem generalize? It's clear from the example of dormitivity, the same problem can arise in different areas — different levels, if you wish. Think of importing Block's problem into the hierarchy of the sort pictured by Lycan: neural realizers of pain are themselves second–order with respect to certain lower–level properties (which preempt their causal powers), the latter in turn are second–order to their lower–level properties, etc., ad infinitum, until you reach the bottom level of microphysics (if such a level exists), which turns out to be the only level where genuine causal powers reside. This strikes us as intolerable. What's more: what if there is no bottom level (as Block challenges us to consider[41])? It looks as though if there is no bottom level in this picture, there wouldn't be any causation *anywhere*!

---

[38] Block, "Can the Mind Change the World?".

[39] I should say that this isn't exactly the way Block formulates the epiphenomenalist threat. I discuss Block's problem and his conclusions in "'Second–Order' Properties and Mental Causation" (unpublished).

[40] In "'Second–Order' Properties and Mental Causation".

[41] In personal communication.

But the realization relation does not generalize downward, or upward, in the micro–macro hierarchy. The first point to notice is that *both second–order properties and these first–order realizers are properties of the same entities.* The pill you ingest has both dormitivity and the chemical property which realizes dormitivity; and a human is in pain and has his/her c–fibers activated (to use the familiar example). It is evident that *a second–order property and its realizers are at the same level; they are properties of the very same objects and systems.* Consequently, when we talk of realizers, there is no movement downward, or upward, the hierarchy of entities ordered by the micro–macro relation. An illusion of downward movement is created, I think, by the fact that often the realizers are *micro–based (or microstructural) properties*, which are *macro–properties* that are generated or constituted (in some appropriate sense that needs to be explained[42]) by the properties and relations characterizing micro–constituents. But it is clear that this is not always the case: having a primary color, as we saw, is a second–order property over the domain of colors, but its realizers are just colors. Whether or not colors are micro–based properties obviously is not an issue here; even if there is a sense in which colors of surfaces depend on their micro–characteristics, that may be so because all, or most, properties depend on microstructure. The question whether something is a realizer of a certain second–order property is independent of issues concerning any micro–macro hierarchy. The situation is the same with functional properties defined in terms of causal roles. Consider the functional second–order property: having a property caused by tissue damage and causing wincings and cryings. There could be a Cartesian functionalist who takes this property to have phenomenal pain as one of its realizers, and it doesn't take much imagination to think of multiple phenomenal/mental realizers of functionally defined psychological properties.

As this last case illustrates, there is no reason to suppose that realizers must themselves have realizers. To think they must involves the supposition that every property is a second–order property over some domain of properties, and I don't think anyone has given a good argument to suppose this to be the case. This applies to what I have called "micro–based properties" as well. The mass of my coffee table is micro–based in the sense that it is the sum of the mass of its top and the mass of its base. It makes no sense —at least, I don't see how it could be helpful for any purpose— to think of having a mass of 50 kilograms as a second–order property.

---

[42]But I cannot attempt it here.

There are two possible errors we must resist here: first, some appear to think that a property is "multiply realized" just because two or more properties fall under them, or its instances can be partitioned into two or more classes. But surely there is no reason to think that, say, the color red has multiple realizers because some red things are round and some are square ("both square objects and round objects realize redness") —being red and square and being red and round are not realizers of red; or that mass of one kilogram has multiple realizers because there are one–kilo gold lumps, one–kilo copper lumps, and so on.[43] This would trivialize the very notion of realization and render it useless. Second, certain properties enter into nomic relations with other properties, and it may be possible to characterize these properties in terms of their nomic/causal relations with other properties.[44] But from this it does not follow that they are second–order functional properties, causal roles, or "job descriptions". Having a mass of one kilogram has causal consequences (that is, it is a "causal power"), such as accelerating at a certain rate when a force of a certain magnitude is impressed upon it, etc. This gives mass of one kilogram a certain causal role. But this doesn't make mass a functional property;[45] for suppose we define a second–order functional property in terms of this causal role; what would be its realizer? Obviously, having a mass of one kilogram!

So the worry about a bottomless regress of levels of realization seems unfounded. For the realization relation has nothing intrinsically to do with the micro–macro relation; it doesn't track the micro–macro hierarchy of levels. As we saw, the realization relation obtains between properties *at the same level* in the standard micro–macro hierarchy of entities and their properties. At each level there are properties that might be thought to be first–order properties (for certain purposes), and these properties generate second–order

---

[43]Cf. Donald Davidson: "It is often said, especially in recent philosophical literature, that there cannot be a physical predicate with the extension of a verb of action... because there are so many different ways in which an action may be performed. Thus a man may greet a woman by bowing, by saying any number of things, by winking, by whistling; and each of these things may in turn be done in endless ways. The point is fatuous. The particulars that fall under a predicate always differ in endless ways, as long as there are at least two particulars. If the argument were a good one, we could show that acquiring a positive charge is not a physical event, since there are endless ways in which this may happen". In "The Material Mind", reprinted in *Essays on Actions and Events*, pp. 251-52.

[44]See David Lewis, "How to Define Theoretical Terms", *Journal of Philosophy*, 67 (1970): 427-46.

[45]I believe this is the error made by certain positivistic thinkers (e.g., Ernst Mach, operationalists).

properties through existential quantification ("the property of having some property or other meeting specification H"), which can be iterated to generate third– and higher–order properties. This means that even if the exclusion problem arises for a second–order property in relation to its realizers and its causal powers are preempted by the latter, this does not have the consequence of depositing causal powers at a lower level in the standard hierarchical model. Causal competition between a property and its realizers is an intra–level affair, not a cross–level rivalry. However the competition is resolved, causal powers stay at the same level; there is no reason to worry that somehow causal powers will be drained away downward, either to be deposited wholly at the deepest microphysical level or to be sucked away into a bottomless abyss.

6

Let us return to the generalization thesis —the claim that the problem of mental causation generalizes to other levels. It seems clear that if, say, biological properties are second–order functional properties (whatever their first–order realizers may be), problems similar to those for mental properties (as functionally construed) will arise. As has been emphasized, there is no downward motion in the second–order/realizer relation. If G is a second–order property, whether functional or otherwise, any instance of G is in fact an instance of the first–order realizer which is instantiated on that occasion, and the former therefore has all the causal powers of the latter (more on this below). There is no downward seepage of causal powers: it is only that instances of G, hence G as a kind, are causally heterogeneous —as causally heterogeneous and efficacious as its realizers are; this doesn't render them causally inefficacious.

The situation is the same when a realizer is a micro–based property. Consider the mass of my table again: this property is micro–based but it is a macroproperty that no proper part of the table instantiates, and its causal powers differ from the causal powers of the mass of any of its parts. The point of saying that the mass of the table is a micro–based property is merely that it is analyzable in terms of properties and relations characterizing its parts. This does not mean that it, or its causal power, is reductively identified with, or gives way to, properties or causal powers of its parts.

What we have said is true of the supervenience relation as well: as standardly conceived both supervenient and their subvenient base properties are properties of the same entities. Thus, the hierarchy

of properties generated by the supervenience relation does not correspond to the standard micro–macro ordering of levels.

While these observations help clarify the picture, they do not constitute a full response to the generalization thesis, the claim that the problem of mental causation, if it is a problem at all, generalizes to properties at all levels except those at the bottom level. What we have seen is that there is no reason to think that the generalization thesis follows from the model of hierarchically organized levels of entities and their properties. Also, our considerations show that the full generalization thesis is probably false; there is no reason to think that problems similar to the problem of mental causation should beset *all* properties merely in virtue of their being located at a level other than the most basic level. However, this clearly does not rule out the existence of analogues of mental causation in other areas. If mental properties are threatened with the loss of their causal powers to their realizers in virtue of being second–order in relation to physiological–behavioral properties, the same threat confronts all second–order properties everywhere. Further, if the causal closure of the physical creates a problem for mental properties because they are outside the physical system, the same problem has to be faced by other properties outside the system. And so on. What should we say about this?

The first thing we should say, I think, is that if there are such cases outside the mental realm we need to account for them as well; we need an account of how causal relations and causal explanations involving such properties are possible. What we shouldn't take for granted is that all these different cases call for a single uniform solution, that since the problem is so vexing in the case of mental causation and resists solution, the situation will be the same across the board. In other cases, solutions may be easier to come by. In particular, reductive solutions of various forms are likely to be more viable in nonpsychological cases, because at least they do not entangle us in the intractable issues of intentionality and subjectivity.

Let us briefly return to the issue of the causal closure of the physical domain. I believe it is crucially important not to construe the physical domain too narrowly. The standard micro–macro hierarchical model encourages the idea that the causally closed physical domain includes only the basic particles and their interactions, but this is another groundless assumption associated with the hierarchical picture. Obviously the physical domain must also include aggregates of basic particles, aggregates of aggregates, and so on without end; molecules, cells, tables, planets, and biological organisms must all belong in the physical system. What then of properties? What properties, in addition to the basic properties and relations of the

microparticles, are to be allowed into the physical system? Clearly, mass of one kilogram should be allowed in, although no microparticle has this property. Similarly, all micro–based properties, such as being composed of two hydrogen atoms and one oxygen atom in a certain chemical bonding, must be considered part of the physical domain. Otherwise, the system won't be causally closed; mass of one kilogram has causal powers no smaller masses have.

What of second–order properties? It seems to me that properties of the physical domain, or any domain, should be considered closed under the operation of generating second–order, and higher–order, properties. The operation involved is merely existential quantification over properties already in the system, and a logical operation of this sort should not take us outside the system. Actually, the operation involved goes beyond quantification; to generate functional properties in our sense we need the causal relation as well. But I don't see that this should make a difference: any second–order specification picks out first–order properties disjunctively (as logicians tell us, existential quantification can be parsed in terms of disjunction).

The physical domain on this conception is probably wider and more comprehensive than what most philosophers have taken it to be. It will include chemical properties; I believe they are either micro–based properties (based on fundamental physical properties of particles, atoms, etc.) or second–order properties generated from these micro–based properties. This means that dispositional properties, like transparency and ductility, will find their way into the physical system. The physical domain includes biological organisms, since they are aggregates of basic particles; but what of their biological properties? Many people believe they are functional properties; as Burge and others have noted, being a heart is having some property that causes blood to be pumped (when embedded in a certain causal context). This seems pretty plausible. I think that biological properties are like chemical properties: they are either micro–based properties (perhaps relative to nonbiological physical and chemical properties) or second–order properties generated from them. If that is the case, Burge and others needn't worry about biology or biological causation: biological properties, along with chemical properties, will be found *inside* the physical domain! And if mental properties are what the functionalist says they are, they, too, will find a place in the physical system. That, I think, is one of the virtues of functionalism, at least from the point of view of physicalism.[46]

---

[46]This is another way, perfectly appropriate and legitimate in my view, of understanding "physical reduction", i.e., to be incorporated into the physical

Finally, let us return to the exclusion problem besetting second–order properties in relation to their first–order realizers. If this is a genuine problem, as I think it is, getting chemical properties and biological properties into the physical domain isn't enough; physical or not, they are still threatened with epiphenomenalism by their realizers. So consider, again, dormitivity and its chemical realizers, $C_1, C_2, \ldots$ The strategy I suggest in order to resolve this problem is to *identify* the causal powers of dormitivity *disjunctively* with those of $C_1, C_2, \ldots$ Any instance of dormitivity is either a C_1–instance, or C_2–instance, or..., and the causal powers of given instance of dormitivity will be identical with the causal powers of the C_k–instance where C_k is the first–order realizer of dormitivity on this occasion. On this view, second–order properties turn out to be causally heterogeneous, as heterogeneous as their first–order realizers are, and this raises doubts about their nomological–projectible character. But these doubts have to be accepted: if we want to insist on the existence of diverse and heterogeneous realizers for certain second–order properties, as is often urged for mental properties, we must accept the causal–nomological heterogeneity of these properties.[47]

I believe that this solution to the exclusion problem is adequate, and also inherently plausible, for second–order properties. The only question is whether mental properties are properly regarded as second–order properties —more specifically, whether functionalism is a correct view concerning mental properties. While the appeal of the functionalist conception of the mental is undeniable, there is also a strong pull toward the opposite view that mental properties, especially phenomenal properties, are first–order properties in their own right (whatever this really means), with their distinctive intrinsic characters, not causal/relational properties defined by job descriptions. This is what has spawned the so–called problem of qualia and has made the problem of consciousness look so intractable. But on this intrinsic view of mental properties, the exclusion problem seems to me unsolvable: there is no way to accommodate them within the causal structure of the world that is essentially grounded in physical properties.[48]

---

system in the way described here. Many current antireductionist stances are taken on the basis of an unrealistically stringent notion of what a reduction ought to accomplish.

[47]For more details on these points see my "Multiple Realization and the Metaphysics of Reduction" and "Postscript on Mental Causation", both in *Supervenience and Mind*; and "'Second–Order' Properties and Mental Causation".

[48]I discuss this in some detail in my "'Second–Order' Properties and Mental Causation".

There is an alternative to the iteration of the realization relation to generate a hierarchy of properties, ordered micro–to–macro. Suppose that pain is realized by a certain neural micro–based property N. We do not ask "What realizes this first–level neural realizer?" (as Lycan perhaps would). Instead we ask: "What realizes pain at the molecular level?" That is, there may well be (presumably are) micro–based properties at the physicochemical level that realizes pain. And there may be pain realizers at a more fundamental level than these, or levels intermediate between them. Although they are all macroproperties in the sense that they are properties of the subject to whom pain applies, each is micro–based on a different micro–level. How are these realizers of pain related to one another? How are we to understand mental causation in this picture? These are interesting questions, but I must leave them for another occasion.[49]

In conclusion, then, on the functionalist view, the problem of mental causation has analogues in other second–order properties, but the problem is solvable. If mental properties are taken as intrinsic, first–order properties in their own right, the problem of mental causation may well be a special problem about mentality.[50]

---

[49]Discussion of this issue will require a more precise characterization of "micro–based properties".

[50]I am much indebted to Ned Block and Brian McLaughlin for discussing with me many of the issues taken up in this paper. I have received helpful comments from participants at the 1994 SOFIA conference in Lisbon —in particular, from my three commentators, Louise Antony, Manuel Campos, and Jim Tomberlin— and from colleagues at Brown, in particular Justin Broackes and Ernest Sosa.

PHILOSOPHICAL ISSUES, 6
Content, 1995

Mental Causation: A Query for Kim

James E. Tomberlin

Of the vexing problems Jaegwon Kim has raised and explored surrounding mental causation, I will tackle just one.[1] The issue at stake is fascinating and deeply troublesome.

1 The Problem

Consider:

(1) If there is genuine mental causation, the causation in question is either nomic necessity or something weaker.

---

[1]Besides his paper in the present volume, Kim's important body of work on mental causation includes Kim (1984, 1989, and 1993a). These last three essays are reprinted in Kim (1993b).

There are excellent (but jointly conflicting) treatments of mental causation in Antony (1989, 1991), Audi (1993), Baker (1993), Block (1990), Braun (1995), Burge (1993), Davidson (1963, 1980, and 1993), Dretske (1989), Fodor (1989), Honderich (1982, 1991), Horgan (1989), Johnston (1984), LePore and Loewer (1987), McLaughlin (1985, 1989, and 1993), Pereboom (1995), Sosa (1984, 1993) and Van Gulick (1993).

(2) But nomic necessity won't do as an account of mental causation.

(3) Thus, either there is no genuine mental causation or else the causation at issue is weaker than nomic necessity.

Now the lead premise of this valid argument seems beyond reproach. And I think Kim has made an eminently plausible case for the second premise. If so, we appear to face exactly two options: deny genuine mental causation altogether, or find the causation involved weaker than nomic necessity. In what follows, I pursue the latter alternative.

2 Causation as Nomic Necessity: A Review

Let '$O(e)$' abbreviate 'the proposition that event e occurs' and let '$O(e) \rightarrow O(c)$' stand for '$O(e)$ causes it to be the case that $O(c)$'. Then, by the nomic necessity view, we have:

(4) $O(e) \rightarrow O(c)$ is true if and only if $O(c)$ is true in every physically possible world in which $O(e)$ is true,

where it is understood by the set of physically possible worlds that proper subset of the set of all logically possible worlds whose members are exactly those worlds in which the laws of nature hold. Or equivalently, where '$\boxed{P}$' abbreviates 'It is physically necessary that':

(5) $O(e) \rightarrow O(c)$ is true if and only if $\boxed{P}(O(e) \supset O(c))$.

This account is not without plausibility. It properly observes that the relation or connection between the antecedent and consequent of a causal conditional is modally *stronger* than that of the material conditional. At the same time, it seeks to accommodate the fact that causal conditionals are modally *weaker* than the conditional of logical necessity. What is more, this view incorporates the intuition (shared by many) to the effect that there is an intimate connection between statements of causality and propositions formulating nomological laws.

3 Against the Necessity of Nomic Necessity

Although plausible, I submit, the nomic necessity account is too strong: $O(e) \rightarrow O(c)$ may be true even though $O(c)$ is not true in every physically possible world in which $O(e)$ is true.

Take, for example, Sam's death last week. Now Sam suffered a total kidney failure, but he otherwise was in moderate health. With this information and the assurance of Sam's doctors, I take it, Sam's death last week was caused by his total kidney failure. That is, where '$O(e_1)$' and '$O(c_1)$' stand for 'Sam's having a total kidney failure' and 'Sam's death last week' respectively, we have the truth of $O(e_1) \rightarrow O(c_1)$. And yet, it seems just false that $O(c_1)$ is true in every physically possible world in which $O(e_1)$ is true: no law of nature or statement of nomological necessity is violated under the assumption that Sam suffers a total kidney failure but he continues to live a long and productive life. Still, in his actual situation, Sam's death was caused by a total kidney failure.

(Observe: I am *not* contending that causation is never nomic necessity. Quite the contrary, I readily affirm

(6) If $O(c)$ is true in every physically possible world in which $O(e)$ is true, then $O(e) \rightarrow O(c)$.

Every instance of nomic necessity is a case of genuine causation. Instead, by the picture I mean to endorse, causation comes in (modal) *grades*, where nomic necessity is but one though presumably the modally strongest. But if (as I firmly believe) cases like the above involving Sam's death are examples of genuine causation, like it or not there must be one or more grades of causation modally weaker than nomic necessity. As it will become clear in the remaining sections, I opt for at least one such grade.)

4 A Second View of Causation

David Lewis has supplied an ingenious account of causation.[2] Here is a brief summary of his proposal.

Lewis confines himself to causation among events. What he offers is a counterfactual treatment of such causation. It is important, therefore, that we begin with his distinctive view of counterfactuals. Given any two propositions A and C, there is the counterfactual $A \mathbin{\Box\!\!\rightarrow} C$: the proposition that if A were true, then C would be true. The operation $\Box\!\!\rightarrow$ is defined by a rule of truth as follows:

(7) $A \mathbin{\Box\!\!\rightarrow} C$ is true (at a world w) if and only is either (a) there are no possible A–worlds (in which case $A \mathbin{\Box\!\!\rightarrow} C$ is vacuously true) or (b) some A–world where C holds is closer (to W) than is any A–world where C does not hold.

---

[2]See Lewis (1973), reprinted in his (1986).

According to (7), a counterfactual is nonvacuously true just in case it takes less of a departure from actuality to make the consequent true along with the antecedent than it does to make the antecedent true without the consequent.

Next, Lewis characterizes a pair of notions, counterfactual dependence and causal dependence, as follows. Let $A_1, A_2, \ldots$ be a family of possible propositions, no two of which are compossible; and let $C_1, C_2, \ldots$ be another such family of equal size. Then, Lewis has it, C's depend counterfactually on the A's if all the counterfactuals $A_1 \mathbin{\Box\!\!\rightarrow} C_1$, $A_2 \mathbin{\Box\!\!\rightarrow} C_2, \ldots$ are true. Turn now to events. Let $c_1, c_2, \ldots$ and $e_1, e_2, \ldots$ be families of distinct possible events such that no two of the c's and no two of the e's are compossible. Then the family $c_1, c_2, \ldots$ depends causally on the family $e_1, e_2, \ldots$ if and only if the family of propositions $O(c_1), O(c_2), \ldots$ depends counterfactually on the family $O(e_1), O(e_2) \ldots$.

From this a relation of causal dependence is defined:

(8) Where c and e are two distinct possible events, c depends causally on e if and only if $O(e) \mathbin{\Box\!\!\rightarrow} O(c)$ and $\sim O(e) \mathbin{\Box\!\!\rightarrow} \sim O(c)$.

And for causation itself among single events Lewis offers:

(9) If c and e are two actual events such that c depends causally on e, then e is a cause of c.

(Observe: Lewis goes on to extend (9) to a bi–conditional. But for my purposes here (9) will do.) Given all of the above, of course, we have:

(10) If c and e are two actual events such that $O(e) \mathbin{\Box\!\!\rightarrow} O(c)$ and $\sim O(e) \mathbin{\Box\!\!\rightarrow} \sim O(c)$, then e is a cause of c.

Lewis's counterfactual account of causation —hereafter *LCA*— has a number of salient features: *first*, given the Lewis rule of truth for counterfactuals $A \mathbin{\Box\!\!\rightarrow} C$ does *not* entail $\boxed{P}(A \rightarrow C)$: where '$A$' and '$C$' stand for 'The New York Rangers won the cup' and 'John will be unhappy', in turn, $A \mathbin{\Box\!\!\rightarrow} C$ is true even though $\boxed{P}(A \rightarrow C)$ is surely false; *second*, inasmuch as $\boxed{P}(A \rightarrow C)$ trivially entails $A \mathbin{\Box\!\!\rightarrow} C$, it follows that *LCA* is a modally weaker account of causation than the nomic necessity view; *third*, unlike the nomic necessity view, *LCA* nicely accommodates examples such as the one involving Sam's death from a total kidney failure as cases of genuine causation; and *fourth*, I take it there is no doubting the existence of mental causation under *LCA*.

These features make *LCA* ever so theoretically attractive. But alas:

5 Kim's Examples: Against the Sufficiency of LCA

Thanks to Kim,[3] it is perfectly clear that *LCA* as formulated won't do as an adequate account of causation among events. A single example will serve: Where '$O(e)$' and '$O(c)$' do duty for 'Today is Friday' and 'Tomorrow is Saturday', respectively, we have the satisfaction of (10)'s antecedent. Contra *LCA* and (10) in particular, however, there is no genuine causation here. And consequently, a mere truth of the pair $O(e) \mathbin{\Box\!\!\rightarrow} O(c)$ and $\sim O(e) \mathbin{\Box\!\!\rightarrow} \sim O(c)$ will not suffice for e's being a cause of c.

6 A Proposal

Lest advocates of counterfactual causation abandon ship, I offer a reminder followed by a comparison.

A Reminder. Before Gettier,[4] it should be recalled, nearly every epistemologist found comfort in something like

(11) S knows that P if and only if (a) P is true, (b) S believes P, and (c) S's belief that P is justified.

With Gettier came the onerous realization that justified true belief is simply not sufficient for propositional knowledge. Like it or not, (11) has to give way to

(12) S knows that P if and only if (a) P is true, (b) S believes P, (c) S's belief that P is justified, and (d) C.

As nearly everyone knows, finding this missing condition C has proved to be notoriously difficult. But surely the endeavor is a requisite one: The proper lesson of the Gettier Problem is *not* that there is no bona fide notion of propositional knowledge to be fleshed out; quite the contrary, we should know that there is such a notion given the vital lesson that (mere) justified true belief is insufficient for its exemplification.

A Comparison. Like the situation with Gettier, prior to Kim, advocates of counterfactual causation enjoyed *LCA* or something similar, confident that there is a legitimate form of causation modally weaker than nomic necessity. With Kim, alas, comes the firm verdict

---

[3] Kim (1973).

[4] See Gettier (1963).

that Lewis's pair of counterfactuals are not sufficient for causation. Fine. But friends of counterfactual causation ought nevertheless to remain steadfast in the observation that examples like the one of Sam's death from a total kidney failure are genuine cases of causation not covered by the nomic necessity view. Accordingly, just as (11) yields to (12), owing to Gettier, so too (10), after Kim, should be replaced in favor of

(13) If c and e are two actual events such that $O(e) \mathrel{\Box\!\!\rightarrow} O(c)$, $\sim O(e) \mathrel{\Box\!\!\rightarrow} \sim O(c)$, and K, then e is a cause of c.

And once again a demanding but needful task is to supply the condition K. Still, regardless of how K is to be finally unpacked, this theoretical constraint is in force: Sam's death as a result of his total kidney failure is to be accommodated as legitimate causation; and so K cannot imply the modally stronger grade of causation embodied in nomic necessity. With Sam's unfortunate death becoming a bona fide member of the causal order, however, there seems no denying that room will have been provided for mental causation, after all.[5, 6]

References

Antony, Louise 1989: Anomalous Monism and the Problem of Explanatory Force. *Philosophical Review*, 2, 153-87.

Antony, Louise 1991: The Causal Relevance of the Mental. *Mind and Language*, 6, 295-327.

Audi, Robert 1993: Mental Causation: Sustaining and Dynamic. In Heil and Mele (1993).

Baker, Lynne Rudder 1993: Metaphysics and Mental Causation. In Heil and Mele (1993).

Block, Ned 1990: Can the Mind Change the World? In Boolos (1990).

Boolos, George (ed.) 1990: *Meaning and Method: Essays in Honor of Hilary Putnam*, Cambridge, Cambridge University Press.

Braun, David 1995: Causally Relevant Properties. In Tomberlin (1995).

---

[5]In future work, along with exploring various ways of spelling out condition K, I plan to apply the modal grade view of causation to the thorny problem of causal determinism versus free will.

[6]A version of this essay was presented at the Seventh Sofia Conference on Content, held in Lisbon, Portugal, May 1994. I am grateful for a stimulating discussion with special thanks to Jaegwon Kim, Tyler Burge, Roger Gibson, Bill Lycan, Ernie Sosa, and Crispin Wright. For rewarding additional discussion, I am indebted to David Cowles, Greg Fitch, Terry Horgan, Ed Zalta, and my colleagues Frank McGuinness, Jeff Sicha, and Takashi Yagisawa. I do not mean to imply, of course, any agreement on their part with my views.

Burge, Tyler 1993: Mind–Body Causation and Explanatory Practice. In Heil and Mele 1993.

Davidson, Donald 1963: Actions, Reasons, and Causes. *Journal of Philosophy*, 64, 691-703. Reprinted in Davidson (1980).

Davidson, Donald 1980: *Essays on Actions and Events*, Oxford, Clarendon Press.

Davidson, Donald 1993: Thinking Causes. In Heil and Mele (1993).

Dretske, Fred 1989: Reasons and Causes. In Tomberlin (1989).

Fodor, Jerry 1989: Making Mind Matter More. *Philosophical Topics*, 17, 59-80.

Gettier, Edmund 1963: Is Justified True Belief Knowledge?. *Analysis*, 23, 121-23.

Heil, John and Mele, Alfred (eds.) 1993: *Mental Causation*, Oxford, Clarendon Press.

Honderich, T. 1982: The Argument for Anomalous Monism. *Analysis*, 42, 59-64.

Honderich, T. 1991: Better the Union Theory. *Analysis*, 51, 166-73.

Horgan, Terence 1989: Mental Quasation. In Tomberlin (1989).

Johnston, Mark 1985: Why Having a Mind Matters. In LePore and McLaughlin (1985).

Kim, Jaegwon 1973: Causes and Counterfactuals. *Journal of Philosophy*, 70, 570-72.

Kim, Jaegwon 1984: Epiphenomenal and Supervenient Causation. *Midwest Studies in Philosophy*, 9, 257-70.

Kim, Jaegwon 1989: Mechanism, Purpose, and Explanatory Exclusion. In Tomberlin (1989).

Kim, Jaegwon 1993: The Non–Reductivist's Troubles with Mental Causation. In Heil and Mele (1993).

Kim, Jaegwon 1993: *Supervenience and Mind*, Cambridge, Cambridge University Press.

LePore, Ernest and McLaughlin, Brian (eds.) 1985: *Actions and Events: Perspectives on the Philosophy of Donald Davidson*, Oxford, Basil Blackwell.

LePore, Ernest and Loewer, Barry 1987: Mind Matters. *Journal of Philosophy*, 84, 630-42.

LePore, Ernest and Loewer, Barry 1989: More on Making Mind Matter. *Philosophical Topics*, 17, 175-91.

Lewis, David 1973: Causation. *Journal of Philosophy*, 70, 556-67. Reprinted in Lewis (1986).

Lewis, David 1986: *Philosophical Papers*, II, Oxford, Oxford University Press.

McLaughlin, Brian 1985. Anomalous Monism and the Irreducibility of the Mental. In LePore and McLaughlin (1985).

McLaughlin, Brian 1989: Type Epiphenomenalism, Type Dualism, and the Causal Priority of the Physical. In Tomberlin (1989).

McLaughlin, Brian 1993: On Davidson's Response to the Charge of Epiphenomenalism. In Heil and Mele (1993).

Pereboom, Derk 1995: Conceptual Structure and the Individuation of Content. In Tomberlin (1995).

Sosa, Ernest 1984: Mind–Body Interaction and Supervenient Causation. *Midwest Studies in Philosophy*, 9, 271-81.

Sosa, Ernest 1993: Davidson's Thinking Causes. In Heil and Mele (1993).

Tomberlin, James E., (ed.) 1989: *Philosophical Perspectives*, 3, *Philosophy of Mind and Action Theory*, Atascadero, Ridgeview Publishing Company.

Tomberlin, James E., (ed.) 1995: *Philosophical Perspectives*, 9, *AI, Connectionism, and Philosophical Psychology*, Atascadero, Ridgeview Publishing Company.

Van Gulick, Robert 1993: Who's in Charge Here? And Who's Doing All the Work?. In Heil and Mele (1993).

PHILOSOPHICAL ISSUES, 6
Content, 1995

I'm a Mother, I Worry

Louise M. Antony

I agree with almost everything Prof. Kim has said. In particular, I share his frustration with the Alfred E. Neumans[1] of the philosophical world —philosophers who simply see no problem about reconciling our folksy conviction that what we *think* matters to what we *do*, with our more tutored views about the structure of reality and the nature of causation. Kim identifies and discusses two "deflationary" strategies for dealing with philosophical puzzles about mental causation:

a) "dissolving" the problem by attending (or attending more carefully) to, or by taking more seriously, folk psychological explanation.

b) "diffusing" the problem about mental causation by showing that it generalizes to all of the special sciences.

I agree with everything Kim says about (a), but I think I disagree with some of what he says about (b) —so let me say a little about (a) as a way of working my way around to my disagreements.

---

[1]Professor Kim's title alludes to the popular American humor magazine, *Mad*: "What me worry?" is the motto of the magazine's mascot character, a gap–toothed grinning idiot named Alfred E. Neuman.

With regard to the "dissolution" strategy: I agree with Kim that the problem of mental causation is not an artifact of screwed–up philosophical priorities. But not only that —I think the problem actually *arises* from our taking ordinary folk psychological practice very seriously. Treating psychological explanations seriously as *causal* explanations is what *generates* the problem of mental causation.

The Neumanians who adopt the dissolution strategy seem to me to be confusing the philosophical challenge of *explaining* folk psychology with some skeptical demand for its *justification.* But for those of us who see an issue about the metaphysics of mental causation, it's not our *warrant* for psychological ascriptions that's in doubt. It's just because we *are* warranted that the problem seems so compelling.

Neumanians like to emphasize the practical and psychological indispensability of folk psychology. But it's not the indispensability that provides the warrant —it's the predictive (and presumed explanatory) success.[2] And it's the predictive success that demands explanation. I sometimes get the feeling that these deflationists are trying to insulate folk psychology from empirical risk[3] —but I think that the folk psychology that can be so insulated is not the kind that we're interested in preserving.

Still, deflationists have forced those of us who do think there's a problem about mental causation to get clearer and more precise about the metaphysical commitments that a healthy realism about the mental is supposed to carry. For example, does it really require that mental things satisfy some criterion of "*causal relevance*?" I confess to having very feeble intuitions about this —maybe, since "causal relevance" is a term of art, the problem of demonstrating the "causal relevance of the mental" *is* artifactual.

So what do we really care about here? What does realism about the mental require that brings in the notion of causal relevance?

Here is where I see things slightly differently from the way Kim sees them. I agree that it takes a pretty minimal set of assumptions to get the problem going, and there may be several such minimal sets. But if we take my set rather than Kim's, I think considera-

---

[2]I've been informed by reliable Sunday supplement magazines that it's a characteristic of highly successful individuals that they consistently overestimate the quality of their own performances on simple cognitive tasks. Apparently a belief can be both useful and psychological indispensable without being justified.

[3]Although it would be wrong to classify Davidson as a deflationist, this is clearly his strategy. For discussion of the strategy and its shortcomings, see Antony, forthcoming.

tions will emerge that count against Kim's proposed solution to the explanatory exclusion problem.

Let me here emphasize something Kim said —that there are two kinds of worries one might have about how the mental is supposed to get integrated into the causal order:

i) worries that stem from presumed characteristics of the mental that it shares with the non–mental, such as the second–orderness of mental properties (if some version of functionalism is correct), or the non–localness of intentional properties (if intentional content is external, or historical).

ii) worries that stem from characteristics that are presumed to be peculiar to the mental, such as the normativity of intentionality, or the qualitative character of some conscious mental states.

Those who adopt the diffusionist strategy, strategy (b), are presumably responding to worries of the first sort. But I agree with Kim that if the puzzles about mental causation can be shown to arise for non–mental phenomena as well, then we simply have to solve them there, too. The generalization thesis, if true, can't offer us any more comfort than to provide us some company in our misery.

I have nothing to say here about worries of the second sort, except that I'm sanguine about qualia, and close to clueless about normativity. *My* worries are all of the first sort —I'm worried about how 2nd–order properties, especially functional properties, can be causally relevant.

Kim sees the problem emerging from an apparent competition between two distinct causal explanations of the same event; I rather see it as emerging from two apparently conflicting demands of realism about the mental: one is that mental events have to make a difference, and the other is that they have to make a difference *in virtue of* their mental properties. In both cases, we're trying to avoid some form of epiphenomenalism —in the first case, ontological (or what Brian MacLaughlin calls "token") epiphenomenalism, and in the second, property (or "type") epiphenomenalism.[4]

In the first case, we want to be assured that mental things are not only *there*, but that their being there "makes a difference" to the world. This suggests the counterfactual criterion of causal relevance:[5] if I hadn't wanted some coffee, I wouldn't have gone

---

[4]See McLaughlin, 1989.

[5]LePore and Loewer (1987) formulate two distinct criteria of causal relevance, one counterfactual and one nomic.

downstairs. Eliminativism is the position to beat here. However, this isn't enough —it's not enough for there to just *be* things in the world that are causes and that *happen* to be mental. We realists want it to be the case that these mental things cause what they cause (at least some of the time) *in virtue of being mental.* And now it's the reductionist who's lurking in the background.

"*Sure*", says the reductionist, "sure there are baseballs, and *sure* baseballs cause things to happen, but not in virtue of *being baseballs.* They break windows in virtue of their masses and their velocities; their being or not being baseballs doesn't matter. And so it is with mental things. It's the physical properties of their realizers that do all the causal work; the mentalistic properties just come along for the ride".

Now to see whether or not this is so, we might try to apply the counterfactual test. But it turns out that there are problems when we try to use the test this way. The test seems designed to apply to *tokens.* The murderer fired the gun and thereby killed his victim. When we ask whether it was the loudness of the shot, or the emitting–of–the–bullet–ness that was really causally responsible for the death, we are not interested in whether loudness can, in general, ever cause death, but rather whether it was the loudness *in this case.* To apply the counterfactual test, we must be able to make sense, then, of that particular event's *lacking* each of the properties under consideration. But how are we to do that in the case of a particular mental event if a) its mental properties supervene on its physical properties, and b) if its physical properties are nomically or even metaphysically constitutive of its mental properties?

Perhaps the thing to do is to switch to a different conception of causal relevance —let's try the nomic regularity conception, according to which a property P is causally relevant if there are true causal laws that invoke P. We'll have to be liberal about what we count as laws —we can't require that they be exceptionless— but let's set aside the issue of the legitimacy of *ceteris paribus* laws. Let's assume that a "law" is simply a true generalization that supports counterfactuals and is confirmed by its instances.

We might find such that there are laws in this sense that invoke mental properties, and thus that these properties are certified as causally relevant by the nomic criterion. But this won't be quite enough. We're trying, after all, to defeat the *reductionist* challenge, so what we need to show is that the reference to mentalistic properties in these laws is *essential.* We have to show that psychology is not going to become redundant once we learn exactly how it's realized in the biology.

This brings me to my dissatisfaction with Kim's proposed solution to the exclusion problem as it arises for 2nd–order properties *vis a vis* their first–order realizers. Kim's proposed identification of the 2nd–order properties with the disjunction of its first–order realizers gives the 2nd–order property a very tenuous ontological status —it makes it merely "abbreviatory" (if I can talk that way about properties, as opposed to predicates). Kim acknowledges that his proposal makes 2nd–order properties unprojectible, but says that we'll have to live with that. Well, I just can't. It makes mental properties (and 2nd–order properties generally) too cheap; it makes them look bogus.

Consider: gather up a bunch of things that are (intuitively) heterogeneous in their causal powers, and pair each causally relevant first–order property of each thing with the causally relevant first–order property of the thing's effect. So property C_1 is paired with property E_1, and C_2 with E_2, and so on. Now we can easily —too easily— construct two second–order properties that we can guarantee will be invoked by a true causal law:

Let $\mathbf{C} = C_1 \vee C_2 \vee C_3 \vee \cdots \vee C_n$
(i.e., the property of having a property that causes **E**'s)

and

let $\mathbf{E} = E_1 \vee E_2 \vee E_3 \vee \cdots \vee E_n$
(i.e., the property of having a property that results from **C**'s).

It is guaranteed that '**C**'s cause **E**'s' will be true, but my sense is that **C** and **E** aren't even *real*, much less causally relevant. I don't believe that there's anything that the C_i's *really* have in common, nor the E_i's. Nor does the new, higher–level generalization really add anything to the list of facts in the world. If we don't require projectibility, this is the sort of thing that can happen.

Some have argued[6] that second–order properties can never be causally relevant, and if that's so, then that would account for the difficulty of trying to demonstrate that they are. But I'm not without hope. Consider *dormativity*, and consider the following three singular causal claims:

1) The sleeping pill caused me to fall asleep.

2) The sleeping pill caused my nasal passages to dry out.

3) The sleeping pill caused me to have a car accident.

---

[6]Ned Block, for instance, in Block 1990.

Now what generalization (again allowing *ceteris paribus* laws) backs each of these claims? In each case, there will be some first–order property that in fact mediated the transition from the sleeping pill to the effect. But if we can make a convincing case that the relevant generalization for at least one of them invokes dormativity rather than the first–order property, then we will have made a case for both the reality and the causal relevance of a second–order property.

What certifies (1)? It can't be the generalization that "dormative agents cause people to fall asleep" because the connection between dormativity and somnolence is analytic.[7]ed at all. It won't help much to say that "dormative things cause people to fall asleep" is true, *ceteris paribus*. It probably is, but who ever heard of *ceteris paribus* analyticities? (Maybe I should invent them.) Still I think there's a salvageable intuition here, even if I can't right now work out precisely what it is. What about (2)? No, again —this time because the drying effect was due entirely to properties specific to the realizer— it happened to have been an antihistamine, and the fact that the pill had dormative qualities was utterly irrelevant to the production of the effect.

But what about (3)? Here I think we have what we need — although the particular sleeping pill was an antihistamine, and though it's being an antihistamine was certainly causally relevant to my falling asleep, and hence to my running off the road, we'd still miss an important generalization if we simply said that "antihistamines cause people to have car accidents". For the fact is that tricyclics have very much the same effect when people take them and then go driving —there is something that antihistamines and tricyclics *have in common*, and that is precisely their dormativity. They both cause car accidents *by causing people to fall asleep*.[8]

This, at any rate, is the direction in which I think we must go to solve the explanatory exclusion argument. We have to show that the reference to mental properties is somehow ineliminable, and I suggest the way to do this is to make a case that there are nomic regularities that make essential mention of them.

---

[7]I realize in saying this that I'm glossing over some very tricky issues. I recognize, for example, that the truth of the claim "taking this dormative agent will cause me to fall asleep" is not guaranteed by the meanings of words, since it's not guarante

[8]I wish I'd thought of this example myself, but unfortunately Joe Levine did. So I guess I have to thank him.

REFERENCES

Antony, Louise (Forthcoming). "The Inadequacy of Anomalous Monism as a Realist Theory of Mind", *Zeitschrift Protosoziologie.*

Block, Ned (1990) "Can the Mind Change the World?", in Boolos, G., ed., *Meaning and Method: Essays in Honor of Hilary Putnam*, Cambridge University Press, 137-70.

LePore, E. and Loewer, B. (1987). "Mind Matters", *Journal of Philosophy*, 84: 630-42.

McLaughlin, Brian (1989). "Type Epiphenomenalism, Type Dualism, and the Causal Priority of the Physical", *Philosophical Perspectives*, 3: 109-35.

PHILOSOPHICAL ISSUES, 6
Content, 1995

Kim on the Exclusion Problem

Manuel Campos

I am going to be essentially concerned with Jaegwon Kim's solution to the exclusion problem.[1] I am interested in this point because I think that there is a certain tension between Kim's solution and his seemingly straightforward acceptance of the existence of mental causation. This acceptance, patent throughout the paper, is most clearly expressed at the beginning of it. Thus, Kim explains there that he doubts that many of the people concerned with mental causation are actually worried "that our thoughts and desires might turn out to have no causal powers to move our limbs". He goes on to say that the problem of mental causation is a problem "of showing *how* mental causation is possible, not *whether* it is possible". This latter problem is not at issue for him.

I take this acceptance of mental causation to entail something along the following lines: mental events are causes of the behavior of individuals in virtue of the (mental) properties instantiated in these events. If there is mental causation, mental properties have to be causally endowing properties —properties whose instantiation by particulars confers on these particulars the powers to cause behavioral events.

As Kim explains, the exclusion problem results from having two competing causal explanations for a single event. This is precisely

---

[1]See Kim's "Mental Causation: What? Me Worry?", in this volume.

what seems to happen in the case of mental causation: we have two causal stories —one involving mental properties and the other involving physical properties of the agent's nervous system— that would seem to provide a full account of the causation of behavior. What is then the relation between these two stories? In particular, what is the relation between the mental and neural properties mentioned in them?

Kim discards several possible accounts of this relation. He rejects the overdetermination of behavior by mental and physical events, that mental and neural properties be partial causes in the same chain of events, and that mental and neural properties be part of one another. The solution he finally proposes takes mental properties to be a variety of functional properties, and the causal powers of functional properties to be the disjunction of the causal powers of the realizers of these properties. Thus, from the point of view of causation, mental properties are taken by Kim to be disjunctive.

Causally disjunctive properties are a problematic metaphysical category, as Kim himself has argued in other places (see his note 47). One central reason for this is that they do not seem to endow the entities that instantiate them with causal powers. Rather, it is down at the level of the particular disjunct properties on which the disjunctive property at stake supervenes that one has to look for real causal relevance.[2] Kim seems to acknowledge this when he says that as a consequence of his account of the nature of their causal powers, functional properties turn out to be causally heterogeneous, and that this heterogeneity raises doubts about their nomological–projectible character.

In what sense can one talk then about disjunctive *properties*? It seems that one can talk about them at least in the sense that there are some features shared by the different entities of which these properties are predicated —for instance, in the case of functional properties, the shared feature is a certain pattern of causal dispositions. One can then talk about these entities as having something in common, or as sharing a *property*, but this should not be understood as to entail that this property is causally homogeneous. To illustrate the point, brittle things share some feature; namely, their brittleness (i.e., the fact that they shatter when struck). But the brittleness of these objects is not what causes their shattering. There doesn't seem to be anything common to all brittle objects that we call *brittleness* and such that it causes the breaking of them when struck. It is rather

---

[2]If the disjuncts turn out to be disjunctive as well, then the search for causal relevance has to proceed to a still lower level.

the realizer properties of brittleness that are causally responsible for the shattering.

Kim's solution to the exclusion problem offers an account of what the relation between mental properties and neural properties is. If we look at it from a perspective of interest in causation (which is the relevant one when it comes to the exclusion problem), the relation between mental properties and their neural realizers is taken by Kim to be the relation that holds between disjunctive properties and the disjuncts on which they supervene. Does this eliminate the competition between the two causal stories mentioned before? Don't we then still have two sets of properties, mental and realizer properties, supposed to account for the same contribution to the causation of behavior? The answer to this should be that though we still have two sets of properties, properties in one of these sets are not causally relevant because they are disjunctive. Hence, they don't compete with the properties of the other set.[3]

The catch of this solution is, of course, *epiphenomenalism*; i.e., the questioning of mental causation. The reason for this is, as I said, that if mental properties are as causally heterogeneous as Kim thinks they are, then it is at the level of their realizers that we have to look for the causal powers we associate with the mental. Naturally, this consequence seems very difficult to mesh with Kim's apparent acceptance of mental causation (and, I take it, of the causal relevance of mental properties).

There are some possible alternatives worth considering to try to avoid the former dilemma that I would like to mention. Thus, for instance, functional properties might be thought to have a casual relevance of their own. However, I guess that this possibility would be difficult to combine with the strongly reductionist stance Kim is known for. On the other hand, as Kim himself points out, some mental properties might turn out not to be functional properties (in the 'causal specification' sense of "functional"), but, for example, a variety of extrinsic properties of brain states. I would guess, however, that trying to attach some type of causal relevance to such properties would also conflict with Kim's reductionist position. Finally, maybe there is a way of understanding mental causation (and causation in general) that doesn't entail that mental properties have to be causally endowing properties. In any case, these are points that go beyond the scope of Kim's paper, so I will conclude here.[4]

---

[3]Kim has argued along this lines other times (see his note 40).

[4]I'd like to thank Fred Dretske and John Etchemendy for their comments.

PHILOSOPHICAL ISSUES, 6
Content, 1995

Ruritania Revisited[1]

Ned Block

Perhaps you are wondering what I mean by 'holism'. After all, everyone seems to use the term in a different sense. Even if we restrict ourselves to holism of meaning and content, we have many different

---

[1]This paper is very similar to my "An Argument for Holism", forthcoming in *The Proceedings of the Aristotelian Society.* It appears here with the kind permission of The Aristotelian Society. The argument of the paper is descended from Hilary Putnam's famous Ruritania argument in "Computational Psychology and Interpretation Theory", in *Realism and Reason, Philosophical Papers* Vol. 3, (Cambridge: Cambridge U. Press, 1983.) I published a very short version of the revised Putnam argument in "What Narrow Contents Are Not" in B. Loewer and G. Rey, *Fodor and His Critics*, (Oxford: Basil Blackwell, 1991). I discovered later that Hartry Field had come up with a similar revision of Putnam in an unpublished paper. I am very much indebted to Field's paper and to conversations with Field; the paper could reasonably have both of our names on it. Field rejects the conclusion, arguing that instead one of the premises (what I am calling Field's Principle) should be rejected. (Maybe it should be Field's Anti–Principle to make it clear that he rejects it.) Versions of this paper were given at a conference in honor of Tyler Burge in Vancouver, in the Fall of 1993; and at the following conferences and meetings in the summer and spring of 1993: the NEH Summer Institute at Rutgers, the meeting of the Sociedad Filosofica Ibero–Americana in Tenerife and at a conference at the University of Maryland. I am grateful to audiences at those occasions. I hope to write a longer version to be included in a volume of papers from the conference in honor of Burge.

holisms. Some take holism about meaning to be the doctrine that if you've got one meaning, you've got lots of them.[2] On other views, to say meaning is holistic is to say that the meaning of each term depends on the meanings of all or most other terms.[3] Others take meaning holism to be the doctrine that there is no real distinction between language and theory or between the "dictionary" and the "encyclopedia".[4]

Although everyone seems to favor a different definition of meaning holism, there is widespread agreement that some versions of meaning holism are *extremely implausible*, and for a specific reason, namely that they lead to the following: Meaning depends on belief; if any of my W–beliefs change (that is, if any of the beliefs that I would standardly express using the word W change), then my word W changes meaning. And if any of my W–beliefs differ from your W–beliefs, then what you mean by $W \neq$ what I mean by W. Now suppose I accept and you reject "The Pantheon's lead was used to make a canopy." It follows that we don't share meanings of any of the words in this sentence.[5] So the meaning of the sentence that I accept isn't the same as the meaning of the sentence that you reject. So how can people ever disagree? Moreover, we may both accept "Lead is heavy", but since we don't mean the same by "lead", agreement is problematic too. Further, if I accept a sentence and later reject it, then the meaning of what I accept is not the same as the meaning of what I later reject, so how can I ever change my mind?

---

[2]Jerry Fodor and Ernest LePore, *Holism: A Shoppers' Guide*, (Cambridge: MIT, 1992).

[3]Michael Devitt, "A Critique of the Case for Semantic Holism", in *Philosophical Perspectives* 7, *Language and Logic, 1993.* (Atascadero: Ridgeview: 1993). Devitt gives some objections to the line of argument I give here which are keyed to views that do not appear in this paper. I will discuss some of Devitt's points in the longer version; in my view, the immunity of the points in this paper to Devitt's criticisms show that the criticisms do not get at the heart of the argument. See also Devitt's *Coming to Our Senses: A Naturalistic Program for Semantic Localism*, forthcoming.

[4]Quine is the leading exponent of these views; sometimes he is taken to say that the unit of meaning is the whole language rather than words or sentences, but this form of words has a variety of interpretations, some of which are the ones just mentioned in the text. Putnam has used 'meaning holism' to mean that meanings have an identity through time, but no essence. See for example *Representation and Reality*, (Cambridge: MIT, 1988), but this notion seems very different from the family of ideas mentioned in the text.

[5]This needs qualification. What follows is that your 'lead' differs in meaning from my 'lead', your 'canopy' differs from mine, etc. But your 'lead' could still mean the same as my 'canopy' and conversely. Then it would be possible that we didn't disagree at all.

I propose to avoid the issue of whether one or another version of holism really has this horrible consequence by simply *defining* 'holism' as a version of the horrible consequence. Let holism be the doctrine that any *substantial* in difference in W–beliefs, whether between two people or between one person at two times, requires a difference in the meaning or content of W. I propose to argue *for* the horrible consequence, that is, for holism in this sense.[6]

Of course, the interest of the conclusion hinges on the interpretation of 'substantial'. The argument itself will fill in what I have in mind. I hope you will agree that I am not putting the kind of restrictions on 'substantial' that would make the thesis uninteresting.

So far I have not said what I mean by 'meaning' or 'content'. I will use 'meaning' and 'content' more or less interchangeably.[7] Meaning and content in this paper are *narrow* meaning and content, if such there be. I will not assume that there is such a thing. Rather, I am arguing for a conditional: if there is such a thing as narrow content, it is holistic in the sense described. One person's *modus ponens* is another's *modus tollens*, so the upshot for some readers may be that narrow content is holistic, and for others that there is no such thing as narrow content.

At this point it would be nice if I could tell you what narrow content is. All I will say is this: Narrow content is internal content, content "inside the head". Arguably, some beliefs supervene on the non–relational, physical properties of the body. These are beliefs that I necessarily share with any döppelganger, any molecular duplicate of myself, no matter how different the duplicate's environment or language community. For example, perhaps the belief that $2 + 5 = 7$ is one I must share with any döppelganger. If so, then the belief is narrow and its content is a narrow content. Some proponents of narrow content suppose that every belief has both wide and narrow content, the narrow content being what is needed for explanation of behavior; others suppose that beliefs have only narrow content.[8]

---

[6] Jane Heal argues that this doctrine (as well as a version of the third one mentioned above) can be ascribed to Fodor and LePore. See her "Semantic Holism, Still a Good Buy" in *Proceedings of the Aristotelian Society* XCIV, 1994, 325-340.

[7] I tend to use 'meaning' in connection with words and 'content' with sentences.

[8] Jerry Fodor and David Lewis argue for both narrow and wide content, but they give rather different accounts of why one should accept narrow content. See Fodor's *Psychosemantics: The Problem of Meaning in Philosophy of Mind*, (Cambridge: MIT Press: 1987) and Lewis, "Reduction of Mind" in Samuel Guttenplan (ed), *A Companion to Philosophy of Mind*, (Oxford: Basil Blackwell: 1994). Fodor now rejects narrow content. See his *The Elm and the Expert*, (Cambridge: MIT: 1994).

Perhaps you are disappointed with this meager account. I don't say more for two reasons. First, I don't have an account of narrow content. Second, even if I did, I'd be reluctant to offer it because my argument does not depend on any specific theory of what narrow content is, but rather only that it is narrow and explanatory (and other conditions to be spelled out below). To give an account would be to encourage the idea that my argument depends on it.

If you believe that content *is* narrow content or that there is both narrow and wide content, then presumably you will regard the conclusion that I will argue for as interesting (though it would be rational to wait to see what the "substantial" qualification comes to). But even if you are agnostic about narrow content or feel at sea about what it might be, still the conclusion ought to be of interest to you. If we can establish for sure that any content that is narrow is holistic, that may help us to think about whether wide content is holistic or why it has seemed to be holistic. There has long been a gulf between holists and anti–holists. Perhaps it will turn out that they have been talking past each other: holists are right about one thing, anti–holists about another.

1 The Premises

The assumptions made so far are relatively uncontroversial and will remain in the background. Here are the assumptions that will be in the foreground:

NARROWNESS. Narrow content supervenes on non–relational physical features of the body. Or in slogan form, narrow content is narrow. This is just a definition.

DIFFERENCE. If at one time, a person has substantially different beliefs associated with term t_1 and t_2, then t_1 differs in narrow content from t_2 for that person at that time. So for any normal person, words like 'cat' and 'dog' and 'panda' have different narrow contents.

EXPLANATION. Narrow content's main purpose is its role in psychological explanation.

INCOMPLETE UNDERSTANDING. Incomplete understanding and full mastery of a concept are completely compatible. This idea should be familiar from the work of Tyler Burge. I can have full mastery of the concept of arthritis, so that it is correct to

ascribe to me beliefs such as that arthritis is a disease that I expect but don't want, even if my understanding of the concept is incomplete. I will also assume that where there is a wide concept there is a narrow concept.

INTER/INTRA: FIELD'S PREMISE.[9] The relation of *same narrow content* that holds between people is the same relation of *same narrow content* that holds within a single person. So if my word X has the same narrow content as my word Y, and my word Y has the same narrow content as your word Z, it is legitimate to conclude that X is the same in narrow content as Z.

These principles are not independent. For example, one of the roles of EXPLANATION is to bolster DIFFERENCE. If a theorist holds that a single person's 'dog' and 'cat' have the same narrow content, one should wonder what this theorist thinks narrow contents are *for*. Any narrow contents that are usable for psychological explanation will have to be more fine–grained than that.

2 The Example

The argument will be based on a version of Putnam's Ruritania example. There are two parts of Ruritania, B and W. Bruce lives in B, Walter, his twin, lives in W. The dialects of the two parts are exactly the same save for the fact that in B, they use 'grug' instead of 'beer', whereas in W, they use 'grug' instead of 'whiskey'. The B dialect lacks 'beer'; the W dialect lacks 'whiskey'. Ruritanian = English except for the use of 'grug' in the two dialects. So in B, 'whiskey' means whiskey and in W, 'beer' means beer. At age 10, Bruce and Walter alike in every relevant respect. We could suppose that they are molecular duplicates. In particular, the difference between the different substances known as 'grug' in their communities has not impinged on them at all. They share all beliefs, images, recognitional dispositions and the like having to do with the referent of 'grug' in their communities. If asked about grug, the following is as much as could be squeezed out of them:

- "Grug is a brownish liquid."
- "Drinking grug makes grownups act funny."

---

[9]Field points out that this premise is assumed in the argument and argues that we should reject it rather than accept the conclusion.

- "Grownups like to drink grug at social occasions."
- "Grug is bought in "liquor"stores."
- "Grug is often served before dinner."
- "It would be peculiar to drink grug with breakfast."

Two years pass in which Bruce and Walter become more integrated into their societies, learning more about the different items called 'grug' by their language communities and the connection between those terms and English. Here is what they know at age 12: Bruce could give voice to the following:

1. "'Grug' translates in English to 'beer'."

2. "Grug comes in small cans (in both parts of Ruritania)."

3. "More than six cans makes people very drunk."

4. "Grug is relatively cheap."

5. '"Grug' in the other dialect is characterized by 1-4 below."

Walter could give voice to the following:

1. "'Grug' translates in English to 'whiskey'."

2. "Grug comes in liter bottles (in both parts of Ruritania)."

3. "One glass knocks you out."

4. "Grug is expensive."

5. "'Grug' in the other dialect is characterized by 1-4 above."

Both twins are bilingual, and they have all the same beliefs —at least they would utter all the same sentences— except for indexical 'grug' beliefs. In contexts in which B–terminology is appropriate, both say "Grug is cheap and comes in 6–packs." In contexts in which W–terminology is appropriate, both say "Grug is expensive and comes in liter bottles". In contexts in which neither dialect is singled out, they use indexicals. Walter says "'Grug' is *our* word for whiskey," Bruce says "'Grug' is *their* word for whiskey."

3 The Argument

The basic idea of the argument is simple. At age 10, Bruce's 'grug' has the same narrow content as Walter's 'grug'. But at age 12, the narrow contents of their native or home 'grug's differ. So at least one must have changed, and symmetry requires that both have changed. Why is any further argument needed? The first premise (the identity at 10) depends only on the definition of 'narrow'. But what is the justification of the second premise, that the narrow contents are different at age 12? If we have wide content in mind, the second premise will seem just obvious —it is our practice to take reference and beliefs into account in deciding about translation, and in this case both the reference and the beliefs are different, and further, there is a competitor: each twin's foreign 'grug' seems the right translation of the other twin's home 'grug'. But if it is wide content we have in mind, the first premise is false.

Concentrating on narrow content, it may still seem obvious that the twins' home 'grug's differ in narrow content at age 12. The twins' home 'grug' beliefs are very different. E.g., using their home 'grug's, one says "Grug is expensive," and the other says "Grug is cheap." As I said above, in regard to a different matter, if we are to regard these as the same in narrow content, what would narrow content be for? Any narrow contents usable for psychological explanation will have to be more fine–grained than that. But now it looks as if there is very little difference between the premises of the argument and the conclusion. After all, 'grug' at age 10 and 'grug' at age 12 differ in associated beliefs too. If belief differences between people count for differences of narrow content, why shouldn't belief differences between two stages of one person count as well? And then we could dispense with the argument altogether!

So the purpose of the added complexity in the argument to follow is to justify the claim that at age 12 their home 'grug's differ in narrow content. The idea is to justify this inter–personal claim by appeal to an *intra–personal* claim, the claim that each twin's two 'grug's differ in narrow content from each other, a consequence of the DIFFERENCE principle.[10] But isn't this just a matter of a difference in beliefs too? There is an important difference, but I'll wait until later to say why.

One final preliminary: the argument is complicated, so to make it

---

[10]My 1991 rendition of the argument (mentioned in footnote 1) involves an unhappy compromise between the simple and the complex versions, in which I attempted to justify the second premise by appeal to the intra–subjective difference between 'beer' and 'whiskey'.

easier to follow, I will adopt a simplification. I will speak of sameness and difference of narrow meaning and content as identity and difference, simpliciter, using '=' and '≠' as symbols. '$\text{Word}_1 = \text{Word}_2$' is to be understood as saying that the two words have the same narrow content.

I. At age 10, Bruce's 'grug' has the same narrow content as Walter's 'grug'. At age 10, Bruce and Walter are molecular doppelgangers, so all their words have the same narrow content, by the principle that narrow contents are narrow (NARROWNESS). But do the boys' 'grug's have narrow content at all? Here is where the INCOMPLETE UNDERSTANDING principle comes in —to justify the claim that they have narrow contents despite a large measure of ignorance.

II. At age 12, both boys understand the home 'grug' and the foreign 'grug', and they attach different narrow contents to them. Bruce's 'grug_B', that is, his word 'grug' in his home dialect used to mean 'beer', has a different narrow content from his 'grug_W', the word that as a bilingual he uses to mean what 'grug' means in the other dialect, namely, whiskey. As explained above, I will put this by saying that for Bruce, 'grug_B' ≠ 'grug_W'. (See Figure 1.) Here I appeal to the DIFFERENCE principle, the principle that says that substantial differences in belief associated with two terms at one time make for different narrow contents, as with 'cat' and 'dog' for a normal speaker. For Bruce, 'grug_B' and 'grug_W' are as different in narrow content as your 'whiskey' and 'beer'. For example, he knows that 'grug_B' translates in English to 'beer', and 'grug_W' translates in English to 'whiskey'.

III. At age 12, one boy's foreign 'grug' is the same as the other's home 'grug'. Bruce's 'grug_W' = Walter's 'grug_W', that is, Bruce's 'grug_W' has the same narrow content as Walter's 'grug_W'. (See Figure 1.) Recall that Bruce and Walter are döppelgangers at age 12 except for indexical 'grug' beliefs having to do with *whose* 'grug' is in question. Except for the indexicals, they are the same with respect to the word 'grug' used to mean whiskey. I appeal to the NARROWNESS principle, and to the EXPLANATION principle to justify the claim that the indexical difference doesn't make a difference to narrow contents. (More on this later.) This is the step in the argument that requires the example's two 'grug's. If not for the need to make Bruce and Walter döppelgangers, I could have run the argument with different words for the two 'grug's.

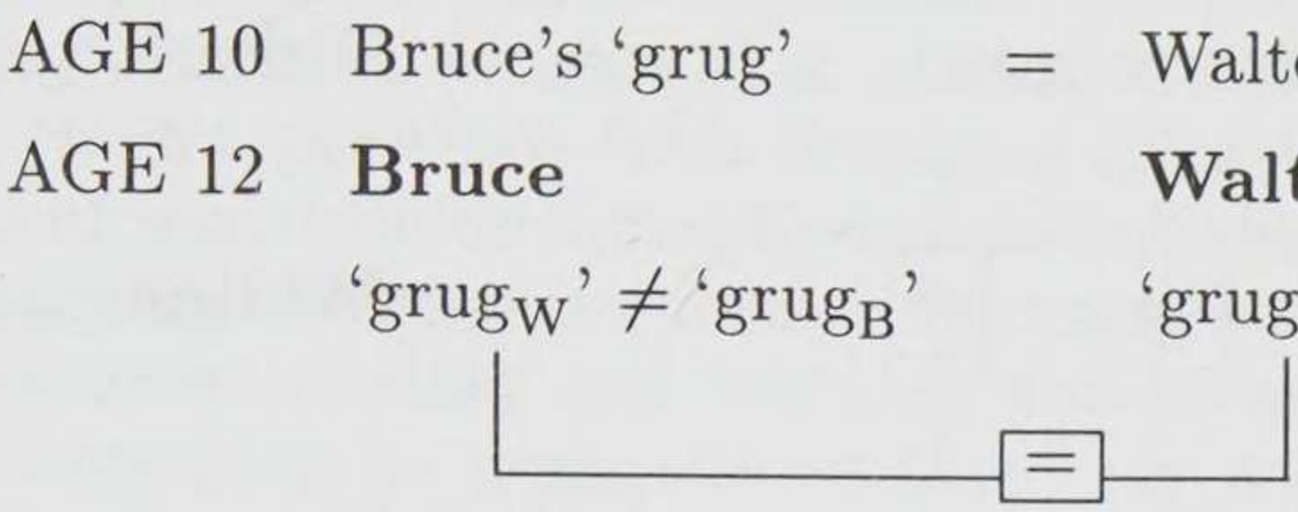


FIGURE 1. At age 10, Bruce's 'grug' = Walter's 'grug'. At age 12, Bruce's 'grug$_W$' ≠ Bruce's 'grug$_B$', and Bruce's 'grug$_W$' = Walter's 'grug$_W$'. The situation is symmetrical as between Bruce and Walter, but superfluous detail on the right hand side has been left out to avoid over–complicating the diagram.

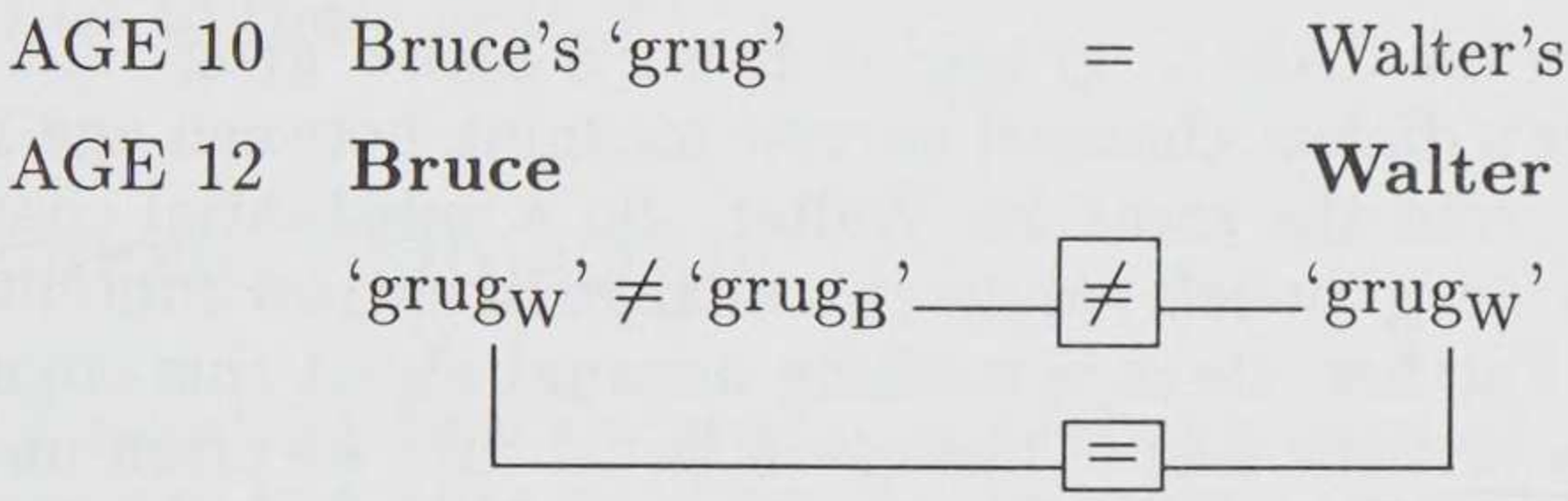


FIGURE 2. Bruce's 'grug$_B$' ≠ Walter's 'grug$_W$'.

IV. It follows that at age 12, one boy's home 'grug' ≠ the other boy's home 'grug'. Since Bruce's 'grug$_W$' ≠ Bruce's 'grugB', and since Bruce's 'grug$_W$' = Walter's 'grug$_W$', it follows by the logic of identity (or rather identity of narrow meaning or content) that Bruce's 'grug$_B$' ≠ Walter's 'grug$_W$'. (A representation that's not synonomous with one of a pair of synonomous representations can't be synonomous with the other either, and the same holds for narrow synonomy.) See Figure 2.

V. At age 10, Bruce's 'grug' = Walter's 'grug'. At age 12, Bruce's 'grug$_B$' ≠ Walter's 'grug$_W$'. It follows by the logic of identity that either Bruce's 'grug' at age 10 ≠ Bruce's 'grug$_B$' at age 12 or Walter's 'grug' at age 10 ≠ Walter's 'grug$_W$' at age 12. And since there is no asymmetry in the details of the case and therefor no reason to treat one child differently from the other, *both* Bruce's 'grug' at age 10 ≠ Bruce's 'grug$_B$' at age 12 *and* Walter's 'grug' at age 10 ≠ Walter's 'grug$_W$' at age 12. See Figure 3. But Bruce's 'grug$_B$' at age 12 just is the word

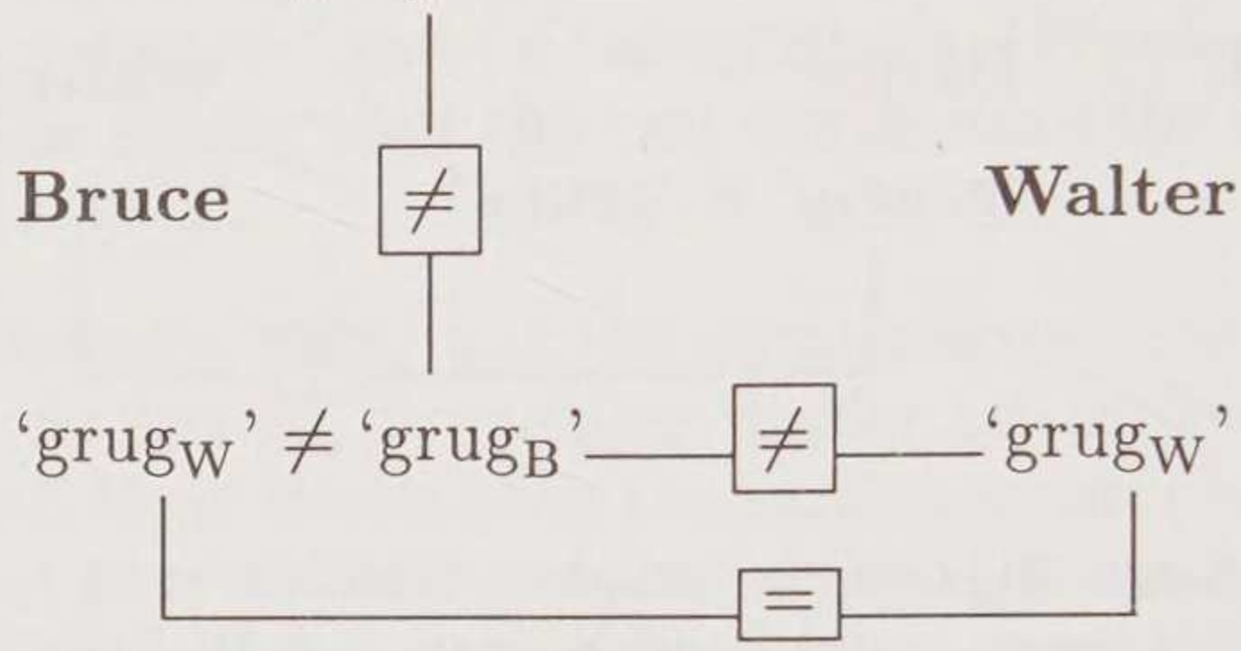


FIGURE 3. Bruce's native 'grug' at age 10 ≠ Bruce's native 'grug' at age 12.

'grug' in his dialect; 'grug$_B$' is Bruce's home 'grug'. So 'grug' in Bruce's dialect changed narrow meaning between age 10 and age 12. And the same for Walter. So a substantial change in Bruce's 'grug' beliefs results in a change in narrow content of his 'grug'. Further, there is nothing unusual about this change. It involves just the sort of change in belief that we often undergo. Reading the New York Times can induce this sort of change in a single sitting.

Note that the argument assumed no particular *theory* about what narrow content is. Some say that narrow content is functional role, others that it is observational content, others that it is specified by a function from contexts of language acquisition to the contents acquired in those contexts and others that it is the same as wide content. But I have not assumed the truth or falsity of any of these views. Of course, the list of assumptions given above does put some restrictions one what narrow content could be, but I hope these restrictions will not be controversial.

4 Essential and Analytic Properties

In arguing against Putnam on Ruritania in *Psychosemantics*,[11] Fodor suggests that holism can be avoided. He says "Learning what anything really is changes one's narrow concept of that thing", but learning other sorts of features do not change narrow concepts. The

[11] *Psychosemantics*, op. cit., pp. 94-95.

suggestion is that we can avoid holism by distinguishing between two classes of truths —those that attribute some sort of essential property (Fodor prescinds from declaring allegiance to any particular form of linguistic or metaphysical essentialism) and those that don't. Learning the former features do change narrow concept, but that is unexceptionable, and learning the latter features do not. But this suggestion is powerless to avoid my argument, for in my example Bruce and Walter do not learn anything that might count as learning what "grug" really is. Here's what beer really is: a fermented alcoholic beverage, brewed from malt and flavored with hops that is less than 20% alcohol. Here's what whiskey really is: an alcoholic beverage distilled from fermented mash of grain (corn, barley, rye), aged in wood, and roughly 40%-50% alcohol. Note that none of the things that Bruce or Walter learn are very closely connected to these facts.[12]

5 Indexical Objection

In III, I said that Bruce's 'grug_{W}' = Walter's 'grug_{W}'. I noted that the twins are just alike with respect to 'grug_{W}' except for indexical beliefs, appealing to NARROWNESS, the principle that says that narrow content is supervenient on the body. Though Bruce and Walter are no longer perfect duplicates by age 12, they are the same with respect to 'grug_{W}'. Well, almost the same. There is the indexical difference I mentioned. But does the indexical difference make a difference to psychological explanation? If not, EXPLANATION, the idea that the main purpose of narrow content is psychological explanation, dictates that the indexical difference doesn't make a difference. But now I'm in trouble, for indexical differences are famous for making an explanatory difference. Famously, if I think my pants are on fire, I jump into the pool, but if I think your pants are on fire I push you in.

True, and there will be important behavioral differences between the twins that hinge on the indexical difference. If they are both told that the part of the country in which 'grug' is used for whiskey has

---

[12]They do learn the translation of 'grug' into English, but learning that isn't learning an essential property unless they know an essential property under the English description. I make use of their knowing the translation only in making them perfect twins at age 12 aside from indexical beliefs. This helps to motivate the idea that one boy's foreign 'grug' has the same narrow content as the other's home 'grug'.

been invaded, Walter, but not Bruce, will be worried about the fate of his house and family. But the difference comes from the indexicals, not from the narrow content of 'grug' itself. Bruce says "'Grug' is their word for whiskey," whereas Walter says "'Grug' is our word for whiskey." It is the indexical difference that makes the difference.

6 Splitting Objection

Does the argument commit a well–known fallacy about the notorious "splitting" cases made famous in recent discussions of personal identity? Suppose that next year you split into two persons more or less just like you are now. (Each successor has half of your cells combined with duplicates of the other half.) Call the two new people 'A' and 'B'. $A \neq B$, since A and B occupy distinct places at one time. But then you can't be identical to *both* A and B, since one thing can't be identical to non–identical things. And since there is no relevant difference between A and B, you $\neq A$ and you $\neq B$. But no one should conclude that the mere possiblity of a split shows that I am not the same person who wrote the first word in this sentence. Even if an actual split undermines identity over time, that does not show that the mere possibility of a split does so. That would be a fallacy. But is my argument an instance of the same fallacy? How can the mere possiblity of a split of 'grug' into 'grug_B' and 'grug_W' show that cases of substantial changes in belief without any such split are changes of narrow meaning or content?

But the 'grug' example, spelled out in the way I would fix on in response to this objection, is not a case of 'grug' splitting into 'grug_B' and 'grug_W'. Bruce's word 'grug' at age 10 is his native or home 'grug'. I have called his home 'grug' at age 12 'grug_B', but this terminology should not mislead us. These are the same words at different times. Words can maintain an identity over time just as people can. (Note that I have temporarily abandoned my terminology of talking about identity of narrow content in terms of identity of words; 'grug' at age 10 is the same word as 'grug_B' at age 12, but they have different narrow contents.) To use a popular metaphor, he has a 'grug' "file" at age 10. As he learns more about beer, he puts more information in his 'grug' file. At age 12, he has quite a bit of information in his 'grug' file —"Grug is cheap," "Grug comes in 6–packs", etc. The 'grug' file at 10 is the same file as the 'grug_B' file at age 12. When he learns about whiskey, he opens a new file, the 'whiskey' file. I hereby stipulate that this all happens before he learns about the other part of Ruritania or their dialect. When

he learns about the W–Ruritanian dialect, and the fact that they use 'grug' to mean whiskey, that's just another bit of information in his 'whiskey' file. Of course, I'm filling in details that were left vague in the original description so as to fit this case, but that is perfectly legitimate because it does nothing to weaken the original argument and if I am right, it allows a firm rejection of this objection.

Suppose you find out that the term 'beer' is used in Outer Mongolian to mean whiskey. Does that make your 'beer' "split" in any way that raises one of these identity over time cases? I assume not, and the same is true for our twins. I could have run the example without Bruce and Walter having learned the word in the other dialect of Ruritanian at all. My purpose in running the example the way I did was to make it easier to justify the idea that Bruce's 'grug_W' = Walter's 'grug_W'. I appealed to the idea that they are exactly alike except in indexical beliefs. But even if the "exactly alike" had to be weakened a bit, I think that premise would not be very much weakened.

7 Small Distance Objection

I mentioned earlier that there is a simple version of the argument that suffers from an excessively small difference between premises and conclusion. The simple version is: the twins' native dialect 'grug's are the same in narrow content at age 10 and different at age 12. So at least one must have changed, and given symmetry considerations, both changed. The problem is: What's the justification of the claim that the twins' home 'grug's are different at age 12? Well, its true that they have very different beliefs connected with them (e.g. "expensive" vs "cheap") But if differences in belief make for a difference in narrow content, why not just appeal to that directly to support the difference between 10 and 12 and dispense with the argument altogether? Further, the DIFFERENCE principle just appeals to a difference in belief, and so doesn't it just beg the question in the same way?

Let me reply by mentioning a justification for the DIFFERENCE principle. So far, all I have said is: what good would narrow contents be if a person's 'cat' and 'dog' (at one time) had the same narrow content? But doesn't that apply between people and within a person over time too? We can see a difference between the other cases and the intra–personal case at a time by considering what we might call incoherent cognition. By this phrase, I mean the kind of confused thinking and acting that a person engages in if he has contradictory beliefs. If I accept "Pandas are cute"and "Pandas are not cute",

that could cause and causally explain incoherent cognition of a sort that would not arise in someone who accepted instead "Pandas are cute" and "Cats are not cute". 'Panda' has the same narrow content in both occurences, so "Pandas are cute" and "Pandas are not cute" are incompatible contents. 'Panda' and 'cat' have different narrow contents, so "Pandas are cute" and "Cats are not cute" are not incompatible contents. The file metaphor may help: "is cute" and "is not cute" are fine in different files, problematic in the same file. In the same file, they lead to incoherent cognition. So here's why the DIFFERENCE principle is right: different beliefs require different files and different files involve different narrow contents. Note that no such reasoning will apply in justifying an inter–personal version of the DIFFERENCE principle or one that applies to two stages of a single person. Incoherent cognition only operates in the intra–personal synchronic case.

But wait! The file explanation doesn't depend on there being different *beliefs* in the two files. Suppose I have two different files headed 'panda'. Both contain "Furry" and "Found in Asia" and "Not identical to the other animal called "panda". Both files have the same beliefs.[13] In fact, this does describe my epistemic situation some years ago. (Now I know more —that one is the great panda and the other is the lesser panda.) All that we need for different narrow contents in a single person are different *files* —the contents can be the same.[14] So a line of thought which seemed to back up the DIFFERENCE principle actually appears to argue for something much stronger, something like this: if at one time a person thinks Xs are distinct from Ys, then 'X' differs from 'Y' in narrow content for that person at that time. Differences in beliefs are only relevant because it would be hard for one to have 'X' beliefs that are different from one's 'Y' beliefs without thinking that Xs are distinct from Ys.

The upshot is that I could have motivated the difference in narrow content between the two 'grug's within each twin at age 12 by ap-

---

[13]Why should we say two files that are the same instead of one file that is written down redundantly? Even if there is no functional difference between the two files there could still be a functional difference between two files and one file.

[14]See Putnam's 'elm'/'beech' point, Kripke's 'Paderowski' example, and Loar's 'chat' example. Putnam, H., "The Meaning of 'Meaning' in *Mind, Language and Reality* ,Philosophical Papers Vol. 2 (Cambridge: Cambridge University Press, 1975). S. Kripke, "A Puzzle About Belief" in *Meaning and Use*, A. Margalit (ed), (Dordrecht: D. Reidel: 1979). B. Loar, "Social Content and Psychological Content" in *Contents of Thought*: Proceedings of the 1985 Oberlin Colloquium in Philosophy, R. Grimm and D. Merrill (eds), (Tucson: University of Arizona Press, 1987).

pealing just to the fact that each twin thought the two words picked out different things. There was no real need to appeal to a difference in beliefs. So if the DIFFERENCE principle is reformulated as just indicated, the premise that says that a twin's two 'grug's at age 12 are different in narrow content does not depend on a difference in beliefs. Hence the step in the argument that says that the twins' native 'grug's are different in narrow content at age 12 is suitably distant from the conclusion about change over time.

8 Field's Principle

Thus far the role of Field's principle (that is, the principle Field points out is required for the argument but, according to Field, should be be rejected), INTER/INTRA has been mainly in the background. This principle says that the intra–personal relation of *same narrow content* is the same as the inter–personal relation of *same narrow content*. One of the places in which the principle was used was in the last stage of the reasoning. At age 10, the native 'grug's are the same. At age 12, the native 'grug's are different. So one or both must have changed between 10 and 12. The content identity in the premises is interpersonal, whereas the content identity in the conclusion is intrapersonal. Are these the same identity relations? Field says no, but I disagree.

Here is one possible justification for the claim that intra–personal content identity is the same relation as interpersonal content identity. Any relation between representations relevant to psychological theory that can obtain between representations of 2 people can also obtain between a representation of or in a person at one time and a representation of the same person at another time. Consider, for example, a difference, D, between representations of Oscar and Elmer that explains why a given stimulus causes Oscar to do one thing and Elmer to do another. D can also explain in just the same way why Oscar does one thing at one time and another thing at another time. From the point of view of psychological theory, any explanatory difference or similarity that can be found operating between representations in two people can also be found operating between two stages of a single person.[15]

---

[15]The converse is not required for the argument. I do not claim the converse, because I think there are psychologically interesting principles of representational change over time. So what I hold is that interpersonal representational relations are a subset of intra–personal relations over time.

This justification is fine as far as it goes. But it doesn't go far enough because it leaves out an important case. There are three relevant content identity relations:[16]

- Intra–personal + Diachronic
- Inter–personal + Synchronic
- Intra–personal + Synchronic

The reasoning in step V just described includes the first two of these. Bruce and Walter are relevantly the same at one time, different at another time, and a conclusion is reached about a difference within each over time. I believe that the justification mentioned can show that this reasoning is OK. But there is no point in going into the matter, because the reasoning in steps II, III and IV includes the latter two relations, and the justification given does nothing to show that these are the same.

In II, I said that Bruce's two 'grug's had different narrow contents. This is intra–personal and synchronic. In III, I said that one's foreign 'grug' = the other's home 'grug': inter–personal and synchronic. And in IV I concluded that at age 12, their home 'grug's are different: again inter–personal and synchronic. Both the premises and the conclusion are synchronic.

Is the last of the three content identity relations listed above the same as the first two? Well,there certainly is a large difference. As I pointed out, no difference in beliefs is required for intra–personal synchronic narrow content difference. All we need are two files, even if the same things are written in both of them. Beliefs seem much more relevant to the other types of comparisons. Consider, for example, the rationale for claiming that Bruce's 'grug_{W}' has the same narrow content as Walter's 'grug_{W}' rather than Walter's 'grug_{B}'. The two 'grug_{W}'s have all the same non–indexical beliefs associated with them, whereas one twin's 'grug_{W}' is the same as the other's 'grug_{B}' only in indexical beliefs.

So are the narrow content identity relations the same or not? If there is a well–defined intra–personal narrow content identity relation and a well defined inter–personal relation, trivially we can compose these relations can get a well–defined narrow content identity relation that is both inter and intra–personal. But that doesn't answer the question. Consider the done–by–one–person relation, $S_{XY} \leftrightarrow X$ and Y are done by the same person. Consider the done–by–cousins

---

[16] I am indebted to discussions with Field and Brian Loar.

relation $C_{XY} \leftrightarrow X$ and Y are done by cousins. The first is internal and the second is external. We can compose them to get the a relation that holds of a pair of actions if and only if they are done by the same person or by cousins. But this trivial exercise sheds no light on whether C and S belong to the same kind. To say that X and Y are the same in narrow content is to say that each has a narrow content, and that they are identical. So the question of one versus two (or three) identity relations is really a question of whether there is more than one *kind* of narrow content. If there is more than one identity relation, then the twins' 'grug's have two or more kinds of narrow content, one of which governs intra–personal relations at a time, another of which governs inter–personal relations.

If the kind of narrow content that comes into intra–personal synchronic comparisons is different from the kind that comes into the other comparisons, the sting goes out of the problems of holism that I started with. We were worried about the idea that you can't change your mind and that two people never disagree. But if there are two or more different kinds of narrow content, each appropriate to different comparisons, we have departed so much from common sense that we can hardly expect common sense ideas to apply.

So I draw a disjunctive conclusion: If there is one kind of narrow content, it is holistic. If there are more, then maybe holism is true but it loses its sting.[17]

[17]I am grateful to Paul Boghossian and Paul Horwich for comments on a previous draft, to Tyler Burge, Martin Davies, Michael Devitt, Jerry Fodor, Brian Loar and Hartry Field for helpful comments at meetings at which earlier versions were delivered, and to comments from the audience in Robert Stalnaker's and my graduate class, especially to Alex Byrne, Ned Hall, Diana Raffman, Robert Stalnaker and Daniel Stoljar.

PHILOSOPHICAL ISSUES, 6
Content, 1995

Ruritania and Ecology

Josefa Toribio

1 Disjunctions

Ned Block has argued for the truth of the following conditional: If there is such a thing as narrow content, it is holistic.[1] His argument, we've been told, doesn't assume any particular theory of what narrow content actually *is*. But, whatever it is, the following five assumptions put some restrictions on it:

(1) Narrowness: narrow content is narrow, i.e., it supervenes on non relational physical properties.

(2) Difference: terms that belong to the same category can have different narrow contents, because one has *substantially* different beliefs associated with those terms. To be *substantially different* is to be different in the same way that beliefs associated with "cat" are different from beliefs associated with "dog".

(3) Explanation: narrow content has to serve the purposes of psychological explanation.

---

[1]Holistic in the following sense: any substantial difference in the beliefs that I would standardly express using the word W, whether between two people or between one person at two times, requires a difference in the meaning of W.

(4) Incomplete understanding: narrow concepts don't exist without wide concepts.

(5) Field's Principle: The relation of *same–narrow–content* includes both a relation that holds between people and a relation that holds in the same person at different times.

The argument has been illustrated by a story of cousins in Putnam's Ruritania. The story aims to show that the learning process that takes place for Bruce and Walter —bilingual speakers from two parts of Ruritania— between the age of 10 and the age of 12 involves a change in the narrow content of 'grug' (In Ruritania–B 'grug' is in used instead of 'beer'. In Ruritania–W 'grug' is used instead of 'whiskey'. Other than that, the language in Ruritania is just like English). The story is long and complex and I would use up my space as a commentator if I gave you the full version. Take this as a simplified presentation. For more details, just follow steps I to V in Block's paper. Let me just focus on the conclusion. At first blush, if Block's argument is sound, we are left facing the following dilemma: either we learn to live with a holistic notion of narrow content or (given the disastrous consequences of holism) we ought to reject narrow content altogether.

That was, in fact, the final conclusion of an earlier version of the paper. But, alas, conclusions, like narrow contents, are always changing as a result of gathering new information, and the final conclusion of the new version of the argument is a much more complex disjunction. To wit: either, if there is *one kind* of narrow content, it is holistic or, if there is *more than one kind* of narrow content, then those narrow contents are holistic. This kind of holism, however, doesn't matter, as, at this point, all common sense intuitions that make holism a *horrible* consequence are lost. The reader is led to believe then that common sense intuitions still apply if there is only one kind of narrow content and therefore that holism does matter. So, in a way, the first dilemma still has its full strength.

Assumption 5 —a principle that Field rejects and Block defends— turns out to be the key premise in the switch to this version of the final conclusion. Block's contention is that interpersonal and intrapersonal content identity relations are the same, and he uses that transfer of *sameness* (and difference) in arguing for the changes in narrow content that lead to holism. Those who don't accept the principle are left with the alternative of multiple narrow contents and therefore with no intuitions. Block also points out that the question about the *sameness* (or, again, difference) of narrow content identity relations is *really* the question of "whether there is *more* than one

kind of narrow content". I wonder, however, why *that* is taken to be the key question instead of, e.g., whether there is *less* than one *kind* of narrow content, namely, whether there is no such a thing (which will take us back to the first dilemma). I'll try to show why I think that's the relevant query and also why I think the answer ought to be that there is no such thing.

2 Farewell Narrow Content

Let's recall first of all the assumption about explanation (3): Narrow content has to serve the purposes of psychological explanation, i.e., narrow content —whatever it is— plays a (causal) explanatory role in the agent's behavior. Now, what I want to argue is, firstly, that the viability of the full–fledged holistic option presented in Block's disjunctive conclusion is seriously affected by this reasonable assumption. So seriously affected that the only alternative seems to be to give up on narrow content altogether. I will argue, secondly, that there may be a way of cashing an alternative notion of content, neither narrow nor wide, that is appropriate for psychological explanation and that doesn't have holism as a consequence. The way to do so, I will suggest, is to locate this notion of content in a new, more *ecological* conceptual frame. If I am right, and this single notion of content is plausible, we won't have to accept the disjunctive conclusion of Block's argument.

In order to establish my first contention, it would be very useful to remember the middle rock in the jump from the changes in narrow content to its holistic nature. This middle rock consists in showing that if narrow content exists, it is a very unstable entity. In fact the picture that the cousins story paints for us involves the following indirect route: If there is such a thing as narrow content, then it changes an awful lot, therefore it is holistic. Now, what does it mean to say that narrow content *changes*? As we are not allowed to invoke any theory about what narrow content actually is, the only way in which we can make sense of that question is by relying on the assumptions displayed earlier. When Block says that the narrow content of 'grug' has changed when Bruce and Walter reach the age of 12, what he must be saying is that —whatever it is that has changed:

(1) The non–intentional, syntactical properties of Bruce's and Walter's brains have changed.

(2) The set of beliefs associated with 'grug' are now *substantially* different.

(3) The role that the narrow content of 'grug' plays in the explanation of their behaviors is also different.

(4) The wide concept of 'grug' is different in Bruce's and Walter's communities.

(5) Not only is Bruce's narrow content of 'grug' at twelve different from Walter's narrow content of 'grug' at twelve, but also both Bruce's and Walter's narrow contents at twelve are different from their narrow contents at ten.

I think it is perfectly right to say that changes in syntactical properties cut no philosophical ice here. As Block himself has argued, if changes in narrow content came down to nothing else but syntactic changes, narrow content wouldn't fulfill its main explanatory task: "The problem is that syntactic narrow contents are far too coarse–grained to serve psychological explanation. Syntactically identical objects can play very different functional roles, and be associated with very different recognitional capacities".[2]

So, what seems to be doing the job of justifying the holistic character of narrow content are the changes involved in 2, 3, 4 and 5. More specifically, if narrow content matters mainly, if not exclusively, to predicting and explaining behavior, then it is clear that the kind of change relevant here is change that affects the role that the set of beliefs associated with 'grug' plays in the explanation of Bruce's and Walter's behavior. But we have to be careful here. One of Block's main points about the difference principle underlying (2) is to argue —in order to avoid the small distance objection— that differences in intrapersonal synchronic narrow content can be motivated without appealing to a difference in beliefs. What Block claims is actually that he "could have motivated the difference in narrow content between the two 'grug's within each twin at age 12 by appealing just to the fact that each twin thought the two words picked out different things" (See Block, this volume). However I don't see how this claim can be sustained. If each twin thought that the two words picked out different things, isn't that a belief that has changed regarding what each twin thought each 'grug' picked out in the past? I am strongly inclined to say so. And hence I am strongly inclined to say that a difference in narrow content is always motivated by (in part) a change in beliefs.

---

[2]Block, N., "What Narrow Content is Not" in B. Loewer & G. Rey (eds.), *Meaning and Mind. Fodor and His Critics*, Basil Blackwell, 1991, p. 39.

If that is the case, as any process of learning involves one way or another a change in the set of beliefs associated with *what* we are learning about, we would have to conclude that there is no way of establishing a proper, i.e., general causal explanation of the kind a scientific psychology requires. Not only will no two people share the same narrow content, but also, within a single person, the narrow content of any particular term will vary as soon as any of her belief–systems involving that term alters, which is all too soon, as far as generalizable psychological explanation is concerned.

Block's argument then, by showing that narrow content is holistic in this sense, has also shown something very undesirable, namely that the basic assumption about the psychological explanatory relevance of narrow content is false. If the main task narrow content was here to perform was to account for the psychological explanatory role that representations play in the agent's behavior, but its holistic nature rules out the successful fulfillment of that function, then either we have to abandon the idea of that explanatory role (i.e., we have to be eliminativists) or we have to find an alternative notion of content that can successfully play that role.

The rearrangement I propose is prompted by the basic idea that any notion of content whose main function is to serve psychological explanation of behavior *can't* be thought of as something to be ascribed to an agent regardless of her being embodied and embedded in a particular environment. The idea is not new. Herbert Simon's famous parable about the ant's path on the beach taught us the lesson that a proper explanation of the ant's or any system's behavior would be doomed if we abstract from the features of the current environment where that behavior has been displayed. What is new —as John Haugeland has pointed out[3]— is the conclusion to be reached from here, namely the claim that if we want to understand an agent's behavior, we would have to regard the agent's internal representations and the agent's environment as an integrated explanatory unit. As he puts it, we would have to regard Mind as "not incidentally but *intimately* embodied and *intimately* embedded in its world".[4] That's the basic constraint I want to use.

The main reason for this constraint lies in the fact that psychological explanations invoke mental states with particular intentional contents in order to justify a given course of action, and that those

---

[3]Cfr. J. Haugeland, "Mind Embodied and Embedded". Proceedings of *Mind and Cognition: An International Symposium.* Institute of European and American Studies, Academia Sinica, Taipei, Taiwan, May, 28-30, 1993, p. 5.

[4]*Ibid.*, p. 34.

intentional contents are individuated by externalist properties. The individuation of a mental state as an allegedly explanatory state with such–and–such a particular content can only be achieved by appealing to properties of external objects and events that can account for a particular behavior. But, if this is so, then those content–involving descriptions can't be of the internalist kind; they can't be narrow content–involving descriptions.

If narrow content is to play its main explanatory role, then we should characterize it in such a way that it couldn't be individuated and/or ascribed to an agent regardless of their being embodied and embedded in a particular environment. But then, of course, it won't be *narrow* content any more, at least in the individualist sense of *narrow* that is at issue here. That's exactly my proposal. I want to defend a notion of content that serves the purposes of psychological explanation in a way that is consistent with both the general supervenience constraint expressed above and an embodied–embedded account of cognition. I call this notion ecological content. The new version of the explanation assumption (3) that I want to defend is then:

(3′) ecologically embodied and embedded content has to serve the purposes of psychological explanation, i.e., ecological content plays a (causal) explanatory role in the agent's behavior.

This new line lets us reconcile the idea of a change in content with the idea of a change in its explanatory relevance. As we saw, when I was trying to find out what "change of narrow content" meant, the differences in the non–relational properties didn't seem to be relevant. Now we can see why. They didn't seem to be relevant because we were thinking all the time of those non–intentional properties just as syntactical micro–properties of Bruce's and Walter's brains. If instead of so doing we think of these non–intentional properties both in terms of the inner micro–properties of an agent's brain and the outer macro–properties of the domain she has to negotiate, then the changes in content and its explanatory psychological role move together. And that is as it should be.

I am aware that this kind of reply is in need of further development (See Toribio, "Ecological Content", submitted to *Mind and Language*). As it stands I foresee two main objections. Someone might argue, firstly, that in giving up the micro–property assumption, I'm just swapping narrow content for wide content, because I force myself to specify that alternative notion of content by means of what's been called a wide conception proposal. In that sense, I would be merely repeating some version of a Tyler Burge style of

argument. The other objection is that my ecological content would succumb to the same holistic considerations as narrow content.

First things first; is my ecological content a version of the good, old fashioned, truth–conditional notion of wide content? It is not. The reason is simple. I'm not trying to depict my ecological notion "as sets of ordered pairs of context and truth conditions".[5] I'm not trying to defend that conception because the realm of the embodied and embedded content I'm defending is not the objective realm in which the semantic notions of truth or reference have their place. The notion of content that I am defending is tied to external conditions, but does not allow us to differentiate between environmental differences the subject cannot detect (or rather, that the subject cannot detect without a good deal of scientific or specialized knowledge. E.g., H_2O/XYZ; real dime/fake dime, etc). That notion (a) is not narrowly individuated, because that leaves out the essential environmental component and (b) is not widely individuated either, because that allows the environmental component to float free of the agent's informational embedding in the world.

Ecological content can be characterized in terms of environment–involving computational descriptions of a system's internal states: descriptions which are partially constituted by the behaviorally relevant macro–properties of the domain it has to negotiate. As such, it is not purely referential or wide, because those macro–properties don't necessarily coincide with the properties that are relevant for the individuation of truth–conditional content. Ecological content is not narrowly individuated either, because while the internal states of the system are computationally characterized (and therefore the supervenience constraint is kept in place, i.e., the supervenience base of content is purely *internal*), those computational states are *externally* characterized.

As regards the second objection, whether or not this modified notion of content still leads us to holism, I have a *conservative* answer that involves an actual and counterfactual account of behavior. What a holistic view of ecological content amounts to is the rejection of the idea that we may ascribe an individual content to each belief of a given set in such a way that it would justify the different forms in which that belief can be established as appropriate or inappropriate in a given environment.

Now, even if ecological content changes through the learning process to which Bruce and Walter have been exposed, we still have the

---

[5]Block, N., "What Narrow Content is Not" in B. Loewer & G. Rey (eds.), *Meaning and Mind. Fodor and His Critics*, Basil Blackwell, 1991, p. 50.

conceptual resources to individuate those contents. The network of abilities developed by Bruce and Walter in the specific and different environments in which they are embedded constitutes those conceptual resources which will justify the ascription of individual contents to each belief in their belief boxes. This netword of abilities may change as their relationship with the environment changes, but that change doesn't turn the ecological content of their beliefs into holistic content. What we have is just a *conservative extension* of the former system of beliefs that would let Bruce and Walter act appropriately whenever they have to respond to any unfamiliar property of the domain, whenever, e.g., they travel to the other part of the country. In this way we can locate psychological explanation in the complex of cognitive activities, both internally and externally constituted, that represent its appropriate realm. With this appeal to ecology in hand, we can stop Block's drift to holism and establish an adequate notion of content for the purposes of psychological explanation.

PHILOSOPHICAL ISSUES, 6
Content, 1995

Can There Be a Rationally Compelling Argument for Anti-realism about Ordinary ("Folk") Psychology?

Crispin Wright

One way of showing that there can be no such argument would be to provide, to the contrary, some conclusive case for realism. Another would be to show that certain areas of discourse, including ordinary psychology, are somehow off–limits for realist/anti–realist debate — that the conditions for a valid such debate are somehow abrogated when the subject matter is intentional psychology. My particular concern here is to explore a case for a distinct but no less intriguing possibility: that it may actually be a consequence of (the best version of) anti–realism about ordinary psychology that it should admit of no rationally compelling support. That, I shall suggest, may well be the situation; and if it is not, it is not clear how the consequence can be avoided save by a much more radical and sweeping view than the psychological anti–realist is likely to want or to have bargained for.

1

I shall try to get by with only the lightest theoretical attention to what may be involved in 'realism' and 'anti–realism' respectively.

For present purposes, a realist about a given region of discourse will be one who holds three things:

(i) that its ingredient statements have a content which fits them for the representation of real states of affairs;

(ii) that the characteristic aim of those who practise the discourse is successful such representation, and

(iii) that the world is furnished to provide states of affairs of the kind which such statements are apt to represent.

Each of these three claims is distinctively denied by a well–known anti–realist paradigm. *Expressivists* and *instrumentalists* characteristically deny that a targeted discourse deals in representational contents. *Fictionalists* distinctively deny that the characteristic aim of competent practitioners of the discourse is the representation of real states of affairs. *Error–theorists* distinctively deny that the objects, or properties, characteristically dealt with in the discourse are real, and hence that the world contains states of affairs of the appropriate kind.

Anti–realists about intentional psychology have typically directly denied the third realist component —the worldly reality of the states of affairs which ordinary psychological discourse seems to call for. Since it is doubtful whether someone who repudiated *only* the second component would properly be described as anti–realist, and since any successful attack on the first component: the representationality of psychological discourse —would enjoin rejection of any appropriate category of corresponding states of affairs,[1] we may take it that anti–realists of whatever stripe must converge on such a denial.[2]

The denial can seem like an affront to the merest common–sense. It seems to fly in the face of the characteristic *evidence* of intentional states —the fact that a subject's being in such a state is, as it seems, in typical cases effortlessly and non–inferentially available to them. Surely each of us does have —really have— beliefs, desires, hopes, intentions, wishes, and so on. How else, save by the self–ascription of such states, are we to make sense of most of what we do? How else, save by assuming certain such states, are we ever to decide rationally what to do?

This protest is open to the retort that the "evidence" of ordinary psychological states to their subjects really comes to no more than

---

[1]The considerations about indeterminacy of radical interpretation and Cognitive Command to be reviewed shortly are one such attack.

[2]Though this will need qualification —see note 15 below.

the phenomenon of *avowal*: that people acquire the propensity, on being educated in ordinary psychological practice, spontaneously to affirm claims concerning their own intentional states which for the most part, both to themselves and others, seem —by the standards of the practice— to make decent sense of their behaviour. It is, so it may be contended, quite another matter whether anything real answers to these claims, particularly in the light of certain well–known challenges to the realist view.

2

I'll briefly rehearse three such challenges. First, there is a challenge from considerations of *Cognitive Command* and the indeterminacy of radical interpretation. The idea of the realist about a given region of discourse —unless pessimistic enough to think that what is true there is altogether beyond our ken— is that soberly and responsibly to practise in that region is to enter into a kind of representational mode of cognitive function, comparable in relevant respects to taking a photograph or making a wax impression of a key. The realist conceives that certain matters stand thus and so independently of us and our practice —matters comparable to the photographed scene and the contours of the key. We then engage in a certain process, viz. we put ourselves at the mercy, so to speak, of the standards of belief–formation and appraisal appropriate to the discourse in question —compare taking the snapshot or impressing the key on the wax— and the result is to leave an imprint on our minds which, in the best case, appropriately matches the independently standing state of affairs. Philosophers such as the early Wittgenstein and J.L. Austin tried to be very definite about this type of conception —probably too definite. But even left vague, it does have certain quite definite obligations. If we take photographs of the *same* scene which somehow turn out to represent it in incompatible ways, there has to have been some kind of shortcoming in the function of one of the cameras, or in the way it was used. If the wax impressions we take of a single key turn out to be of such a shape that no one key can fit them both, then again there has to have been some fault in the way one of us went about it, or in the materials used. The price you pay for taking the idea of representation in the serious way the realist wants to take it is that when subjects' representations prove to conflict, then —(prescinding from certain necessary qualifications, mainly to do with vagueness, which I won't elaborate now),— there has to have been something amiss with the way they

were arrived at or with their vehicle —the wax, the camera, or the thinker.[3]

It follows that one obligation of the realist about intentional psychology will be to hold, and therefore to justify holding, that disagreements about a subject's intentional states, since they involve a clash between what purport to be substantial representations, have to involve defects of process or materials, as it were; —that at least one of the parties to the disagreement has to be guilty of a deficiency in the way he arrives at his view, or to be somehow constitutionally unfit. Contraposing therefore: any suggestion that such disagreement can be *rationally blameless* is a suggestion that the realist —seriously representational— view of intentional psychological discourse is in error. But the well–known thesis of the indeterminacy of radical interpretation suggests exactly that. The claim of the thesis is that, such is the methodology of the discipline, radical interpreters of a given subject's saying and doings who proceed unimpeachably can nonetheless wind up with mutually inconsistent yet unimprovable conceptions of that subject's overall psychological set. If nothing in the methodology of radical interpretation constrains its products to within uniqueness, then it would appear to follow that forming opinions in a manner constrained by that methodology is not, in the sense the realist contends, a substantially representational mode of cognitive function. At least: we must either grant that conclusion or concede that the truth values of such opinions may transcend decision by the methods of radical interpretation, (so that the price of realism becomes what Quine famously stigmatised as the Myth of the Museum.[4])

A second, very familiar challenge derives from a worry about causal over–determination. According to our ordinary way of thinking, the beliefs, desires, and other intentional states of a subject which combine to explain certain aspects of her behaviour contrive to *produce* that behaviour —so that the species of explanation involved is causal. Of course there is a tradition —associated, erroneously as I believe, with the later Wittgenstein— which denies this. But the great difficulty with that view, as Davidson and others have emphasised, is that it seems powerless to explain our intuitions in cases where, although a subject does possess certain beliefs and desires which could make perfectly good rationalistic sense of particular elements in her

---

[3]For fuller specification and discussion of the notion of Cognitive Command, see chs. 3 and 4 of my *Truth and Objectivity*, Cambridge, Mass., Harvard, 1982.

[4]See pp. 27-8 of W.V.O. Quine, *Ontological Relativity and Other Essays*, New York and London, Columbia U.P., 1969.

behaviour, —(she wants, for instance, to kill her husband, and believes that she would have a good chance of doing so by toppling the ladder on which he is standing)— it would be wrong to cite them in the explanation of what she actually does —(if, for instance, she leans on the ladder thoughtlessly in order to remove a painful shoe.) However ordinary thought also has it that any *physical* event or process must admit of a complete explanation in terms of its physical causes if it has causes at all. This is uncomfortable if behaviour is viewed as ultimately physical (and how else?). For then it seems to have too many causes: any purposive action of mine is apparently caused not just by the intentional states which explain it, but by parallel neural and other bodily happenings.

There are various familiar strategies of reconciliation. One is to try to stabilise the view that ordinary psychological explanations are not really causal, notwithstanding the sort of difficulty noted. Another, it hardly needs stating, would be to go for some form of physicalism —to deny the distinctness of the intentional psychological antecedents of a piece of behaviour from all its neural and physical causes.[5] A third —unusual and interesting— response would be to deny the identity of the *explananda* in the two kinds of case —to insist on a distinction, for the purposes of explanation, between *action*, strictly so regarded, which is the province of intentional psychology, and *tokens of behaviour* whose explanation need not be psychological at all. But none of these lines, many would feel, has yet been presented in fully convincing detail. Each is open to plausible extant criticisms. So in the present state of play, it seems that someone might quite reasonably feel pressured to respond to the problem by disputing the reality of the explanations offered by intentional psychology, —and hence, since it is of the essence of intentional states to be explanatory of action, the reality of the states that such explanations purportedly depict.

Finally, there is a challenge from content anti–realism. The putative states of affairs in which intentional psychological explanation trades are individuated by the joint specification of a type of attitude —belief, desire, hope, etc.— and a content [that p], the explanatory potential of such states varying as a function of each ingredient in the pairing. It follows that if a general anti–realism about content is correct, —if the world contains no real semantic properties,— then

---

[5]I take no stand here on the question whether this second strategy might be successfully be implemented non–reductively, by play with (some notion of the) *supervenience* of mental states upon physical ones rather than with any claim of strict identity.

a complete inventory of items in the world, and of the characteristics which they can possess, will contain no mention of the states characteristically featured in ordinary psychological explanations.

Arguments for the unreality of content are very familiar in contemporary philosophy. Kripke's sceptical argument, advanced in Wittgenstein's name, and Quine's arguments for the indeterminacy of translation are among the more interesting. Putnam's model–theoretic arguments against realism have been held to lend themselves to such an interpretation[6] (contrary to Putnam's own belief that their effectiveness is restricted to use against a metaphysical–realist antagonist.) Each of these lines of argument is, familiarly, criticisable in detail, and has indeed been roundly criticised. But the underlying worry does not go away that what they, in effect, exploit are various defective conceptions of what contents and content–properties might be in a natural world; and that we actually have to hand no better conception of that than those exploited by such arguments.

3

To accept that any of these challenges is successful is to accept that ordinary psychological ascriptions[7] serve to represent no real subject matter. That would seem to leave open a choice between, on the one hand, viewing ordinary psychological discourse as hopelessly compromised —the well–known eliminativist response— and, on the other, sticking to it that the discourse as a whole is acceptable, even while conceding that it serves to represent no real matters of fact, by finding for it some other heuristic or instrumental role.

The eliminativist response is deeply unattractive.[8] Indeed, in contrast with cases like astrology, or phlogiston theory, where such a reaction is surely correct, it is hard to feel convinced that there really *is* an intelligible eliminativist option in the present case. To take

---

[6]The principal sources are Saul Kripke, *Wittgenstein on Rules and Private language*, Oxford, Basil Blackwell, 1982; ch. 2 of W.V.O Quine, *Word and Object*, Cambridge, Mass., MIT, 1960; and ch. 3 of Hilary Putnam, *Reason, Truth and History*, Cambridge, Cambridge U.P., 1981. The latter is in many ways a more satisfactory argument than that of Putnam's famous "Models and Reality", *Journal of Symbolic Logic*, 45 (1980), pp. 464-82.

[7]—or at least those in which negation is not the principal operator.

[8]A useful many–handed catalogue of complaints, together with some important qualifications, is provided in the special number of *Mind and Language*, 8, vol. 2 (1993).

it that ordinary psychology is merely a superstition would presumably be a commitment to dispensing not just with all examples of ordinary practical–syllogistic reasoning but also with anything like our ordinary concepts of rationality and cognition, which presuppose the authenticity of content–bearing states and processes. Their elimination would thus threaten to leave us without the resources to make sense of any kind of behaviour which seems to call for explanation in, broadly, information–processing terms. Not just ordinary psychology but considerable sweeps of cognitive psychology would be up for elimination too.[9]

However, although some philosophers have canvassed, and a few have even urged, this profoundly opaque prospect, I think it's fair to say that most of those who have been drawn to an anti–realist thoughts about ordinary psychology have had in mind a different form of anti–realist response: a conservative response, broadly comparable to expressivism in ethics and instrumentalism in the philosophy of science. On such a view, it is not necessary, in order for it to be legitimate to speak of content and of states individuated by content, that there be real states of affairs involving content and content properties; it is enough that such talk is appropriately disciplined, and that it serves some legitimate purpose. There are other things for indicative discourses to do besides state facts.

Many will find this direction unattractive too, though not perhaps as unattractive as eliminativism. What sort of purpose might intentional psychology, divorced of any claim to reality, really serve? One well known kind of instrumentalist conception, pioneered in the writings of Dennett,[10] is that the rationalisation of others' behaviour within the familiar ordinary psychological categories can prove an economical way of anticipating it —that it is much easier to predict the moves of a good chess–playing computer, for instance, if I think of it as an intentional strategist rather than merely as a physical mechanism. Now, it seems fair to object that the model implied by this proposal of the explanatory content and utility of ordinary psychological theorising seems somewhat off–beam: ordinary psychological claims do not generally contribute in any spectacularly successful way to *prediction*, seeming to bestow a more characteristically retroactive and interpretative kind of understanding. It is

---

[9]The point is forcefully made in Barbara Hannan's lead–off contribution to the *Mind and Language* special, cited in note 8.

[10]See especially ch. 1, "Intentional Systems", of Daniel C. Dennett, *Brainstorms*, Montgomery, Vermont, Bradford, 1978 and his *The Intentional Stance*, Cambridge, Mass., MIT/Bradford, 1987.

moreover difficult to understand how the impression of *explanation* which ordinary psychological accounts of people's behaviour contrive to give could survive taking quite seriously the thought that such accounts describe nothing real. However the most awkward aspect of any broadly Dennettian view emerges, as it seems to me, when one puts on one side the other–directed uses of ordinary psychology —by which Dennett was preoccupied— and focuses instead on one's own case. It is not just *difficult* to think of the most ingrained elements of one's own self–conception as accepted merely as the components in a self–directed "stance"; it is not clear that it is even *coherent* to do so. For is not such a stance itself individuated by its content —by the attitudes one ascribes to oneself? And does not the Dennettian take it as a *matter of real fact* that one is taking such a stance? If not, what? —a second order stance? It is one thing to take a broadly instrumentalist view of a particular type of theory; quite another to be implicitly told that one must also take an instrumentalist view of the taking of the instrumentalist view. Self–consciously to deploy a complex of supposed fictions in the Dennettian manner is to engage in a complex attitudinal state which there is then no remaining room to construe as fictional or merely instrumental.

Another very familiar anti–realist paradigm —that provided by the expressivism in ethics championed by such writers as A.J. Ayer and R.M. Hare[11]— also teeters into incoherence when applied in the present instance. It is of the essence of any such view to rely on a robust distinction between genuine assertions and other forms of speech act. But any such distinction must ultimately be explained by reference to certain *characteristic intentional states* of participants in the discourse in question —that is why expressivists have

---

[11]The loci classici are of course the famous "Critique of Ethics and Theology" offered in Chapter six of Ayer's *Language, Truth and Logic* (London: Victor Gollancz, 1936); and R.M. Hare's *The Language of Morals* (Oxford: Oxford University Press 1952). Ayer and Hare proposed the strict expressivist view that moral discourse, properly understood, is only apparently assertoric, and that moral utterances are characteristically governed by a different kind of illocutionary force, serving to fit them for a quite different role than the statement of fact —the expression of attitude, endorsement of norms, or whatever. The strict expressivist line is softened in the more recent treatments of Simon Blackburn and Allan Gibbard. Chapter 6, "Evaluations, Projections and Quasi–realism", of Blackburn's *Spreading the Word* (Oxford: Oxford University Press 1984) remains the best introduction to his view; Alan Gibbard's ideas are developed systematically in his *Wise Choices, Apt Feelings* (Cambridge, Mass,: Harvard University Press, 1990). It may be wondered, however, whether their proposals would not to better to travel under the banner of 'minimalism' in the sense of Truth and Objectivity, or 'non–factualism' in the sense of section IV to follow.

thought they could excuse ethical pronouncements any genuinely assertoric role on the ground that they are characteristically aimed at the expression not of beliefs but of certain distinctive feelings and at shaping the corresponding feelings of others. Genuinely assertoric discourses, that is to say, will be marked off from merely expressive discourses by systematic differences in pragmatics which will simply not be stateable without recourse to the categories of intentional psychology. So the very statement of the detail of such a view will demand the possibility of *assertions* about matters in that province.

4

There is an urgent need, then, to consider what —if not eliminativism, nor Dennettian instrumentalism, nor some analogue of ethical expressivism— is the happiest form for an anti–realism about ordinary psychology to assume. This is a big question. But the key move, I believe, is to realise that the *truth–aptitude* of a discourse, and indeed the *truth* of very many of its characteristic assertions, do not have to be in dispute between views about it that are quite justifiably regarded as realist and anti–realist respectively. We need the resource of a conception of truth which allows us to grant truth–aptitude, and indeed truth, to responsible judgements within a given discourse without thereby conceding a realist view of it.

Such a view will hold that to ascribe truth to a statement need not be to ascribe a property of controversial metaphysical content, that any sentence is a candidate for truth which is possessed of assertoric content, and that possession of assertoric content is essentially a matter of meeting certain syntactic and disciplinary constraints —essentially, sentences have such content which are capable of significant embedding within constructions such as negation, the conditional, and in contexts of propositional attitude, and whose use is subject to acknowledged standards of warrant. When such standards are satisfied that will then suffice, other things being equal, defeasibly to justify the claim that the sentence in question is true. The crucial question thus becomes not whether ordinary psychology deals in truth–apt claims, nor whether those of its claims which are justified in the light of its proper standards may defensibly (if defeasibly) be regarded as true, but rather what *kind* of truth its statements are fitted for. And the claim of the anti–realist will be that they are *not* fitted for the kind of robust, non–deflationary truth aspired to by the realist: that the discourse of intentional psychology does not

deal in contents which are apt for the representation of aspects of objective reality.

Work is needed, of course, to explain what such a claim really comes to: to explain the form an anti–realist conception of truth may assume in detail, and how it may deserve the tag, "anti–realist". We need to say what qualifies something to be a truth predicate, and to explain by having, or lacking, which features such a predicate may qualify as a vehicle for realist or anti–realist intuitions. I have entered into these matters in some detail elsewhere[12] and will not enlarge upon them further here, except to say that matters are complicated by the fact that, as I believe, the standard deflationary conception of truth, according to which "true" is merely a device of 'disquotation', serving to express no real property, is of no help in this context; and that there is no simple crux between realism and anti–realism but a variety of features whose possession by a discourse may give some point to realist/anti–realist thinking about it.

We envisage, then, a form of anti–realism about ordinary psychology which will grant that its distinctive claims, literally construed, are truth–apt, and that we are justified in taking many of them to be true. What will be denied is that the discourse generally, and its truth predicate in particular, exemplify any further features which give point to the kind of imagery characteristic of realism: par excellence, the conception of a correspondence between ordinary psychological claims and any aspects of the real objective world.

The acceptance of such a distinction still leaves a space for a kind of error–theoretic anti–realism about psychology: a view which regards psychological statements as (abortively) representational of certain purportedly real states of affairs. There is space, in other words, for a view which holds that psychological discourse *semantically aspires* to realist truth, as it were, but —whether or not speakers are characteristically deluded on the point— systematically fails to secure it.

Again, I think we should foreclose on this option. The reason is simply that ordinary psychology must in any case incorporate acceptability conditions —standards of appropriateness and inappropriateness for its claims— which have nothing to do with truth and falsity as interpreted by realism That is, among the welter of strictly

---

[12]They supply virtually the whole agenda for *Truth and Objectivity*. For an anticipation of the directions there taken, see my "Realism, Anti–realism, Irrealism, Quasi–realism" in French, Uehling and Wettstein, eds., *Midwest Studies in Philosophy*, vol. 12 (1988), Minneapolis, University of Minnesota Press, pp. 25-49; and "Anti–Realism: The Contemporary debate – W(h)ither Now?" in Haldane and Wright, eds. *Reality, Representation and Projection*, New York, Oxford University Press 1993, pp. 63-84.

false claims which, on the sort of vew mooted, ordinary psychological discourse involves, there will be a distinction between those which are nevertheless appropriate —acceptable, in the light of the ordinary purposes of the discourse,— and those which are not. Given that such a subsidiary set of standards has to be recognised by any account, there then seems no evident point to the play with global falsity, realistically conceived. Better, rather, to construe the practice of the discourse purely in terms of the supposedly subsidiary norms.[13]

Let us identify *non–factualism* about a discourse with the view that there is, when it is correctly conceived, no realist semantic aspiration —that its ingredient claims are merely minimally truth–apt, as I have elsewhere expressed it,[14] and involve no claim, even in a fictionalist spirit, to the representation of real states of affairs.[15] Then the recommendation I am making is that the most competitive

---

[13]I do not mean to assert that there could *never* be justification for an error–theoretic view of a discourse of the kind just outlined. My point is rather that the *default* view should be that just recommended —that the obligation is on the error–theorist to show that any other construal of the content of a discourse's characteristic claims than in terms of a kind of truth–conditions which allow the charge of global falsity, would result in serious distortions of our practice and understanding of it. I do not say that cannot ever be done. It might be accomplished in the present instance, for example, by making good the three claims that our understanding of rational explanation is as causal, that physical effects have only physical causes, and that intentional states are not physical. In that case the received truth–conditions of psychological ascriptions would indeed incorporate metaphysical error, and it would indeed be a distortion of our understanding of ordinary psychology to construe its truth–predicate merely as a construct out of its own internal discipline. But I am sceptical whether such a case can be made, in that way or any other.

[14]In *Truth and Objectivity*.

[15]We now need to make the qualification advertised in note 2 above. Earlier it was suggested that anything worth regarding as a form of psychological anti–realism would involve the denial of the worldly reality of psychological states of affairs. But it is also clear that the identification of truth with 'correspondence to fact' as at one level a platitude, so that (for the anti–realist) to grant that we are justified in taking certain psychological claims to be true is already a commitment to recognising the existence, in some sense, of psychological states of affairs. The necessary qualification, then, has to be that it is *only* in a platitudinous, metaphysically non–committal sense that our anti–realist may countenance psychological states of affairs.

In general, once it is granted that "true" is open to variously more or less 'robust' —i.e., realism–implicating— interpretations, the same will go for its cognates, "fact", "state of affairs", "correspondence to fact", "real", and so on; and it will no longer do to identify anti–realism about a discourse with the range of views converging on the simple denial of the reality of the germane kind of states of affairs. What all such views must deny, rather, is their reality in a

shape for anti–realism about ordinary psychology to assume is that of non–factualism. The anti–realist should allow that psychological discourse is genuinely assertoric and highly disciplined and thereby sustains the introduction over its characteristic claims of a predicate with all the essential features of a truth–predicate. But she should insist that nothing is true of it which licenses the interpretation of this predicate in terms of the imagery of correspondence to external, objective matters, in the fashion characteristic of realism, rather than as a construct out of its own internal discipline.

I have taken space to characterise what I consider to be the best general direction for anti–realism in this area to take because I want it to be clear that it is open to the specific problem I shall now proceed to describe. But actually —though I shall not try to do so on this occasion— it would not be difficult to develop analogues of the problem which would directly engage the other —instrumentalist or eliminativist— forms of psychological anti–realism which we have reviewed, —and which of course have in any case additional difficulties of their own,— since any form of denial of the reality of psychological states of affairs may be expected to have ramifications along the general lines we shall explore.

5

We begin with the following Lemma (1):

> It is not possible consistently to be a non–factualist about intentional states but realist about *linguistic* content —about semantics.

Obviously, no–one who favours any broadly Gricean story about linguistic content, according to which the content of any expression will be a construct from some set of characteristic intentions of those who use it, can have any reservations about this claim. But there are two considerations which, even for one with no sympathy for the Gricean approach, are strongly supportive of it.

First, even if the meanings of expressions are not directly determined by the intentional states of language users, after the Gricean manner, a realist about linguistic meaning who wishes to be a non–factualist about contentful psychological states, will presumably have

---

sense —(not necessarily the same in every case: realism may admit of degrees)— cognate to a realist interpretation of "true".

to suppose that there is not even a *supervenience* of the former on the latter. For the real cannot, presumably, supervene upon the unreal.[16] But there surely is such a supervenience: meanings cannot change without change in the psychology (if only the 'wide' psychology) of language–users.

That's the first thought in outline. And if supervenience is taken to be a *constitutive* relation —if we may simply take it that in every case the subject matter of a supervening discourse is somehow composed out of, or otherwise depends for its existence on that of the discourse supervened upon— then the asserted principle of asymmetry, that the real cannot supervene upon the unreal, may seem incontestable. But that supervenience is thus constitutive is not obvious if the supervenience of one discourse upon another is understood, in the usual intuitive sense, merely to consist in its being (conceptually) necessary that any change in the distribution of truth–values among statements in the former would entrain change in the distribution of truth–values among the statements of the latter.[17] What, on that understanding, is there to be said for the principle of asymmetry? Certainly, in cases which most immediately spring to mind of other asymmetric superveniences in the relevant sense —of the moral upon the natural, for example, or the psychological upon the physical— it is never the supervened–upon discourse, rather than that supervening, which has been prone to provoke the anti–realist instinct. But is there any reason why that has to be the pattern?

I have no space to explore the matter properly here, but I'll venture one suggestion about why there may be a general difficulty in the idea of any *asymmetric* real–on–unreal supervenience —and if the suggestion is right, that will suffice for the present purpose, since the supervenience of the semantic upon the psychological is an asymmetric one. Suppose we are concerned only with *maximal* discourses —discourses capable of representing *all* states of affairs of a certain general kind. (Morals, ordinary psychology, aesthetics, physics, semantics would all count as maximal in this intuitive sense.) Suppose too that we are persuaded that non–factualism cannot coherently be a *global* view.[18] More specifically, suppose it accepted that

---

[16]Here and in what follows, the reader should take "real" and "unreal" as a blanket terms of art denoting whatever status particular forms of realism or anti–realism respectively assign to the states of affairs which the truth of statements in the targeted disourse would require.

[17]Naturally, "truth–value" is not, in this context, to be understood as importing a realist conception of truth.

[18]As argued at pages 769-70 of my "Kripke's Account of the Argument Against Private language", *Journal of Philosophy*, LXXI, 759-778, (1984). For misgivings,

> Any non–factual assertoric discourse must supervene upon some factual discourse.[19]

Then a case —assumed for reductio— where a factual discourse, D_1, asymmetrically supervened upon a non–factual discourse, D_2, would in turn —by the transitivity of supervenience— require the supervenience of D_1 upon some factual discourse, D_3, upon which D_2 supervened. But it is not intelligible how one factual discourse could supervene upon another unless, as a matter of (conceptual) necessity, there was some common range of states of affairs which each could serve to depict.[20] And if that were so, it would be obscure why D_2, qua supervenient on D_3, did not also supervene upon D_1 —contrary to the hypothesis that the original supervenience is asymmetric— unless some of the facts representable in D_3 were not representable in D_1 —contrary to the hypothesis that the discourses concerned are maximal. We may conclude that someone who would wish to hold to the factuality of semantics alongside the non–factuality of psychology, and who acknowledges the evident supervenience of the former on the latter, must make a case *either* that some non–factual discourses supervene on no factual one *or* that semantics is a merely a restriction of some wider factual discourse on which psychology supervenes. Neither option seems promising to me (though of course the latter is just what is involved in the Fodorian programme of 'naturalising'.semantics.)

The second consideration on behalf of Lemma (1) is less intricate and should be less contestable. It is simply the reflection that, whatever the correct account in detail, linguistic meanings cannot exist without *conventions*. And conventions, whatever the proper analysis of the notion, have to be constituted in the beliefs and intentions

---

see Paul Boghossian at p. 525 of "The Rule–Following Considerations", *Mind*, XCVIII (1989), 507-49. Some of the issues to do with the tendency of non–factualism (minimalism) to globalise are treated in detail in ch. 6 of *Truth and Objectivity*.

[19]I don't claim this is obviously correct; rejection of it may provide one escape route from the argument to follow. But it seems very plausible. One motivation would spring from the thought that to count as assertoric at all, a discourse must at least to disciplined to a degree; that change in the distribution of truth–values among the statements in a non–factual discourse has to be a matter of change in which of its statements comply with its proper disciplinary constraints; and that such change is unintelligible —when the constraints remain the same— unless in a context where certain real circumstances have changed.

[20]Otherwise, their respective subject–matters would constitute Humean 'distinct existences', and the suggestion of supervenience would violate the Humean intuition that such existences must be logically independent.

of those who are party to them. So if linguistic meanings were real worldly items, and intentional states were not, then reality would, *per impossibile*, be in part constituted by what was unreal.

That seems compelling. There seems every prospect, accordingly, that Lemma (1) should be sustained: non–factualism about intentional psychology must embrace non–factualism about semantics as well.[21]

6

On consequence of Lemma (1) is that any demonstration of the *incoherence* of non–factualism about linguistic content would represent an actual refutation of (what I have suggested is the best form of) psychological anti–realism Paul Boghossian's "The Status of Content"[22] contains precisely such a purported demonstration: specifically, an argument to show that non–factualism about linguistic content must wind up committed to incompatible claims about the interpretation of the truth predicate. That would be a rather blunter, more conclusive problem than the one I have prefigured!

I have explained elsewhere why, as it seems to me, Boghossian's argument does not perform as advertised and will not recapitulate those discussions now.[23] What will concern us is rather something they brought out, which I shall here elicit in a more direct manner.

We have spoken of the sort of truth to which a realist about a given discourse would have its statements aspire and contrasted it with a more minimal conception of truth, apt for the purpose of the anti–realist, which will be some form of construct from the standards of appropriateness governing a discourse about which a non–factualist view is taken. In order to proceed, it will now be convenient to regulate our terminology, reserving "true" for a substantial, realist notion, and "correct" for the minimal conception. Non–factualism about semantics will thus have it that statements about linguistic content, and all cognate matters, will lack substantial *truth*–conditions. So instances of "S says that P" and "S means that P", for

---

[21] For complementary though, as it seems to me, less conclusive considerations in this direction, see pages 170-3 of Paul Boghossian's "The Status of Content", *The Philosophical Review*, XCIX (1990), pp. 157-84.

[22] See note 21.

[23] See the Appendix to ch. 6 of Truth and Objectivity and my "Eliminative Materialism: Going Concern or Passing Fancy", in the *Mind and Language* volume cited in note 8 above. The present paper is in effect a sequel to the latter discussion and is prefigured in its concluding paragraphs.

example, while governed by conditions of correct and incorrect assertion, will not be apt for truth and falsity in any full–bloodedly realist sense.

Now, to speak of the truth–conditions of a sentence (which has truth–conditions) is, on one view, simply a way of talking about that sentence's content itself; and even if that identification is contested, talk of truth–conditions is certainly "cognate" talk in the relevant sense —there is absolutely no way a non–factualist about semantics could coherently cling to a factualist view of ascriptions of truth– or correctness–conditions. So statements of the form,

> S has the truth–condition that P,

will come within the scope of semantic non–factualism, which will therefore be committed to the following:

> For all S and P: "S has the truth–condition that P" is not truth–conditional

—ascriptions of realist truth–conditions are not up for realist truth and falsity.

Now we observe that if statements of the form "S has the truth–condition that P" do not themselves have truth–conditions, then the corresponding statements of the form, "S is true" cannot be truth–conditional either. One way of seeing the point is as follows. What truth values statements take depends, obviously, on their truth–conditions. Suppose we are Gods who have it in our power to fix everything *real* about the world —to create all genuine facts as it were. If statements' truth–conditions are not part of the real furniture of the world, then we shall not, just by determining everything real, have determined —in any save, possibly, a causal sense— what truth–condition any particular statement has. But then we shall not have fixed statements' truth–values either, since, to repeat, they depend on truth–conditions. Since, by hypothesis, we have fixed everything real, it follows that truth–values are not real characteristics. Statements of the form, "S is true" are accordingly not factual.

Remember that we are reserving "true" for realist interpretation. So in supposing that a range of statements are true —or, for that matter, false— we are saying that they are amenable to realism —that a realist conception of their content and subject matter is appropriate. It follows, accordingly, that the latter claim is *itself* a non–factual one: that the distinction between truth–apt and merely correctness–apt assertoric discourses is one the details of whose extension are not themselves stateable by *truths*, but only permit of *correct* statement; that (Lemma (2))

> Non–factualism about semantics enforces non–factualism about the factual/non–factual distinction itself.

Really, this is rather obvious. For however exactly the distinction is drawn, which side of the factual/non–factual divide a given discourse falls is going to be, in general terms, a function of the *type of content* which its sentences possess. So the non–factuality of claims which place a discourse to one side or the other of the factual/non–factual divide must follow from a general non–factualism about matters to do with content. If, resuming our God–like role, we fail, when determining everything real, to determine —in any save, possibly, a causal sense— any matters to do with content, then we likewise fail to determine anything which functionally depends on matters to do with content —including the detail of the distinction between the factual and the non–factual itself.

More simply: reflect that, with "true" and its cognates restricted to realist interpretation, statements of the form, "S has the truth–condition that P", serve simultaneously both to make a semantic claim and to classify their subjects as factual. So to attempt to hold simultaneously that ascriptions of factuality are themselves factual while ascriptions of content are not, would be —when possession of *truth*–conditions is taken as a hallmark of factuality,— a direct commitment to contradictory claims about the factuality of such statements.

7

Before we can move on, we must confront what may seem to be an evident lacuna: essentially, that to treat the foregoing argument as establishing Lemma (2) is illicitly to generalise a conclusion demonstrated only for the *metalinguistic* classification of discourses. We may grant that claims about the factuality, or non–factuality of discourses, construed as ranges of *sentences*, cannot themselves be factual if claims about linguistic content are not. But surely nothing follows about the factuality of the distinction between the factual and the non–factual when it is drawn at the level of the *contents* —propositions, or Fregean thoughts— themselves. It may be a non–factual question whether S is a non–factual sentence only and precisely because it is a non–factual question what S means. But there simply is no question, factual or otherwise, about what the proposition that P means: P is an entity already individuated as a content, and nothing that has been said bears on the question whether there

is not a factual distinction among such entities between those which are apt to represent real facts and those which are not.

There is a similar objection in the case of Lemma (1): that too only concerns the metalinguistic case: it asserts the commitment of the psychological anti–realist to the non–factuality of *semantics*. No parallel commitment has been disclosed to the non–factuality of the properties and relations of propositions. Neither of the considerations adduced —viz. the supervenience of linguistic content upon psychology, and its dependence upon convention— has any evident transposition to discourse not of the semantics of sentences but of propositional contents themselves.

These objections are technically correct. But they are also, obviously, to no particular purpose unless some form of realism about the contrast between the factual and the non–factual, drawn at the level of propositions, is a serious possibility. One immediate reaction is to wonder how such a realism could avoid what is effectively an (objectionably) *Platonist* conception of propositions and their properties. For if the classification of propositions as factual and non–factual answers directly to real states of affairs, then presumably the entities so classified must themselves be real in a correspondingly robust sense.

It would not do to suggest that the credibility of any view which implied a Platonist conception of propositions is hopelessly compromised just on that account. There has been much respectable philosophy which, beyond being prepared to countenance talk of propositions as useful and legitimate, has involved a kind of realism about them of the sort to which Frege was drawn, whereby propositions are conceived as objective, mind–independent entities and successful communication is viewed as achieved by dint of our shared cognitive relations to them. I myself would foresee no objections of principle to an attempted application here at least of the kind of moderate platonism, reliant on Frege's Context Principle as a principle about reference, which the provision of satisfactory explanations of the meanings of statements about propositions, grounded in discourse of a different, presumed relatively unproblematic sort, might contrive to justify.[24] The question, however, is not primarily whether some form of platonism about propositions —that is, realism about content–entities and their (essential) characteristics— might be acceptable; or even whether some acceptable form could effectively defend against

---

[24]For detailed discussion of this deflationary form of platonism, see my *Frege's Conception of Numbers as Objects*, Aberdeen, Aberdeen University Press 1983; and Bob Hale, *Abstract Objects*, Oxford, Basil Blackwell 1987.

the suggestion that the factual/non–factual distinction, even at the level of propositions, is a non–factual one. Rather it is whether platonism of this kind is an option *in the context of psychological non–factualism.* Certainly the marriage would be an extremely tense one if the motivation for the latter were of either the second —physicalist— or third —content–sceptical— kinds earlier distinguished. But is the combination so much as coherent in any case?

One clear corollary of the combination would be that any psychological state which 'embeds' a proposition, including most basically the state of *grasping* it, cannot consist in a real relation between the human subject and the proposition in question. Since there is no doubt about the reality of the terms, —propositions would be real by hypothesis, and human subjects are presumably no less real for the psychological anti–realist than for anyone else— the source of the unreality must lie in the type of relation concerned. So in particular there are no real facts about subjects' grasping —let alone believing in, or hoping for, or intending to affirm the truth of, etc.,— particular propositions. Since our conception of a proposition is initially arrived at as precisely the conception of something which may be affirmed or grasped or believed, etc., it will come across as extremely odd metaphysics to be told that none of these canonical forms of relationship, as it were, can be reckoned as part of the real furniture of the world, although their object terms —the propositions— are real enough.

Let me try to be more specific about the difficulty which that suggests. Discourse of propositions, like all discourse of abstract entities, raises certain initial epistemological problems which can be assuaged only by an attempt to explain how it is properly understood: to explain, specifically, how it may be so understood that the appraisal and justification of its characteristic statements comes to be within human epistemological compass. Now the use of singular terms —par excellence, that–clauses— which refer to propositions originates in talk about sentence–meanings and the propositional attitudes, and it therefore seems certain that any attempt to address systematically the question of the meaning of proposition–discourse would have to assign a key role to psychological and semantic contexts in its *explanantia.* Yet the present context is one in which —by hypothesis and by Lemma (1) respectively— *both* the latter types of discourse are conceived as non–factual. How could explanations constructed out of such materials possibly provide the means to legitimate the idea that propositions are robustly real entities, or that any contexts concerning them and their properties —in particular contexts serving to classify the factual and the non–factual cases— deal in robustly real states of affairs?

To fix ideas, consider the familiar though controversial form of explanation of statements about certain species of abstract object which I adverted to above: that given by Frege's *Grundlagen* method of *abstraction*[25] whereby contexts of identity for a new kind of object —say, directions— are explained by reference to the holding of some equivalence relation —parallelism— on items of some antecedently understood kind —straight lines— and predications on the former are explained in terms of the possession of certain antecedently understood properties by the latter. In certain cases, the result may be a system of rigorous contextual definitions whereby all overt reference to and quantification over the new kind of objects may be eliminated; in other cases —crucially that of cardinal numbers, which was Frege's primary concern— the most that can be claimed is that the content of reference to and quantification over the new kind of objects is *non–eliminatively* explained by the relevant abstraction and other associated principles.[26] But what is utterly unclear in all cases is how the *explananda* —statements about the new kind of object— could finish up as factual unless their *explanantia* were factual in the first place. How, for instance, could discourse of directions and their properties deserve realist construal unless discourse of lines deserved it already?

Admittedly, if, as I have suggested, it is a grounding in intentional–psychological and semantic discourse which provides the route into the concept of a proposition, then it would seem that the explanation is structurally quite different to that involved in a Fregean abstraction: the concept of proposition will be given by explaining the use of, and then giving a certain cast to the resultant understanding of the explanatory contexts, rather than by exploiting them in the explicit introduction of a novel vocabulary. But the difficulty remains no less impressive: how can entities our initial understanding of reference to which is acquired by a mastery of certain non–factual contexts comprise the subject matter of contexts of a different, factual kind? The suggestion seems merely incoherent unless the explanation of propositions as the objects of the attitudes and the contents of sentences is importantly incomplete. But then, what is

---

[25]Discussed at § 62 and following; see pp. 73-80 of J.L. Austin, tr., Gottlob Frege: *The Foundations of Arithmetic*, Oxford, Basil Blackwell 1950.

[26]This is a consequence of the fact that the abstraction for cardinal numbers mooted at Grundlagen § 63 —often called Hume's Principle— is formulated in terms of a *second order* equivalence relation: an equivalence relation on concepts, whose relata may include numbers among their possible instances and whose expression may itself involve arithmetical vocabulary.

the omitted extra whose inclusion in the explanation might somehow allow propositions to be robust?[27]

I shall not try to take the issue further here. While acknowledging that a defence of the factuality of the distinction between the factual and the non–factual might yet be mounted on the back of a realist conception of propositions, I think that enough has been said to justify scepticism about the prospects when they are constrained by non–factualism about the intentional and the semantic. So while accepting that Lemma (2) has not been proved for the case where the factual/non–factual distinction is applied not to discourses but directly at the level of contents, we may justifiably proceed on the assumption that the ontological materials needed by factualism about the factual/non–factual distinction at that level are denied to the non–factualist about psychology. Granted that Lemma (2) has been made good for the metalinguistic case, the upshot is accordingly that the thesis of psychological anti–realism —in its best form— is a commitment to its own non–factuality, and that any argument for it is consequently an argument for a non–factual conclusion.

8

If this is right, the situation in which our anti–realist finds herself is apt to impress as very uncomfortable. For what possible rationale can realist/anti–realist debates have unless we think of them as answerable to *real* distinctions? Doesn't one have to be a meta–realist, as it were, about the realism/anti–realism distinction —to believe that protagonists in such debates are disputing a factual matter—

---

[27] Someone of reductionist leanings might query with what right I have assumed that any factualist construal of claims concerning the factuality of particular propositions must entrain a realist view of propositions. Might not some reduction of proposition–discourse be effected which both allowed it to be construed as factual and, at the same time, showed that its apparent ontological commitments need not be taken seriously?

This line of thought does not really broach new ground. Such a reduction would —for one in sympathy with the moderate, Context Principle–based form of platonism mooted above— effect not a way of avoiding a commitment to propositions but rather precisely the —presently missing— additional explanation which would justify regarding them realistically. Note moreover that the reduction base could in any case not be provided by discourse of any non–factual kind if it was to deliver the desired result. So the obvious place to try —viz. discourse of sentences and their semantic properties and relations— is excluded in the present context.

before any interest can attach to the question on which half of the distinction ordinary psychology falls?

We can envisage the short answer that it is no more a precondition of the interest of philosophical debate about realism that one take a realist view of the discourse of such debate than it is a precondition of the interest of ethics, or mathematics, or indeed ordinary psychology that one takes a realist view of the discourse of those disciplines. I think that answer is a little *too* short: there is something deeply disorientating about the thought that although there is an intelligible distinction to be drawn between factual and non–factual discourses, there are no real facts about the classification of discourses under the aegis of that distinction. For one thing, non–factualism about a discourse standardly implies some form of unflattering comparison with factual cases —how can that implication be coherent if the comparison itself is similarly disadvantaged? For another, the stability of the non–factualist line about ethics, or mathematics, etc., depends on providing an account of a legitimate role and purpose for such discourses dissociated from the project of representing the world. But what role and purpose might *metaphysics* have if not the attainment of insight and understanding into how things *really* are? —how can those goods be the product of *non–factual* enquiry?

The principal question, however, concerns the status of our leading issue: the possibility of rationally compelling argument for psychological anti–realism.. And a very simple thought is immediately salient: that debates about *non–factual* matters are *eo ipso* never susceptible to rationally compelling resolution —that to characterise a question as non–factual, if it means anything at all, must carry the implication that opinions about it are at bottom rationally unconstrained— that competent interlocutors must be willing, if need be, to agree to differ without the imputation to each other of errors of reasoning or cognition.

I think this simple thought is very likely correct, though I will not argue for it directly now.[28] What is certain is that a commitment to non–factualism about *all* realist/anti–realist taxonomy leaves the psychological anti–realist with a lot of explaining to do if she wants to insist that her views about psychology are rationally mandatory. I shall conclude by outlining one aspect of her dialectical predicament.

The terms: "factual" and "cognitive" are, as Wittgenstein said of "rule" and "same", made for each other: roughly, factual matters are

---

[28]It is right if, as suggested in *Truth and Objectivity*, Cognitive Command represents the first substantial realism–relevant hurdle which minimally truth–apt discourses may fail to clear.

those which may be appraised just by cognitive capacities; cognitive capacities are those which enable the appraisal of factual matters. More precisely, one inclined to take the distinction between factual and non–factual discourses seriously is likely to see it as mirrored in the respective ranges of abilities which competent practitioners of such discourses will draw upon: *cognitive* abilities should be those belonging to a smallest psychological endowment sufficient to enable a competent view on any particular (appraisable) *factual* matter; and any matter should count as *factual* which can be competently determined just by the exercise of *cognitive* abilities.

It follows that rejecting the factuality of a claim is a commitment either to the contention that non–cognitive abilities are essentially involved in its appraisal or to maintaining that the case is one of indeterminacy —that no particular verdict on it is mandated by the appropriate standards. Since the latter option —the claim of indeterminacy— is not available to a *proponent* of the (non–factual) claim that intentional psychology is not factual, her contention will have to be that whatever mandates that view of psychology, it is something beyond appreciation purely by the exercise of cognitive capacities.

Obvious next question: how should the ability to *reason* rate when the question is which matters are factual and which abilities cognitive? Are the judgements of reason —claims of logical consequence, for instance, or of the soundness of a philosophical argument— apt for realist truth and falsity? Should the reason count as a purely cognitive ability?

Assume so. Then a conclusion to our title question seems immediate. For if the reason is a purely cognitive ability, and the mandate, whatever it is, for the claim that intentional psychology is non–factual cannot be appreciated by the exercise purely of cognitive abilities but demands something else, then, whatever sort of cogency it possesses, that mandate cannot be *rationally* compelling, cannot be cogent for a subject who merely acknowledges all relevant facts and reasons correctly.

Strictly, that goes too fast. The conclusion is, so far, only one for the anti–realist to draw —it is conditional on an acceptance that psychology *is* non–factual. What has been argued, that is, is that *if* psychology is non–factual, then all claims about what is or is not non–factual are non–factual and hence —on the assumption that the reason is cognitive— psychological anti–realism is beyond rationally compelling argument. To be sure, if we only add the assumption that whatever is amenable to rationally compelling argument is true, then we can directly advance to the conditional that

> If anti–realism about psychology is capable of rationally compelling argument, then (it is true and hence) is not capable of rationally compelling argument,

and hence to its consequent as an unconditional conclusion. Still, the assumption needs to be made.

Even without that assumption, however, there is surely a Moore–style incoherence in the position. For —in contrast with the situation of comedy, or the revolting, e.g. where we already self–consciously conceive of warrant for an opinion as never purely a rational matter and an intuitive anti–realism is our existing predisposition— philosophical claims of this kind *are warranted a priori, by pure reason, or by nothing at all.* So, on the assumption of the cognitive character of reason, the psychological anti–realist appears constrained to concede both that her position admits of no rationally compelling support and that it is a view which, if warranted at all, can be so only by the adduction of rationally compelling considerations.

The prospects, then, on the assumption of the cognitive character of our powers of reason, look to be murky. What if the assumption is discharged —if the anti–realist is prepared to regard the reason itself as non–cognitive?[29]

Whether such a stance would allow of coherent explanation and defence must be regarded as very moot. But two consequences are worth highlighting immediately. First, the minimum cost would have to be a severance of questions of factuality from those of *objectivity.* For to propose that a question might be rationally decidable and yet not be fully objective would be merely to surrender all grip on the latter concept. So the dialectical situation would be potentially totally transformed. The classification of ordinary psychology as non–factual provokes opposition not just because of certain pejorative overtones: an association with bad company (claims about the comic and the revolting, and so on) and with a certain optionality (as evinced in the Dennettian notion of a "stance"), but because it seems to threaten the objectivity of our self–conception, of our most distinctively human way of thinking about human beings However if statements whose acceptability is normally thought of as being an entirely rational matter, statements which we conceive as fully objective —statements of logic and arithmetic, for instance— get to be fellow travellers, no stigma of this sort can attach to non–factuality

---

[29] As she must in any case do if she is to reject the simple idea, bruited above, that to characterise a question as non–factual, if it means anything at all, must carry the implication that opinions about it are ultimately rationally unconstrained.

per se, and the friend of ordinary psychology, rather than feeling obliged to argue for its factuality, may take comfort in the reflection that a region of discourse could be "non–factual" even though the appraisal of its statements demands nothing but the exercise of cognitive abilities and (non–cognitive —as it is now viewed—) reasoning. Of course, even that might not be true of psychology. But the important debate would no longer be about non–factualism.

Second, it follows from the characterisation of the relations between "cognitive" and "factual" outlined above that no subject matter can count as factual whose appraisal draws essentially on non–cognitive abilities (since the latter are to suffice for the appraisal of any factual matter.) So if our non–cognitive abilities are to embrace not merely the usual suspects —the sense of humour, maybe our aesthetic sensibilities, etc.,— but also the reason itself, then the domain of the non–factual must correspondingly embrace everything which can only be known by exercise of the reason. And that will include not merely all that can be known purely by the exercise of reason —by pure, rational thought— so all of logic and mathematics and —I suppose— philosophy: every claim in whose assessment reason has some indispensable part to play will be dragged in as well.

It has, of course, long been controversial how inclusive a class of claims that should be reckoned to be. Classical epistemological foundationalism would regard the most immediate reports of observation as exceptions. But the now prevalent view would be that —to the extent that background beliefs condition the acceptability of even the most basic observational statements— reason and inference are literally ubiquitous in the appraisal of scientific evidence; and their role in determining the acceptability of theories and hypotheses is, of course, not up for question. If the prevalent view is right, the startling upshot is therefore that to attempt to maintain that non–factualism might yet be open to rationally compelling support by denying the cognitive status of reason would be a commitment to denying the factuality either of absolutely everything or at least of *all of empirical science*. Since —one would imagine— all actual psychological anti–realists have been inclined to regard non–intentional physical science as the place where the hardest real facts are, that would be an irony indeed.[30]

---

[30]I am grateful for their valuable comments and criticisms to the audiences at a Birkbeck College reading party held at Cumberland Lodge in May 1993, at the SOFIA conference in Lisbon in May 1994, and at the Cincinnati conference on Significance in Semantics held in September 1994; also to Paul Boghossian and Bob Hale who generously gave me their detailed reactions to the penultimate draft.

Philosophical Issues, 6
Content, 1995

A Note on Boghossian's Master Argument

Roger F. Gibson

In his paper in this volume, Crispin Wright cites an argument proffered by Paul Boghossian in his "The Status of Content" which puports to establish that the application of non–factualism to content discourse demonstrates the incoherence of non–factualism. Wright is sympathetic to Boghossian's conclusion, but not to his line of argument (see n. 21 of Wright's paper). I, too, reject Boghossian's line of argument, but my alalysis, given below, dffers from Wright's.

On my reading of Boghossian, what might be called his master argument for the incoherence of non–factualism consists of a sort of chain argument. The first link in the chain is intended to secure the claim that non–factualism presupposes robust conceptions of truth and reference. Limiting our attention here to truth, Boghossian's argument, given in Part I of his essay, proceeds, nearly enough, as follows:

P1: All non–factualist conceptions of content consist of the following two claims regarding significant declarative sentences of the form "x is P": (a) the predicate "P" does not denote a property, and (as a result) (b) no such atomic sentence expresses a truth condition.

P2: But, every significant declarative sentence of the form "x is P" has *deflationary* truth conditions (because being significant and declarative are individually necessary and collectively sufficient for having such truth conditions).

C1: Thus, if P1 and P2 are to be consistent with one another, then P1 must be read as asserting that no significant declarative sentence of the form "x is P" has *robust* truth conditions. In other words, the very intelligibility of non–factualism *presupposes robust* truth conditions.

Having established C1, Boghossian concludes Part I of "The Status of Content" with an announcement that he will argue that the application of the non–factualist model to content discourse reveals that the model runs afoul of the aforementioned presupposition, namely, a robust conception of truth (and reference).

In Part II of his essay, Boghossian argues that one ought not be an irrealist about psychological content unless one is also prepared to be an irrealist about all content, including linguistic content. (This is the thesis that Wright calls the Lemma of Content Non–Factualism, LCNF); Boghossian also describes four kinds of considerations leading to content irrealism: indeterminacy of translation (associated with Quine), holism and contextualism (associated with Stich), failure of naturalistic reduction (associated with Schiffer), and "queerness" of content properties (associated with Kripkestein); and he argues for the view that the "core" meaning of "content" is truth conditions.

Thus, it isn't until Part III of his essay that Boghossian again picks up the main thread of his master argument that the application of non–factualism to content discourse runs afoul of its having presupposed a robust conception of truth (and reference). The argument continues, I think, as follows:

P1′: Applied to content discourse, the non–factualist position consists of the following pair of claims: (a) "S has truth condition P" is not truth–conditional; and (b) "true" does not refer to a real property.

P2′: However, according to the argument of Part I, the claim in P1' *that "S has truth condition P" is not truth–conditional* presupposes a robust sense of truth (and reference), but the second claim in P1', *that "true" does not refer to a real property*, is the denial that truth (and reference) is robust.

C1′: Thus, non–factualism applied to content discourse is incoherent: it both asserts and denies that truth is a real property.

As Boghossian goes on to explain, though non–factualism about *any* domain of discourse presupposes a robust conception of truth (and reference), it is only when non–factualism is directed precisely at truth itself that it entails the denial of that very presupposition and is, thus, incoherent.

One conclusion which Boghossian draws from his master argument is that "if there is a genuine issue about the status of content discourse, it cannot be formulated in accordance with our standard irrealist models [i.e., error–theoretic and non–factualist models]".[1] Moreover, though Boghossian acknowledges that various other philosophers are investigating "the possibility that a cognitively interesting contrast may be drawn in non–truth–theoretic terms",[2] he remains skeptical about the prospects of all attempts to reformulate content irrealism in some non–standard way: "I am inclined to believe", Boghossian writes, "... that the correct moral... is just what it appears to be: that we really cannot make sense of the suggestion that our thoughts and utterances do not possess robust truth conditions".[3]

Boghossian's argument in Part I, the argument that purports to demonstrate that the very intelligibility of non–factualism *presupposes* robust truth conditions, is open to different interpretations. The argument does indeed demonstrate that non–factualism *presupposes* something about robust truth conditions, but just what, precisely? Is the presupposition merely that the expression 'robust truth conditions' is *meaningful*? Or is it that the expression 'robust truth conditions' is both meaningful and denotes a *property*? Boghossian takes it to be the latter. That he does is what powers his master argument against non–factualism. However, it seems to me that the first option is open to the non–factualist. That is, she could presuppose that the expression 'robust truth conditions' is meaningful (has a use), without also presupposing that it denotes a property —in the same way that one could presuppose that the expression 'round–square' is meaningful (has a use), but fails to denote a property. And, in the absence of the stronger presupposition, the way is cleared for a non–factualist to maintain that for some significant

---

[1] Paul A. Boghossian, "The Status of Content", *The Philosophical Review*, Vol. XCIX, No. 2 (April 1990), p. 176.

[2] *Ibid.*, p. 183.

[3] *Ibid.*

declarative sentences of the form "x is P", both (a) the predicate "P" does not denote a property, and (as a result) (b) no such atomic sentence expresses a robust truth condition, but now without this latter claim, (b), presupposing that the expression 'robust truth condition' expresses a property. And if this construal of the non–factualist's presupposition were permitted —and why shouldn't it, since meaning is distinct from reference?— then Boghossian's master argument against non–factualism would not go through. It would not because, when Boghossian applies non–factualism to content discourse in Part III of his essay, P1′ would no longer be inconsistent: the claim *that "S has truth condition P" is not truth conditional* would no longer be inconsistent with the claim *that "true" does not refer to a real property*, for now the first claim does not presuppose what the second claim denies.

It seems to be a presumption of Boghossian that intentional psychological discourse is either entirely factual or entirely non–factual. However, it is not clear to me why this must be so; why couldn't some intentional psychological discourse be factual and some non–factual? Or why couldn't there be degrees of factuality? I suppose that one's view on these matters is a function of the line of thought that led one to non–factualism in the first place. And, one purported line, cited by Boghossian, is Quine's thesis of indeterminacy of translation; for example, Boghossian writes:

> It is a famous claim of Quine's that for any mental state or linguistic expression, a pair of content ascriptions can always be devised which would be such that, although they could not both be true, no rational considerations could decide between them. He took this to show that ascriptions of meaning were not a genuinely factual matter.[4]

If this were Quine's line, then I can see how it would lead to a sort of blanket non–factualism about intentions and intensions. But I know of no place in Quine's writings where he makes this "famous claim". Quite the contrary, there are many places in his writings where he disavows it.

Let me conclude by citing one of Quine's most comprehensive disavowals. It is found in his recent but not widely read essay titled "Let Me Accentuate the Positive". Therein Quine writes:

> In ascribing to me the 'claim that there is no "matter of fact" involved in attributions of meaning to utterances, beliefs to people, and aspirations to cultures', Rorty overstates my negativity. How words and

---

[4] *Ibid.*, p. 172.

> sentences are used, in what circumstances and in what relations to one another, is very much a matter of fact, and moreover I cheerfully call its study a study of meaning. My reservations concern rather the ascription of a distinctive meaning or cognitive content to each separate sentence, as something shared by the sentence and its correct translation. I hold that two conflicting manuals of translation can do equal justice to the semantic facts, while distributing the meaning load differently sentence by sentence. The manuals can be counted on to agree over sentences whose affirmation is pretty regularly linked to concurrent sensory stimulation, but they may diverge over others.
>
> When we turn to attributions of beliefs, I see factuality as grading off from case to case. Some beliefs can be ascribed even to dumb animals, in light of behavior. Some beliefs can even be measured, in human subjects, by laying bets and offering odds. But the grammar of the general belief idiom, 'x believes that p', outruns the idiom's factuality. The idiom counts as grammatical no matter what declarative sentence we put for 'p', but for some sentences there is nothing factual about holding the belief: nothing but pious lip–service.
>
> Partly because the grammar of the belief idiom outruns its factuality, the idiom is not acceptable as an idiom of an austere scientific language. It is this exclusion, evidently, that leads Rorty to suppose that I find no matter of fact in attributions of belief. I often do, and I would want to see it conveyed in scientifically more acceptable idioms.[5]

It is pretty clear from this passage that Quine's indeterminacy thesis is not, after all, a line of argument leading to a blanket non–factualism about content discourse. That is not to say, of course, that the other lines that Boghossian associates with Stich, Schiffer, and Kripkestein don't lead to blanket non–factualism about content.

---

[5]W.V. Quine, "Let Me Accentuate the Positive", in *Reading Rorty*, Alan R. Malachowski, ed. (Oxford: Basil Blackwell, 1990), pp. 117-119. Other places where Quine addresses this topic are: *Pursuit of Truth* (Cambridge: Harvard University Press, 1992), paperback, pp. 66-67; *Quiddities* (Cambridge: Harvard University Press, 1987), pp. 18-21; "Reply to Hilary Putnam", in *The Philosophy of W.V. Quine*, L. Hahn and P. Schilpp, eds. (La Salle, Ill.: Open Court Press, 1986), p. 429.

PHILOSOPHICAL ISSUES, 6
Content, 1995

Content, Computation and Externalism*[1]

Christopher Peacocke

This paper starts from the desire to resolve an apparent inconsistency. There is an apparent inconsistency between a widely–held conception of computation, and a plausible view of psychological explanation, when taken together with the goals and actual practice of psychology. I am going to be arguing that the inconsistency can be resolved by recognising and making use of a different conception of computation.

This alternative conception is peculiarly suitable for answering certain kinds of questions about explanation. I will be arguing that this alternative conception of computation is not reducible to the widely–held notion. The alternative conception is also indispensable if we

---

*Reprinted by kind permission of *Mind and Language*.

[1]I have been helped by valuable comments from Martin Davies; from the members of our joint seminar in Oxford in Trinity Term 1993; the audience at a meeting at CREA, Paris; and the referee and editors of *Mind and Language*, in which this paper also appears (volume 9 (1994), number 3). I also owe a special debt to the participants at the 1994 SOFIA meeting in Lisbon, in particular my commentators Daniel Andler and Josefa Toribio, and to discussions there with Ned Block, Paul Boghossian, Tyler Burge, Manuel Garcia Carpinteiro, Jerry Fodor, Paul Horwich, Brian Loar, Jerry Katz, Pierre Jacob and Jaegwon Kim.

are to give a proper elaboration and defence of the whole computational enterprise in psychology. I begin by explaining the apparent inconsistency.

1 The Inconsistency

On one view of computation, computations cannot be sensitive to the semantic properties of representations. The *locus classicus* for a statement of this view is Jerry Fodor's paper, written in 1978, "Methodological Solipsism Considered as a Research Strategy for Cognitive Psychology". "I take it that computational processes are both *symbolic* and *formal*". (Fodor 1991 p. 486). For Fodor, formal operations need not be syntactic —he cites rotation of a mental image— but "What makes syntactic operations a species of formal operations is that being syntactic is a way of *not* being semantic. Formal operations are the ones that are specified without reference to such semantic properties of representations as, for example, truth, reference and meaning" (*ibid.*). "If mental processes are formal, then they have access only to the formal properties of such representations of the environment as the senses provide. Hence, they have no access to the *semantic* properties of such representations, including... the property of being representations *of the environment*". (*ibid.* p. 488). This view of computation is in no way unique to Fodor. It is often tacitly taken for granted in much of the literature. It does not presuppose Fodor's other views on content, nor does it require his views about a language of thought. It would prima facie be entirely consistent (though I would not recommend the combination as correct) to hold simultaneously all these three propositions: that some intentional states are realized in non–sentential states of a connectionist network; that some sequences of states of the network are to be regarded as computations; and that semantic properties are never involved in either the explaining or the explained states of a computational explanation.

The contradiction now emerges. If the non–semantic view of computation were correct, it certainly looks as if there would have to be a massive mismatch between means and ends in much of contemporary psychology. It looks for all the world as if much theorizing in psychology attempts to explain particular intentional, content–involving properties of a subject. We seem to be presented with, for example, explanations of how the intentional property of an experience of representing an object as being a certain shape results from computation from two–dimensional information about its motion; with computational explanations of how a person comes to hear a

sentence as meaning this rather than that; with computational explanations of why a person intentionally chooses this course of action over that; and so forth. Yet offering computational explanations of these intentional properties would just involve a mistake of principle, if the non–semantic view of computation is correct.

On the non–semantic view of computation, a computational explanation of a person's coming to be in an intentional state involves one non–semantic state explaining, by some computational procedure, a second non–semantic state. This second state is said to be the basis of (or realization of, or what constitutes) the intentional state to be explained. But if only non–semantic properties are explained, where is the explanation of the intentional properties? It seems that on the non–semantic conception of computation, only non–semantic features of intentional states could be explained. If the explaining conditions of the computational explanation were content–involving, there might still be some room for manoeuvre here —perhaps the explaining conditions could ensure the right relational properties required for the intentional state to have the content it does. But this too is ruled out by the non–semantic conception of computation, according to which the explaining conditions are also non–semantic. On that conception, the internally individuated explaining conditions cannot magic into existence the complex of non–syntactic relations required for an intentional state to have a certain content.

It is true that non–semantic computational explanations, like all other explanations, will have their presupposed background conditions. But when computation is conceived only non–semantically, it seems that these background conditions will have to do only with such matters as whether processors sensitive to formal properties are operating properly; and the like. Such background conditions cannot resolve the fundamental problem that what gets explained under the non–semantic conception are not the intentional properties of the intentional state. Nor does the puzzlement here rest upon failure to appreciate that computational explanations are genuinely empirical explanations, and not a priori elaborations. It is entirely consistent to insist upon the explanation of content–involving states by other states which are either themselves content–involving, or at least presuppose various non–syntactic conditions, while still acknowledging that these psychological explanations must have an empirical character, and cannot be divined a priori.[2]

---

[2]Though computational explanations of action by intentional states are a less–well developed part of the discipline, one could make corresponding points about that direction of explanation too. It is relational, environment–involving prop-

Still, it may be replied that it is enough for the explanation of the intentional properties of mental states that the computed state has syntactic properties for which there is a corresponding semantics. If we explain the syntactic properties, does not addition of the semantics give us an explanation of the semantic properties of the computed state? To assess this, let us consider an example, the computation of the meaning of a particular sentence s of the external, public language from the meanings of its parts. On the non–semantic conception of computation, a final state is computed, with a certain complex internal syntactic property. Given the semantics for this syntax, states with this complex property of the finally computed state have the following content: that the external sentence s means that p. Can we say that, once we add the semantics for the computed internal syntactic properties, the non–semantic computation explains a person's understanding the sentence s as meaning that p? One reason we cannot is that this explanation does not draw on any information about the meaning of the parts of the sentence s. The initial states from which the final state is computed, on the non–semantic conception, are not characterized in the explanation as carrying any information about the meaning of the parts of the sentence at all. So perhaps it will be said that we should supplement them with semantic characterizations too, to the effect that various expressions, which occur in s, have certain meanings. When we have both supplementations, though, it begins to be hard to defend the claim that we are dealing with a purely non–semantic conception of computation, and for two reasons.

The first reason is that it looks very much as if we have the computation of one content–involving state from another content–involving state, and one which supports content–involving counterfactuals. ("If this word had been taken to have a different meaning, the whole sentence would have been assigned a correspondingly different meaning".) The second reason is that the identities of the particular syntactic properties mentioned in the non–semantic computational explanation are not explanatorily important to the content–involving explanation. There seems to be a clear sense in which two persons with different languages of thought may be computing the same meanings for the sentence s from the same information about

---

erties of action which are psychologically explained; and which properties are explained is fixed by the contents of the explaining intentional states. On the non–semantic conception of computation, states which explain computationally have no access to those contents. So it is similarly a problem how those explaining states could explain actions under relational descriptions fixed by those contents.

its parts, and doing so in the same way at the content–involving level. Conversely, two persons engaged in the same computation, as syntactically identified, may be computing very different meanings for a sentence s, and from very different information, semantically identified, about the meanings of its constituents.

These points prompt a more general reflection. Presented with a proposed explanation of a content–involving fact, one should always raise this question: "Could the proposed explaining conditions equally hold when the phenomena to be explained have different contents?". If the answer is "yes", the explanation must be incomplete at best. This applies both to proposed explanations of how a subject comes to be in a content–involving state, and to proposed explanations of why there are certain law–like links between intentional states. It provides a constraint both upon particular psychological explanations of content–involving phenomena, and upon a philosophical account of the nature of such explanations.

The argument so far has not relied upon any particular substantive theses about the nature of content, beyond the unexceptionable presupposition that semantic properties go beyond syntactic properties. The reasoning up to this point is still applicable even under an internalist conception of content. Even if semantic properties are internally individuated, so that content is 'narrow', semantic properties are nevertheless different internal properties from syntactic properties, and are not explained by purely syntactic computational explanations. Logical constants are sometimes thought of as expressions for which a narrow account of content is correct. If their content is inherited from some designated role in inference, that content is not determined just by their syntactic properties. The same even applies to the notion of an "and"-gate, if that is meant to be a semantical characterization. An "and"-gate may be characterized simply relative to certain assignments of 0 and 1 to the nodes to which it is connected. That is not yet a semantical characterization —indeed, if falsity and not truth were the semantic significance of the assignment of 1 to a node, such an "and"–gate would function semantically like alternation (the output node indicating falsity iff both the input nodes indicate falsity). In short, the principle that syntax cannot determine semantics holds for anything recognizable as content. The challenge as formulated so far in this paper applies to virtually any theory which is prepared to use the notion of content at all.

For any externalist about content, as I am inclined to be, the apparent gulf between the resources of the non–semantic conception of computation and intentional states as explananda is of course

widened further.[3] The externalist will add to the preceding argument the observation that explanations involving only syntactic states can explain only internally individuated states, whereas the intentional states to be explained are not internally individuated.

I want now to elaborate a little one particular conception of the nature of externally individuated states. I do so not to argue further for the gap between the non–semantic conception of computation and any explanation of intentional states (the argument has already been given). The two reasons for this elaboration are (i) it states further constraints on a good empirical explanation of intentional states, and (ii) the positive resolution I will be recommending actually draws upon the same general conception as is instanced by the following account of the externalist nature of intentional states.

What an ordinary psychological explanation explains is not, except in very special cases, a bodily movement. What is explained is rather the fact that there occurs a bodily movement which has certain relational properties which involve its more or less extensive surroundings, physical, psychological and social. A given bodily movement has indefinitely many relational properties. A particular movement of the hand may simultaneously be: an interposing of one's hand between the sun and one's eyes to give shade; a gesture to attract a friend's attention; part of a signal in semaphore; and so on. A particular psychological explanation may explain the occurrence of an event with some of these relational properties, while not explaining others. It is symptomatic of different explanations that different counterfactuals are sustained. When the explanation does explain the event's relational property of being an interposing of the agent's hand between the sun and his eyes, then, other things equal, this following counterfactual will be supported: if the sun had been in a different position, his hand would have moved to a different position. If what is explained is rather his making a certain semaphore signal, then other things equal, that counterfactual would not have been supported. Different counterfactuals would have been supported instead.

What goes for explanations by intentional states also holds, correspondingly, for certain explanations of intentional states. To take one particularly clear case: when you form a belief as a result of my saying something, what is crucial to the formation of your belief is, in general, the sense of what I say. Except in special cases, the particular words I utter do not matter —any others understood by you as

---

[3]For externalism about content see, as starting points, Putnam (1975) and Burge (1979).

having the same meaning would have produced the same beliefs. The same goes for other relatively intrinsic properties of the utterance, such as its pitch and volume (within limits!). But for a sentence to have a certain meaning is for it to have a highly relational property. Similar points could be made about knowledge gained by perception.

It becomes very plausible to endorse the following general thesis: the identity of any state with an intentional content is at least partially constituted by the fact that in suitable circumstances it can explain, or be explained by, relational properties of external objects and events. I will be taking it that this general thesis is correct. If it is correct, then a satisfactory computational explanation of someone's coming to be in an intentional state will have to be an explanation of a state with these distinctive powers of explaining relational properties of events.

The benefits of explanation by externally–individuated states cannot be achieved simply by supplementing of an explanation which appeals only to internally–individuated states. Take an event, intrinsically characterized in terms of the bodily movements which constitute it, and consider an explanation of the event so characterized, an explanation by internal states of the subject. These internal states may be neurophysiological, "syntactic", or even states involving narrow content (if such there be). Now imagine this explanation to be supplemented with a true statement of some of the environmental relations in which the explained event stands. Can this supplemented explanation amount to an explanation of the environmental relations in which the explained event stands? It cannot. If it were to constitute such an explanation, then explaining an event under any of its descriptions would be explaining it under all of its descriptions, which must be overshooting. An event explicable as my pointing towards the north star need not be explicable as an event of my pointing in the direction of some further, distant body in the Milky Way. It need not be so explicable, even if for lawful reasons the north star and that further body are always in the same direction. There is elsewhere in the literature some of the necessary further elaboration of this conception of the explanation of relational properties.[4]

In the face of the tension between the need to explain semantic properties of intentional states, and a purely non–semantic conception of computation, one can only have sympathy with Stephen

---

[4]See Peacocke (1993); and for earlier discussions of relational explananda Burge (1986), Hornsby (1986) and Peacocke (1981). The "object–dependent" treatment of singular senses in that last paper need not be part of the present package: see Peacocke (1993, section 2).

Stich's reaction to Fodor's original paper, viz. that "a thoroughgoing acceptance of the computational paradigm entails a rejection of the representational theory of the mind" (Stich 1991). If non–semantic computational explanations were the only sorts of explanation on offer, there would be principled reasons for despairing of explanations of specifically intentional properties. But there is a different conception of computational explanation, which will allow us to steer between the view that computational psychology rests on a mistake, on the one hand, and Stich's radical scepticism on the other.

2 Content–involving Computation

When we describe a system as, for instance, computing what a sentence means from a syntactic description of it, or computing the shape of an object from two–dimensional information about its motion, intuitively we want to say that an event or state with one representational content is explained by the occurrence of an earlier event or state with another content, and is explained in accordance with a certain rule. We can make this intuition more explicit by introducing the notion of a *content–involving computational description* of a pair of events (or particular states). Such a description is to consist of three specifications:

(i) a specification of a content–involving property of the first event (or state);

(ii) a specification of a content–involving property of the second event (or state); and

(iii) a specification of a content–involving rule stating how the content–involving property given in (ii) is computed from the content–involving property given in (i). This third specification is a description of a content–involving algorithm.

Clauses (i)–(iii) can be generalized straightforwardly to the case in which the content of some event or state is computed from the contents of an array of several events or states.

The notion of content employed in these characterizations is almost entirely generic. The content of a state may be something at the level of reference, such as the value of a particular physical magnitude, or the location of an axis of symmetry of an object. It may equally be at the level of sense. Nor is there any requirement that in a correct content–involving computational description of a given pair of events, the two contents be of the same general type. On the

contrary, it is plausible that there are important kinds of case in which they need to be of different types. One example is that in which the concept under which something is perceived as falling is computed from properties and magnitudes at the level of reference.

What falls under a content–involving computational description may and will also fall under descriptions which are not content–involving. So properties which are content–involving and properties which are not content–involving may be properties of one and the same process. It is also worth noting that it is by no means plain from these characterizations that any pair of events or states falling under a content–involving computational description must in some way involve expressions in a language of thought. It would at least need substantive argument to establish that.

If content–involving properties could explain only other content–involving properties, and if they could be explained only by other content–involving properties, then content–involving computational descriptions would be explanatorily insulated from properties of the world which do not involve content. If we are not to be faced with that unattractive prospect, we must acknowledge two kinds of mixed case. A description of a pair of events (or states) in a given system is a mixed case of the first kind if it consists of three specifications:

(i) a property, which is not content–involving, of the first event

(ii) a content–involving property of the second event; and

(iii) a specification of a general principle which states how content-involving properties, including that given in (ii), are explained in the system in question by non–content involving properties, including that given in (i).

A description of how, in visual perception, the first contents are computed from content–free properties of the retinal image will be a mixed case of the first kind.

A description of a mixed case of the second kind is, naturally, the mirror–image case: the first event is described as having a content–involving property, and it explains a property, which is not content–involving, of the second event. Mixed cases of the second kind must exist if there are to be computational explanations which culminate in content–involving explanations of action.

I suggest that the explanations given in true content–involving computational descriptions of events are special cases of a general type I identified in "Externalist Explanation" (Peacocke 1993), and outlined above. They are special cases in which one externally–individuated state explains another. In the illustrations I gave above of

explanation by ordinary intentional states, we have (or have something which contains) a non–psychological environmentally identified explanandum, such as a movement to the line between the subject's eyes and the sun. For present purposes we can call these "Ur–explanda". They are the explananda whose status as externally–individuated does not depend upon their relations to some other external or externally–individuated state —they are already about the environment. On the side of the explaining conditions, we could equally introduce the notion of an Ur–explanans.

An externally–individuated state need not be one which is individuated by its relations to Ur–explananda or Ur–explanantia. It may be a second–level state which is individuated by its relations to first–level states which are in their turn individuated by their relations to Ur–explananda or Ur–explanantia. (Second–order propositional attitudes may fall under this case.) Or the externally–individuated state may be a third–level state which is individuated by its relations to such second–level states; and so forth. It is a sufficient condition for a state to be externally–individuated that it features somewhere in this hierarchy of levels.

A content–involving computational explanation is characteristically an explanation of and by an externally–individuated, content–involving state. When both the explaining and the explained states are content–involving, a distinctive type of counterfactual is sustained by the explanation, one whose antecedent and consequent both have an externalist character. If the system had not been in the first, content–involving state, it would not have been in the second, computed, content–involving state. (I prescind from overdetermination and many–one functions.) Given the algorithm employed in the computation, there will also be support for a particular counterfactual stating that if the system had been in a certain particular different antecedent, content–involving state, it would have been in a certain different, content–involving state. Such counterfactuals do not specify any particular intrinsic, internally–individuated state in which the organism would have been had the antecedent been true. Indeed, one and the same counterfactual property captured in these conditionals can be true of organisms with rather different internal representations, provided their external relations are such as to sustain the same content–involving descriptions.

All these points apply straightforwardly to content–involving subpersonal computations in the visual system. An algorithm which computes depth from specifications of two images can be a correct component of the computational explanation of depth perception for two different organisms whose mental representations of depth,

in respect of their more intrinsic properties, are rather dissimilar. Similar points apply in the explanation of linguistic understanding. Let us take an example in which the antecedent condition is plausibly externalist. Tyler and Marslen–Wilson showed that in a sentence beginning "If you walk too near the runway, landing planes...", the occurrence of "landing" will be heard as an adjective rather than a gerund. The reverse is true of sentences starting "If you've been trained as a pilot, landing planes..." (Marslen–Wilson and Tyler 1987). It is quite implausible that the explanation has to do with the particular words in the antecedent of the sentence, considered independently of their meaning. We would expect the effect to be present in any English sentence with an antecedent sharing the same features involving meaning —and meaning is an externalist notion if anything is. Another example is provided by disambiguation in dichotic listening experiments (Lackner and Garrett 1972). In those experiments, it is the meaning of the sentences heard in the unattended channel which biases the interpretation of the ambiguous sentence heard through the attended channel. Any other sentences with the same meaning would have had the same effect, and so would many other ways of uttering the sentences actually heard in the unattended channel. In these cases, we have explanation of a content–involving state by another content–involving state, together with the sustained counterfactuals of an externalist character.

It is one thing to give illustrations, but another to develop a general account of content–involving subpersonal computational explanation. A general account must comprise at least the following. First, we should want some description of the principles governing correct ascription of content to subpersonal computational states. If we are unclear about that, there will always be some unclarity about the significance of explanations invoking those states. Second, we should want a positive general account of what is distinctive of content–involving computational explanation. This should be accompanied, third, by a description of those goals which can be achieved only by citing content–involving computations as explanations, and a statement of the reasons why this is so. I attempt these tasks in turn, the first in the next section, and the others in §4 and §5.

3 Ascribing Subpersonal Contents

Ascription of the personal–level contents of propositional attitude psychology is answerable to the overarching constraint of making the subject of the attributions intelligible (Davidson 1984; Grandy 1973;

McDowell 1986). That general description would be agreed upon by theorists who otherwise differ on whether the constraint can be spelled out further, and if so, in what way. Is there a corresponding overarching constraint on the ascription of subpersonal contents? A plausible corresponding constraint is this:

> correct ascriptions of content to subpersonal states are answerable to facts about the relational (environmental) properties of the events they explain, and to counterfactuals about the relational properties of the events they would explain in various counterfactual circumstances.

By linking subpersonal ascriptions of content to relational properties of events, this constraint gives subpersonal content an externalist character.

The event whose relational properties a state with externalist content helps to explain may be an action —the case of an Ur–explanandum. But the explained event may also be that of a subpersonal system entering a state which itself has an externalist content. In such a case, the overarching constraint operates recursively.

What is the relation of answerability mentioned in the constraint? Suppose a set of states, including subpersonal states, collectively represents the world as being a certain way. Then, when the subject of the states has a certain goal, certain behaviour, characterized externally by its relations to the world, is made appropriate by those states. Entering various other externally individuated states may also be made appropriate. To a very rough first approximation, and with all sorts of refinements to be detailed below, the content attributed to subpersonal states must be that which makes appropriate the relational properties of the events they explain, and which makes appropriate the relational properties of events they would explain in counterfactual circumstances.

Attributions of contents to subpersonal states made in accordance with this overarching constraint must also be limited from above, so to say. A set of attributions which is more fine–grained than can be justified by the relational truths mentioned in the overarching constraint is to be rejected. This is the natural generalization of what in *Sense and Content* I called the Tightness Constraint (Peacocke 1983).

The mental states referred to in everyday propositional attitude psychology are externally individuated. We have also noted that it is plausible that those everyday states are individuated in part (at least) by their power to explain relational properties of actions. If this is so, then explanation by subpersonal states with content can

include the explanation of propositional attitudes. When it does, the content which is the content of a personal–level state is computed from the contents of some subpersonal states. In such cases, the explaining subpersonal states will, by transitivity, also contribute to the explanation of what the personal–level states themselves explain. This of course does not make the personal–level states explanatorily idle. On the contrary, the explanation of the final states by the initial states in this chain is dependent upon their causing the personal–level states.

The description given so far of the overarching constraint may make it sound as if this account is concentrating on the ascription of content on the output side of the organism's economy, to the exclusion of the subpersonal contents computed in perceptual processes. But in fact the two go together, for reasons of principle. Suppose an action is explicable under the relational description "picking up a 50p coin from the tray of coins". Its being so explicable involves the truth of certain counterfactuals: other things equal, the agent would have picked up a 50p coin even if it had been in a different position on the tray; and so forth. If the truth of this range of counterfactuals is also not to be a miracle, the subject must have some perceptual means of identifying 50p coins; and his being in the perceptual states which yield identification of the 50p coins is something for which a content–involving computational explanation is appropriate. This point is quite general, and applies to any perceptually identified property, relation, kind or magnitude which enters the relational characterization under which an action is explained. We cannot make sense of a range of states explaining behaviour under environment–involving descriptions unless some of those states are themselves, on suitable occasions, explained by certain environment–involving conditions.

It is far from evident that the overarching constraint restricts explanation by states with subpersonal content to cases in which the ultimate target of the explanation is something at the level of propositional attitude psychology, or its distinctive explananda. There certainly appear to be relational explananda which are of the right kind to receive treatment by states with content, even though they do not concern matters which are normally, or even ever, intentionally controlled. An example of the first sort —not normally intentionally controlled— would be the explanation of how an upright organism keeps its balance. This certainly involves relational explananda, for it involves the organism's relations to the gravitational vertical. It is not apparent that there is an a priori reason of principle that subpersonal computational explanations should not be appropriate in

this case. A bipedal organism may well contain a subsystem which computes whether the line between its feet is gravitationally below its centre of gravity, and if not where it is (and what it should do about it). More generally, for a device which has an ability which is environmentally individuated, it is an open possibility that a content–involving computation may be the appropriate explanation of how it has the ability. Devices produced by natural selection come to mind in the first instance. But an explanation of a persisting environmentally–individuated capacity must exist even in a device initially produced by random mutation.

On the overarching constraint I am defending, a contentful state which is reliably produced in a certain kind of environment need not correctly represent that environment. According to the overarching constraint, it can be legitimate to assign contents in a way which results in a computation of an incorrect magnitude, provided that assigning the correct magnitude would lead to unexplained failures of the agent's actions to satisfy the descriptions one would expect if the content had been correctly computed. Misperceptions of distance or direction will have their consequences, notably for the relational properties of actions intentional under suitable spatial descriptions involving distance or direction. We can call such cases *forward–looking certification* of assignment of an incorrect magnitude.

Equally, there can be *backward–looking certification* of an unfulfilled content in the production of action. A person may hear a tongue–twister uttered, and fail in the task of uttering it himself. If we were to ascribe to him a successful intention to utter a certain sentence, we would have to conclude that he must have misperceived the sentence uttered to him. But to suppose that would be undermined by various other relationally characterized actions he could perform, such as his ability to write down the uttered sentence, to match its components with other heard utterances, and so forth. The overarching constraint is, then, more a subpersonal analogue of the personal–level principle that we should maximize intelligibility, rather than of the principle of charity (that we should maximize correctness).

4 Answering How–questions

I now turn to the second of the issues raised at the end of §2, the question of what is distinctive of content–involving computational explanations. This section and the next are very much concerned with issues in the philosophy of explanation. Readers in the cognitive

sciences who find themselves less exercised by such problems should skip to §6.

Sometimes workers in a given science have to address a question of the following kind. They already know that something has a given property, and what they then want to know is: how is it able to have that property? We can call such questions "how–questions". How–questions include such examples as: "How is the human body able to avoid waste products building up in the blood?" and "How is a person able to understand a sentence he has never previously encountered?". A how–question must have an answer if the property it asks about is not to be a miracle. The need to find an answer is particularly pressing in the case of properties which involve the relations of the object in question to other things.

Questions of the form "How is a certain kind of organism able to be in states with such–and–such kind of content?" form one subclass of how–questions. I am continuing to presuppose the soundness of the arguments for the view that content is externally individuated, in the sense that what makes a given mental state have the content it does is some complex family of relations in which it stands to things, properties and relations in the environment of its subject. If this is so, then questions of the form "How is a certain kind of organism able to be in states with such–and–such kind of content?" are more specifically how–questions about the relational states of the organisms in question. This point applies not just to perception, but to linguistic understanding, belief, intention, desire, the emotions, and to all other states with externally–individuated contents.

The theory of DNA and its powers provides an answer to a how–question, 'How are humans capable of conforming to the principles of Mendelian genetics?'.[5] The example repays a little reflection, for it illustrates certain points of structural similarity to the psychological case, and also certain points of structural dissimilarity. To have one of the properties identified in Mendelian genetics is to have a highly relational property. For a person to have a recessive gene for red hair —or a recessive 'factor' for red hair, as Mendel would have said— is to have something which involves his relations to hair colour, to other genes (or factors), and to parents and empirically possible descendants. Mendelian theory spells out exactly what the relational property is. A first point to note is that a satisfactory answer to the question "How is a human able to have a recessive gene for red hair?" must cite something with relational properties sufficient to

---

[5]For a very clear introductory discussion of the parts of Mendelian genetics pertinent to the present point, see Kitcher (1982).

account in turn for the relational properties constitutive of having a recessive gene for red hair. The identification of a gene with a certain sequence on a DNA molecule is a good answer to the how–question only because a certain normal chemical environment for the operation of DNA is taken for granted, a chemical environment in which a certain sequence will be causally influential, in the appropriate circumstances, in the development of a human with red hair.

Quite generally, an answer to a how–question about a relational property of an object will be satisfactory only if it adverts at some point either to states which are similarly relational, or to states for which certain relational properties are presupposed. If the states mentioned in a subpersonal psychology are to answer how–questions about content–involving properties, the explaining states must stand in certain external relations. An attempted answer to a how–question about relational states that does not mention relational states would be at best incomplete —that was precisely the worry about the non–semantic conception of computation back in §1 above. A proposed explanation of an inherited trait which cites properties of DNA molecules would hardly be satisfactory if these presupposed normal chemical environmental conditions were not to obtain. What, without those normal conditions, could be the connection between properties of the DNA and the properties of the resulting developed organism?

Marr's celebrated three levels of description can also be regarded as incorporating a distinction between what is computed and how it is computed (Marr 1982, Chapter 1). The correct description, at Marr's second level, of the algorithm is an answer to a how–question: "How is the function specified at level one computed?". Marr's conception of his three levels has come under various attacks in the past few years, but none of the attacks known to me has damaged the distinction between the function computed and the algorithm used to compute it. In particular, the claim that the relation between the function computed (level 1) and the algorithm which does the computation (level 2) is one–many should not be thought to imply the epistemological thesis that we can always discover which function is being computed without investigating the algorithm at the second level. Patricia Churchland and Terrence Sejnowski write:

> In Marr's view, a higher level was independent of the levels below it, and hence computational problems could be analyzed independently of an understanding of the algorithm which executes the computation, and the algorithmic problem could be solved independently of an understanding of the physical implementation. (Churchland and Sejnowski 1990, p. 368)

There is evidence that the view Churchland and Sejnowski criticize was indeed Marr's view. Marr did write that "an algorithm is likely to be understood more readily by understanding the nature of the problem being solved than by examining the mechanism (and the hardware) in which it is embedded" (Marr 1982, p. 27). But Marr could and should have insisted upon the significance of the three levels of description without commitment to the epistemological thesis he endorses in passages such as that just quoted. Accepting the three levels of description does not commit one to endorsing a move from a non–epistemic to an epistemic sense of "independence". Maybe one cannot, for a particular process, discover the correct description at level one without first investigating both of the lower levels appropriate for that process. All the same, once the descriptions are discovered, one can say that the algorithm and implementation answer the question of how the system computes the function identified at the level–1 characterization. This seems no more problematic than someone's not discovering the function of the kidneys until they investigate the detailed chemical processes which they carry out.[6]

While a well–posed how–question must have a tractable description of the property to be explained, a description can be tractable without providing the fullest possible analysis of what it is to have the property. It would be the fulfilment of one sort of dream if, for each type of mental state with content, we had a full constitutive theory of what properties and relations a subject must stand in to be in that state, where a full constitutive theory would be one derivable from substantive and complete theories of the type of state and the identity of the contents in question. It would further be a fulfilment of a second dream if we could give a subpersonal psychological explanation of how an organism is able to have the properties and relations identified in that full constitutive theory. But explanatory progress is possible before any completeness is attained. Indeed it is not even clear that completeness has to be possible, even as a regulative ideal. Just to have an account of the sorts of content involved

---

[6]I should emphasize that I am, though, in complete agreement with Churchland and Sejnowski that "the idea that there is essentially one single implementational level is an oversimplification" (p. 369). It has also been questioned whether it always has to be possible to construct a Marrian level 1 description of a computational process (thus Boden (1988, p. 227)). But when we are concerned with content–involving computation, it is a necessary truth that where there is a computation of a content–involving function, there is a function–in–extension which is the content–involving function computed. Wherever there is computational explanation, there will be a properly answerable how–question which is answered by a correct statement of the algorithm involved in the process.

in a given mental state, with a statement of some of their relations to one another and to various other states, is already to have something for which a well–defined how–question can be addressed and answered. It seems obvious, for instance, that progress is made in the psychology of perception by theories which take for granted that perceptual states have certain kinds of content, and then go on to provide answers to various empirical questions, including how–questions, which arise for those states. This can be done before we have anything like a full constitutive philosophical account of perceptual content. And if there is no such thing as a full account, it is still desirable, for each constitutive feature of the content, to have an explanation of how organisms of a given kind are capable of standing in the relations it requires.

I now turn to answer a charge that there is an inconsistency in what I have said. I have been emphasizing that in content–involving computational explanation, we have explanation of externalist states by externalist states, and I said a little about what is distinctive of such explanation. I offered computational explanations of perceptual experience as examples of content–involving computational explanation, whilst also noting that these explanations begin, in the visual case, with properties of the retinal image. But retinal states are *not* externalist states; and corresponding points will apply to other computational explanations of perception (and of linguistic understanding). So it seems that in all these cases, there will be a first content–involving state, externalistically individuated, which is not explained by an externalist state. Hence either the principle that externalist states must be explained by externalist states is false; or else, perhaps, contrary to appearances, even the earliest states in such explanations should be regarded as externally individuated. (This second alternative has the air of an incipiently Gibsonian position.) Corresponding points apply *pari passu* on the action side.

There is, though, a better response to the charge, a response with some theoretical motivation. Any content–involving computational explanation presupposes some *normal environment* for the organism which has the computational states (or at least it presupposes normality in certain respects). This has long been recognised and emphasized in defences of externalism about content (thus Davies (1986)). When the ultimate goal of a content–involving computational description is to contribute to the explanation of the organism's being in the content–involving states of a propositional–attitude psychology, such a presupposition is only to be expected. The very same normal environment is presupposed by that propositional attitude psychology too. Provided the organism is in the

presupposed normal environment, in the visual case the first content–involving externally individuated state which is produced by retinal states will then stand in certain environmental relations. So the explanation does not involve any commitment to that content–involving state resulting whatever an organism's normal environment may be. Correspondingly, the content–involving state will have the property that, in the organism's normal environment, it is explained by, and can in suitable auxiliary circumstances explain, environmental states of affairs. As always, and in accordance with the principles for ascribing subpersonal content in §3 above, the correct content to be ascribed to the first content–involving state will be dependent not only upon what environmental conditions cause it, but also upon the nature of the other processes to which it is or can be an initial input, and upon the environmental effects of those processes too.

It might be replied that if there is a presupposition of a normal environment, then a chain of wholly internalist states could in some sense explain the occurrence of an externally–individuated content–involving state of propositional attitude psychology. Certainly it should be agreed that certain modal objections to which such a position would otherwise be exposed are neutralised by the presupposition of a normal environment. However, an explanation by such a sequence of internalist states would also be much less general than the explanation by content–involving states. The content–involving computational explanation has the property familiar in other cases, that it can apply across a variety of internalist states which different organisms can enjoy. This, incidentally, is the promised point of difference between content–involving computational explanation and genetic explanations involving DNA. The latter are not conceived as applicable whatever the biochemical structure of an organism. Though (I suppose) there might in principle be a more abstract genetics, which operates with a notion of the "instructional content" of a gene rather than its particular chemical composition, explanations which specifically mention DNA do in fact operate at a level lower than that.

So far we have been considering the power of content–involving computational explanations to answer questions about how an organism is able to be in a particular externally–individuated state. But there is also a second kind of how–question to whose answer a content–involving computational explanation makes an essential contribution. Questions of this second kind are of the form: "How does such–and–such state reliably represent correctly (or, indeed, incorrectly) in such–and–such circumstances, within this organism's normal environment?" Here what has to be explained is the property

of correctness (or incorrectness), a relational property of a relational state. Content–involving computations are by their very nature well–suited to provide such explanations. Suppose each of the content–involving subcomputations of a process is a *correct* method of computing the magnitude (or property, or whatever) which that substep computes, from any given initial input values. Then the final step of the process will compute its magnitude (or property, or whatever) correctly. Similarly if there is an incorrect step at some point, then we can see why in certain circumstances the system will produce a final state with an incorrect content (cancelling errors aside). The understanding such content–involving computations can provide of why correctness or incorrectness results is distinctive of explanation by content–involving states, for it relies essentially on the content–involving character of the individual subcomputations. A description of a computation which does not mention the contents of computational states could never yield such understanding. Exactly the same can be said of semantic descriptions of subcomputations within a complex algorithm in arithmetic. Only with the semantic descriptions do we have an explanation of why it reliably yields a correct answer to questions about the value of a certain arithmetical function.

5 Objectivity, Irreducibility and Explanation

It would be widely agreed that a process can be a computation only if it has some true content–involving description. We cannot make sense of the possibility of a computation in which there is no saying what is computed, or what it is computed from. This widely–agreed view is well–expressed by Churchland and Sejnowski: "in the most general sense, we can consider a physical system as a computational system when its physical states can be seen as representing states of some other systems, where transitions between its states can be explained as operations on the representations" (Churchland and Sejnowski 1992 p. 62).

This semantic dimension is not a trivial definitional "add–on". Rather, it is motivated by the original point and interest of the notion of computation. Though Turing described his machines in purely formal and mechanical terms, they were of interest for the theory of computability precisely because the symbol "1" can be regarded as referring to the number 1, and juxtaposition on the machine's tape can be regarded as having a certain significance. We rely on the semantic dimension in regarding each Turing machine as yielding an arithmetical function. We can draw conclusions about

arithmetical functions on natural numbers from results about Turing machines only because of this referential dimension. It is not true that the original, core notion of computation was purely syntactic.

However, to agree that any computation must have some true content–involving description is not yet to incur any commitment to the objectivity, irreducibility or explanatory significance of the content–involving level of description. The possession of each of these properties by content–involving computations has been questioned in one way or another, and I now turn to address these issues.

In Churchland and Sejnowski's own thought, the agreed characterization of a computation is combined with the thesis that whether something is a computation, or a computer, is interest–relative. "We count something as a computer because, and only when, its inputs and outputs can usefully and systematically be interpreted as representing the ordered pairs of some function that interests us" (*ibid.* p. 65). Because of what they call this "interest–relative dimension", computation is not, according to them, a natural kind.[7] The view I have been presenting is very different. The ascription of subpersonal contents is answerable to counterfactuals which are environment–involving. At the most basic level, the ascriptions are answerable to the truth of various counterfactuals involving what in an earlier section I called Ur–explananda and Ur–explanantia. The truth of an ascription of a subpersonal content to a state could be interest–relative only if the truth of these counterfactuals is interest–relative. In the case of the dichotic listening experiment, one of the counterfactuals sustained would be "If different words with the same sense had been heard in the unattended channel, the subject would still have assigned the same interpretation to the sentence heard in the attended channel". I cannot see that the truth–value of this counterfactual is interest–relative. The same applies to the underlying explanatory relation of which its truth value is an indication.

Why should the claim of the interest–relativity of computation be found tempting? Perhaps it is the influence of such a thought as this: "even a sieve or a threshing machine could be considered a computer, since they sort their inputs into types, and if one wanted to spend the time at it, one could discover a function that describes the in-

---

[7]Strictly, what they say is that computers do not form a natural kind. But if this is not to be extended to computation, the claim would be of limited significance. The context of the claim in a general discussion of computation in *The Computational Brain* certainly suggests that it is meant to have a more general significance.

put–output behaviour" (*ibid.* p. 66). But in those two examples, the notion of representational content does no explanatory work. If we say of the example of the sieve and the threshing machine that (for instance) that kinds of shape and size are representing themselves, then in attributing those spatial contents and treating the processes involving them as computations, in the nature of the case we gain no explanatory power that we could not have had without such spurious attribution of contents.[8] It is indeed true that virtually any concrete object at all can be considered as embodying some input–output function or other. This fact poses no threat to the conception of computation as a theoretically significant general kind, just because in the cases we want to exclude, ascribing content is either explanatorily redundant, or overshoots. In genuine computational explanation, neither of these is the case.[9]

The position I have been defending supports the irreducibility of content–involving subpersonal computational explanation to neurophysiological explanation. The initial argument for irreducibility is that the computational explanations can form part of an explanation of a relational explanandum, which concerns the subject's, or the system's, relations to the environment. Neurophysiological explanations form part of explanations with explananda that concern bodily properties or movements, and not their environmental relations. The irreducibility of computational to neurophysiological explanations does not conflict with the thesis that there is some kind of dependence relation between computational and neurophysiological explanations. Proposed computational explanations of facts about humans which have no foundation in neurophysiology cannot be accepted. But if the computational explananda are different from those of neurophysiology, it does not even look as if the case can be assimilated to such an example of macro/micro reduction as that of the gas laws to the mechanics of particles. What is explained by the gas laws —particular instances of a general relation between pressure, volume and temperature— is also explained by the reducing theory (plus identities or "bridge laws"). Yet even a description of a state as representing the location of something in head–centred coordinates —the sort of representational content which is the bread–

---

[8]In the example of the slide–ruler, mentioned by Churchland and Sejnowski, the device is counted as a computer because of the human intention that it is to be used in a certain way.

[9]On this point at least, I am entirely at one with Churchland and Sejnowski. "By contrast with systems we conventionally call computers, the *modus operandi* of some devices are such that a purely causal explanation, without reference to anything having been computed or represented, will suffice" (p. 68).

and–butter of a computational approach —is an environment–involving description.[10] Similarly, the computational explanation of a system's being in such a state can form part of a larger explanation with the explanandum that the subject reaches out in a certain head–relative direction (say, to switch off a device which is emitting a sound).

Am I relying on an overly demanding standard for reduction? The reasoning just given certainly needs elaboration, for there are indeed other good cases of reduction in which we do not demand strict identity of the explananda of the reduced theory with some of the explananda of the reducing theory. The explananda of Newtonian mechanics do not involve any relativization to a frame of reference when they concern temporal relations, whereas there is always such a relativization in the explananda of the special theory of relativity. Yet the theory of special relativity does explain the approximate correctness of Newton's laws for objects moving much more slowly than light. So the question arises: are there some kinds of parameters which can be fixed, or relativizations made, of such a kind that once they are in place, we can legitimately say that computational explanation reduces to the neurophysiological?

To a particular neurophysiological explanation of a bodily movement or a bodily property, we could add two things. The first addition would be a statement of the general principles for ascribing contents to computational states, insofar as they can be made explicit. The second addition would be a statement, for each of the states mentioned in the given neurophysiological explanation, of the relevant environmental relations in which they stand and their relations to other relevant states. With this double supplementation, we

---

[10]Churchland and Sejnowski's position often comes over as reductionist: they give as a "negative version" of the working hypothesis of their book that "it is highly improbable that emergent properties are properties that cannot be explained by low–level properties..., or that they are in some sense irreducible, causally *sui generis*, or as philosophers are wont to say, 'nomologically autonomous', meaning, roughly, 'not part of the rest of science'" (p. 2). On the other hand, their discussion of particular examples often insists upon points which the supporter of irreducibility will highlight. Thus on the role of neurons whose behaviour can be explained as computing head–centred coordinates from retinal position of a stimulus and position of the eyeball in the head, they write: "Knowing that some neurons have a response profile that causes other neurons to response in a certain way may be useful... but on its own it does not tell us anything much about the role of those neurons in the animal's visual capacity. We need additionally to know what the various states of the neurons represent, and how such representations can be transformed by neural interactions into other representations" (p. 68).

could arguably derive an explanandum which relates to the environment in the way that the conclusion of a computational explanation does. Such derivations give a mapping from neurophysiological explanations to computational explanations by relying upon the principles of attribution of computational contents to subpersonal states, together with the relational, environmental properties of particular neural states. Can the inclusion of the principles of attribution be regarded as analogous to the "rules for attributing pressure and temperature" when these connect pressure and temperature with molecular magnitudes? After all, any successful reduction will have some form of bridge laws. So can the inclusion of the rules of attribution be assimilated to this more familiar phenomenon?

The fundamental difference is that the rules for attributing content to the computational states relate contents not (or not only) to neurophysiological states, the states of the putatively reducing theory, but to environmental matters too. This is why the case is disanalogous to other cases of reduction. It is also why the second addition, which includes a statement of certain environmental relations of the particular neurophysiological states mentioned in a particular explanation, is required. There would be no need for that second addition were the principles for attributing contents to allude only to neurophysiological matters. The existence of such a relationally supplemented neurophysiological explanation, for each computational explanation, is better seen as an elaboration of what is involved in computational explanation having a dependence upon the neurophysiological which falls short of reduction. In fact, once the two additions are made, it becomes clear that the identity of the particular neurophysiological state is irrelevant to the explanandum of the computational state. Any other neurophysiological state standing in the right relations to have the same content would equally have sufficed.

There is a different position, one which does not need to involve neurophysiological reductionism, but which still questions the significance of a content–involving conception of computation. This position holds two theses: (a) that the benefits of a computational explanation which involves externalist contents can be obtained by appropriate supplementation of a computational explanation which does not involve externalist contents; and (b) that this is a preferable approach, because it achieves the goal of computational explanations, that of exhibiting the mechanisms underlying cognition. This position is not committed to neurophysiological reductionism, since it may insist that there is some species of content which is not externalist, and which is well above the neurophysiological level. This

species of content could be used in the computational explanations, which, according to claim (a), can be suitably supplemented.

The position consisting of (a) and (b) is close to one defended by Francis Egan, who writes:

> the explanation of organism/environment interaction is not the primary goal of computational theorizing, and such explanations are forthcoming only when a computational theory is supplemented by further assumptions about the normal environment in which the described cognitive mechanisms are deployed. (Egan 1992, 447).

Egan also writes that her treatment "depends in part upon the view that the goal of such theories is to characterize the mechanisms underlying our various cognitive capacities, and further, that this goal is best served by theories which taxonomize states individualistically" (pp. 444-445). Egan's position may be only close to, and not identical with that consisting of (a) and (b), because what the supporter of content–involving computation wants is underdescribed as an "explanation of organism/environment interaction". Perhaps supplementation can help to explain that interaction, but what was wanted was different —it was an explanation of the organism's being in states whose individuation involves relations to the environment. Nonetheless the line of thought is very similar, and I want to consider whether the kinds of reasons Egan advances can be used to undermine the arguments I have been developing. I take (a) and (b) in turn.

The objection that the explanatory benefits of externalist, content–involving computations can be gained simply by supplementing a description of the computation which is not externalist is the exact parallel for this subpersonal domain of a response to externalist arguments in the case of the states of folk psychology. When an externalist about folk–psychological states argues that they explain actions only as characterized in environmental terms, internalists are inclined to reply that internalist states are the only truly explanatory states. Internalist states explain bodily movements individualistically described, they say. And, they continue, to get an explanation of the relational properties of the action on which the externalist insists, we have only to supplement the individualistic description of the action with some information about its environmental circumstances.[11]

---

[11]See the discussion of this response in "Externalist Explanation" (Peacocke 1993, pp. 208-9).

The rejoinder to the response that supplementation can suffice to meet the need is also the same for subpersonal psychology as it is for personal–level folk psychology. The rejoinder is one we have already given —that by taking an explanation of an event under one of its descriptions, and supplementing the explanation with the information that the same event falls under another description, you do not thereby have an explanation of the event under the second description. Consider a computational explanation of an organism's being in a state which represents an object in front of him as having a certain shape and colour. In the context of other attitudes on the part of the subject, when all is working well, this computational explanation can contribute to a larger explanation of why the subject grasps the thing in front of him which has that shape and colour. If that same object was the 1,000th off the factory production line, we do not equally have a larger explanation of which the computational story forms a part, a larger explanation of the following fact: that the subject grasped the 1,000th object off the production line. (Here of course the definite description "the 1,000th object" has narrower scope than the "that..." clause containing it.) In particular, the necessary counterfactuals are not supported. It is not true that if some other object, possibly in a different location relative to him, had been the 1,000th off the line, he would have grasped *it*. But he would have grasped some other object, possibly in a different location relative to him, had it had the same shape and colour.[12]

Part (b) of the rival position in question states that only internally individuated computational states can meet the goal of identifying the mechanisms underlying our various cognitive capacities. If a mechanism is treated definitionally as something which is not individuated in content–involving terms, it is indeed an immediate a priori truth that content–involving computations are not mechanisms. That we should aim to identify the content–free mechanisms which make cognition possible is certainly something I endorse. But endorsing this goal cannot show that externalistic content–involving computation does not also have a distinctive explanatory role. Nor does it undermine the earlier arguments that these roles cannot be played by purely individualistic states.

---

[12] "But suppose we consider a property nomologically coextensive with the conjunction of shape and colour: then this argument you have just given will not apply". Agreed —to cover that case, we should appeal rather to the principles governing the ascription of subpersonal contents, which are discussed in §3 above. The counterfactual test is only an approximate guide to the fulfilment of what is required by those principles.

If the term "mechanical explanation" is restricted to states which are not described as content–involving, there remains a different, significant sense in which a content–involving explanation can be mechanical. In an ordinary, personal–level intentional explanation of a person's actions or mental states, we take for granted a huge range of a person's cognitive capacities. We take for granted his ability to perceive, to reason, to remember, to understand, amongst much else. Even if a subpersonal explanation uses a notion of content, as I have argued it must, it does not take those personal–level cognitive capacities for granted. Rather, it seeks to explain them. A computational psychologist is certainly not taking ordinary perceptual capacities for granted when he argues that, for instance, perceptual grouping phenomena are explained by the computation of minimal translational invariances in the presented scene. The ability of a subpersonal device to carry out the computation will be explained in terms of the operation of a certain algorithm. It will or should be accepted as a constraint on the enterprise that it would be quite illegitimate for the explanation at some point to rely on the subpersonal system's "noticing a certain grouping". So there is middle ground between the personal–level, intentional explanations in which cognitive capacities are presupposed, at one extreme, and the purely mechanical, or content–free explanations, at the other. It is in this middle ground that content–involving computational explanations operate. Reserving the term "mechanical" for the content–free explanations does not make this middle ground disappear.

Despite all these arguments, there may remain a residual influential impression that the total causal story is still that told by the syntactic description of a computational process, and in the nature of the case the syntactic process will be blind to any semantic properties. I think this impression results from an unexamined use of the notion of "the total causal story". If we take a sequence of particular events, such as the handing over of some coins in exchange for a product, it is hard to defend the application of the notion of *the* causal story to be told about it. The sequence of events will have a physical description, an intentional description, an economic description, amongst many others. Perhaps we can make sense of the notion of *the* causal story if this is relativized to descriptions of a specified kind. The total causal story involving neurophysiological descriptions would then be distinguished from the total causal story involving intentional descriptions; and both of these from the total causal story involving economic descriptions; and so forth. But the applicability of these relativized notions of the total causal story has no bite against the conception of content–involving computation.

The total story of a sequence of events at the syntactic level will indeed say nothing about content. This is entirely consistent with the causal story at the content–involving level making use of a notion of computation.

In the philosophical literature, a distinction is properly drawn between what are, on the one hand, broadly epistemic theories of explanation, and on the other, broadly "ontic" or "metaphysical" theories.[13] Epistemic treatments of explanation hold that what makes something an explanation is wholly explicable in terms of states of knowledge or belief, or what is found to be informative, or even conversationally apposite.[14] Non–epistemic, 'ontic' theories by contrast state that genuine explanations must always involve some notion of causation, or law, or counterfactuals which, they claim, cannot be reduced to epistemic notions. This is a very broad category: two theorists may disagree radically about the nature of laws, or causation, or counterfactuals whilst both being non–epistemic theorists. Regularity theorists of causation, and those who reject pure regularity theories, may both be non–epistemic theorists. I mention the distinction in order to raise this question: do the points I have made about the significance of content–involving computational explanation require one to hold an epistemic view of explanation?

While the conception of content–involving computation I have been developing could be accepted by someone favouring an epistemic treatment of explanation, an epistemic treatment does not seem to be required for this conception. It is true that it is very easy to slip into formulating some of my recent claims as points about knowledge or information, as when we say that a neurophysiological explanation "*tells* us nothing" about the environmental properties of the movements explained. However, if we hold that explanation is not fundamentally an epistemic matter, these formulations should not be taken as fundamental, but as simply contrasting the explananda of content–involving computational explanation with those of explanations which are not content–involving.

Combining the present treatment of computational explanation with a generally non–epistemic approach to explanation does impose certain constraints, however. If we hold that combination of views, we cannot also accept both that explananda of temporal states of affairs have the form "particular event e occurs", and also identify the event of a particular bodily movement's occurring with the event

---

[13]For more on this distinction, see Ruben (1993).

[14]So for present purposes, I include pragmatic approaches to explanation under the heading 'epistemic'.

of the person's pointing towards a certain house. That would undermine the distinction I tried to draw earlier between the case in which we have explanation of a particular relational property of an event and that in which we do not. It sits better with the present approach to treat an explanandum as more fact–like, as in general being of the form "$x_1, \ldots, x_n$ stand in relation R at t". This treatment also squares better with the points about the explanation of content–involving phenomena urged back in §1. It would be reasonable to claim that the particular event of a device's creating a mental representation with certain syntactic properties is one and the same as its creating a mental representation with certain semantic properties —much as uttering a certain sentence on a particular occasion can be identical with the event of making an assertion with a particular content. If the explanandum of a psychological explanation is regarded as being just the occurrence of a particular event, then one might regard the semantic descriptions of states which explain the event as not explanatorily relevant to it, and justly see nothing missing in the explanatory power of syntactic descriptions of a process. But when the explanandum is treated as being of the form "$x_1, \ldots, x_n$ stand in relation R at t", this deflationary reaction is not an option. An explanandum that an event with a certain syntactic property occurred is distinct from an explanandum to the effect that an event with a certain semantic property occurred. For the latter explanandum, content–involving explaining conditions will be appropriate (when we are not concerned with the "mixed" cases). For enthusiasts of formulations in terms of philosophical logic, we could say: the characterizations of states as content–involving occurs inside, and not outside, the scope of "explains" in true statements about what explains a person's being in a content–involving state.

If on the contrary, we were to treat a temporal explanandum as just the occurrence of a particular event, and were also to identify a bodily movement on a particular occasion with the action explained, we are in danger of finding ourselves willy–nilly forced into an epistemic view of the explanatory contribution of computational explanations. Consider, for instance, David Lewis's well–known theory that to explain an event is to give some information about its causal history (Lewis 1986). This theory is certainly not in itself an epistemic treatment of explanation. But if the bodily movement, and the person's moving his hand between his eyes and the sun are identical (as they certainly seem to me to be), then their causal histories, understood as the tree–structure of events which stand in the ancestral of the causal relation to them, are identical too. Hence information about

the causal history of either one is information about the causal history of the other.[15] It then begins to appear that it is only within the epistemic domain that content–involving explanation could have any distinctive significance. If this consequence is to be avoided, within the context of a fundamentally non–epistemic treatment of explanation, we must distinguish the explananda.

Sometimes we distinguish even between necessarily equivalent explananda. It is not only on the epistemic view of explanation that we can draw such distinctions. An explanation of why a machine prints out a certain formula is distinct from an explanation of why it prints out the formula with Gödel number **n**, even though being that formula and having that Gödel number are necessarily equivalent. The explanation of the machine's printing out that formula may be that it applied a certain syntactic rule to some other formulae, which then yield that formula. An explanation of its printing out the formula with Gödel number **n** would rather be, for instance, its engaging in a numerical computation yielding the number **n**, followed by a calculation of which formula corresponds to it, followed by a printing of the formula. For an event to be explicable under a certain description, that description must feature in the explanatory principle which underwrites the particular explanation.[16]

Besides showing that non–epistemic views of explanation can draw quite fine distinctions, the example of the preceding paragraph also illustrates another point. Suppose the machine to be operating on some initial formulae, presented as input, and applying a syntactical operation to them to yield another formula which it prints out. Particular counterfactuals of the form "If such–and–such different formulae had been presented, such–and–such formulae would have been printed out" will be true. But so also will counterfactuals of the form "If formulae with such–and–such Gödel numbers had been presented, formulae with so–and–so Gödel numbers would have been printed out". This shows that the truth of certain counterfactuals involving properties of the explaining events does not suffice for those properties to be the ones which are explanatory. The counterfactuals sustained are often a good guide to what is explanatory of what, but they are not strictly sufficient by themselves to establish

---

[15]This point applies not only to content–involving explanation, but to externalist explanation more generally (Peacocke 1993).

[16]It may be only a "de–parameterized" version of the description which features in the explanatory principle, if the underlying principle of explanation includes parameters for context, and various objects in a subject's environment.

explanatory relations, and *a fortiori* not what makes something an explanation.[17]

6 The Contrast with Dennett

I want now to sharpen the characterization of subpersonal content which has been emerging from this position by distinguishing it from the conception of subpersonal content developed in Dennett's important writings. The two positions agree that there is a distinction to be drawn between what it is to be an intentional system, and the empirical explanation of how an organism succeeds in being an intentional system. They agree too that it is one of the tasks specifically of psychology to say how an organism which so succeeds in being an intentional system does so (Dennett 1987, esp. pp. 43-65). But there is a major divergence between Dennett and the present conception over the way in which the content of subpersonal states is to be conceived, and over the appropriate kind of explanation of an organism's success in realizing an intentional system. The issue turns on the right account of the relation between the contents of subpersonal states and the content of personal–level intentional states.

An extreme claim of independence of the contents of personal–level and of subpersonal states would be this:

> States involved in subpersonal computations do not have contents which are of the same kind as the contents of personal–level states, nor can they be ascribed contents whose attribution is justified by their power to explain facts about the contents of personal–level states.

This claim is a claim of the independence of the contents of subpersonal computational states from personal–level contents; I will dub it "the Independence Claim". There are various arguments in Dennett's writings which are either arguments for the Independence Claim, or provide the materials for constructing such arguments. Since I am manifestly committed to disputing the Independence Claim, I will be questioning these arguments. But I should emphasize that I do not want to attribute the Independence Claim to

---

[17]I acknowledge that a full defence of content–involving computation must address the question of the efficacy of states with subpersonal contents. This paper is already very full, but I can just indicate that my position on efficacy would be the analogue, for subpersonal content, of the stance adopted in Peacocke (1993). I suspect that those who hanker for more than that have a desire which is never satisfiable.

Dennett without qualification. This is partly because it seems plain that some of the purposes for which Dennett employs the notion of states with subpersonal content in *Consciousness Explained* (Dennett 1991) require those contents to be of the same kind as feature in personal–level states. Consider, for instance, a Pandemonium–like model, of the sort supported in chapter 8 of *Consciousness Explained*, for explaining why a person says something with one personal–level content, rather than a different content. The competing subpersonal demons or agents must have conceptual contents associated with them if it is the victory of one (or a collection) of them which explains the selection of the particular content of the utterance. Nonetheless, important general arguments and materials in support of the Independence Claim are to be found elsewhere in Dennett.

Three arguments are most prominent.

(a) The first argument starts from the premiss that "the brain is a syntactic engine" (Dennett 1987 p. 61). Nothing could perform the impossible task of extracting semantic properties from syntactic features. Is what I have said about explanation by content–involving subpersonal computation in collision with this obvious truth?

It is worth noting that if there is a threat from this argument, then it is one of quite general application. The argument would tell equally against computational explanations which invoke purely "narrow" contents as well, since narrow contents also go beyond syntactic properties.[18] The argument cannot properly be used to favour computational explanations with only narrow contents over computational explanations employing externalist contents. It could not, for instance, properly be used to support the claim that a subpersonal computational psychology is legitimate if it uses only narrow contents, and not otherwise.

The view I have been proposing does not in any case involve anything's executing the impossible task of getting semantics from syntax. In content–involving computation, semantical states result either from other semantical states, or —in mixed cases of the first kind— from non–semantical states together with a presupposed environmental background (in the presence of other constraints). My position agrees that nothing can be intrinsically a representation of something, and it agrees too that for a representation with an externalist content, internal facts about a subsystem's use of the representation cannot fix its content. It is because the present position

---

[18]Such narrow contents would include the notional contents fixed by the "notional worlds" of Dennett (1987).

acknowledges the soundness of those points that it recognizes that the distinctive explanatory role of content–involving states must in certain respects be different from those which are content–free. Explaining relationally–individuated states of affairs is precisely one such respect.

Dennett's own view is that the brain "could be designed... to *mimic* the behavior of the impossible object (the semantic engine) by capitalizing on close (close enough) fortuitous correspondences between structural regularities —of the environment and of its own internal states and operations— and semantic types" (*ibid.*, p. 61). "An animal needs to know when it has satisfied the goal of finding and ingesting food, but it settles for a friction–in–the–throat–followed–by–a–stretched–stomach detector..." (p. 61). Suppose we fix on this detected internal state, and specify the complex of internal and environmental relations in virtue of which detecting it is pretty much as good as detecting that the organism has ingested food. It seems to me that we are then specifying the relations in virtue of which a detector of that state is able to have the content that food has been ingested. I tried to outline some of the rules for the attribution of content to subpersonal states above. These rules are the analogues for the subpersonal level of something which Dennett recognizes at the personal level —his "rules of attribution" of intentional system theory for assigning intentional contents to personal–level states (Dennett 1987 p. 58). If, as I have argued, the semantic and relational properties of later states can be explained by the presence of a state with this content about ingestion, there is a sense in which the mind is a semantic engine after all, and without any implication of executing the impossible.

(b) Dennett also develops an argument, highly threatening to the conception I have developed, which starts from the point that since subpersonal representations cannot have their contents in virtue of the way they are employed by some "exempt user" outside the system, apparently intelligent use of these representations must be explained by dischargeable homunculi. But, the reasoning continues, the mechanical operation of these dischargeable homunculi can involve only intra–systemic transactions. Here is the argument, in another quotation from Dennett on Fodor:

> [Fodor] fails to recognize that [homunculi] still play the theoretical role of fixing the "topic" and "vocabulary" of the messages they communicate. If viewing messages of the inner code as self–understanding representations in this fashion can save Fodor's enterprise from incoherence —and in principle I think it can— it does so by adding constraints to the notion

> of an internal representation system that emphasize rather than eliminate the distinction between personal level attributions of beliefs and desires and sub–personal level attributions of content to intra–systemic transactions. (Dennett 1978b p. 102)

If this reasoning were sound, it would present my own position with an excruciating dilemma. If homunculi fix the semantics of the subpersonal messages, either they are discharged, or they are not. It is not acceptable for them to remain ultimately undischarged. But if they are discharged by mechanisms which are purely internally individuated, as this passage from Dennett seems to suggest, they will be incapable by themselves of underwriting distinctively externalist explanations.

The dilemma overlooks a possibility. A subpersonal state can have its content in virtue of its *unintelligent* relations to other states which are, nevertheless, still relationally individuated. "Unintelligent" here does not mean "free of all content–involving characterizations". As Dennett's own writings have taught us (Dennett 1978a), it should be understood as meaning "not presupposing the psychological capacities it helps to explain". For instance, in neural–network treatments of many capacities, from reading to stereopsis, the computed representations —of phonemes, depth or whatever— have the contents they do because they are produced in certain ways by certain states which also have representational contents of an externalist character. The ways in which the representations are produced are mechanical in the "middle–ground" sense we noted above, and in no way presuppose the capacity to be explained.

(c) A third, equally threatening, argument aims to establish that "beliefs and desires are not the proper objects of study of cognitive psychology. Put otherwise, cognitivist theories are or should be theories of the subpersonal level, where beliefs and desires disappear, to be replaced with representations of other sorts on other topics". (Dennett 1978b p. 105). The argument runs thus: two computationally different programs subpersonally instantiated in two different people, Mary and Ruth, may both support the ascription of a belief with a given content. Ascription of a belief with that content will be equally explanatory, in context, of certain of Mary's and Ruth's actions. The different programs subpersonally instantiated explain rather different patterns of data —Dennett mentions differences in delays, errors and the like as between Mary and Ruth. Belief ascription is just not meant to explain data of that kind. So, the argument concludes, beliefs and desires are not the concern of a subpersonal psychology at all. This argument too would generalize beyond be-

liefs and desires to other intentional states of the personal level, such as perceptual states.

The argument seems to me to involve an inference we ought not to accept. The inference is of the same form as that in this piece of reasoning: "Two different molecular structures may underlie the property, common to two different liquids, of being viscous. The different molecular structures will explain certain differences between the liquids: their different effects on X–rays shot through them, their different patterns of chemical interactions, and the like. The ascription of viscosity is just not meant to explain data of that fine–grained kind. Hence particular molecular structures do not explain viscosity". The move to that last sentence seems to me to be a non–sequitur. The fact that viscosity is consistent with several patterns of fine–grained data cannot show that molecular structure is unexplanatory of viscosity —it is so explanatory. It seems to me that it would be equally fallacious to offer the parallel argument to the conclusion "Subpersonal states do not explain personal–level content–involving states".

Perhaps I am reading Dennett's argument in too loaded a manner, and talking of propositional attitudes as states is begging the question against his whole position. But the point does not rest on a tendentious presumption that the classifications of folk psychology carve subpersonal states at precisely the same joints as the classifications of a subpersonal psychology. Suppose we just speak very neutrally of the predicate "believes that such–and–such" as being true of both Mary and Jane. We may still want a subpersonal psychological explanation of why it is so. Subpersonal theories are indeed answerable to finer–grained data than are the ascriptions of intentional, personal–level psychology, but they can still also be explanatory of less–fine grained data, including the satisfaction by Mary and Jane of this predicate.

7 Further Tasks

I have been making claims about the nature and significance of content–involving computational explanation. There are many directions of enquiry which start out from the point we have now reached. I mention three. The first kind lies within the psychological realm. We ought to want to know a great deal more about the principles for ascribing particular contents to subpersonal states. What I mentioned earlier in this paper are very general constraints on the ascription of content, together with some illustrative examples. In the

theory of conceptual content, which is fundamentally a personal–level notion, we draw a distinction between general constraints on the form to be taken by an account of mastery of a particular concept, on the one hand, and particular instances of that form for certain concepts on the other. A similar distinction applies to subpersonal contents, and it is more of the instances which conform to the general constraints that we should be wanting. Our ordinary capacities for thought about objects, time, motion, matter, other minds, their emotions, intentions and agency amount to an impressive range of externally–individuated abilities and knowledge. The explanations of how we can be in such states, and why we are correct (or incorrect) when we are, must involve rich subpersonal systems of representation. We need to consider not only what these are, but why it is correct to ascribe their particular distinctive contents. A proper understanding of this will be relevant not only to psychological explanation, but potentially also to constitutive issues about what it is to have the externally–individuated capacities which are explained.

The second path worth following is one which we can hope to give us further understanding of the role of content–involving computation in connectionist networks. It is becoming plausible that we need to attribute content both to large–scale, and to some smaller–scale, patterns of activation in these networks, when our aim is to understand how they do what they do. Outside the classical realm where a computation has genuinely syntactic properties, the classical theory of computation in Turing, Church and Kleene gives us little guidance on how to conceive of computation. We need a theory which gives a general apparatus for describing in content–involving terms the computational activity of a network which carries out a particular task. In developing it, we may need to consider different relations between contents, and perhaps different kinds of content, than those appropriate for the classical, syntactic cases. I am personally optimistic about the prospects for eventually developing such a theory. But until we have attained it, we cannot claim to have a full understanding of the nature and significance of content–involving computation.

A third line of enquiry salient from the point we have now reached begins with the reflection that the relations articulated here between content–involving computational explanation and the external explananda of psychology are instances of a general type which may be present in the other special sciences. In many of the other special sciences, and particularly in the social sciences, we can draw a form of the external/internal distinction. It is frequently the case that the distinctive explananda of a special science are externally individu-

ated, relative to the appropriate way of drawing the internal/external distinction for that science. When this is so, there is equally a question of how the objects or agents in the domain of the special science are able to be in externally–individuated states, for the appropriate internal/external distinction for the domain of that science. When the states of the special science are also representational states, there will also be questions about the explanation of their correctness or of their incorrectness. In all such cases, a generalization of the kind of explanation found in content–involving computational explanation seems appropriate. This suggestion raises a thousand questions. All I can do at this point is to commend this way of approaching some of the general, cross–domain issues about the nature of explanation.

References

Boden, M. *Computer Models of Mind: Computational Approaches to Theoretical Psychology.* Cambridge: Cambridge University Press, 1988.

Burge, T. "Individualism and the Mental". *Midwest Studies in Philosophy*, 4 (1979): 73-121.

Burge, T. "Individualism and Psychology". *Philosophical Review*, XCV (1986): 3-45.

Churchland, Patricia and T. Sejnowski. "Neural Representation and Neural Computation". In *Philosophical Perspectives, 4: Action Theory and Philosophy of Mind*, ed. J. Tomberlin. 1990.

Churchland, P.S. and T. Sejnowski. *The Computational Brain.* Cambridge, Mass.: MIT Press, 1992.

Davidson, D. "Radical Interpretation". In *Inquiries into Truth and Interpretation*, Oxford: Oxford University Press, 1984.

Davies, M. "Externality, Psychological Explanation and Narrow Content". *Proceedings of the Aristotelian Society Supplementary Volume* 60 (1986): 263-83.

Dennett, D. "Artificial Intelligence as Philosophy and as Psychology". In *Brainstorms*, Montgomery, Vt.: Bradford Books, 1978a.

Dennett, D. *Brainstorms.* Bradford Books, 1978b.

Dennett, D. *The Intentional Stance.* Cambridge, Mass.: MIT Press, 1987.

Dennett, D. *Consciousness Explained.* Boston: Little, Brown and Company, 1991.

Fodor, J. "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology". In *The Nature of Mind*, ed. D. Rosenthal. 485-98. New York: Oxford University Press, 1991.

Grandy, R. "Reference, Meaning and Belief". *Journal of Philosophy*, LXX (1973): 439-52.

Hornsby, J. "Physicalist Thinking and Conceptions of Behaviour". In *Subject, Thought and Context*, ed. P. Pettit and J. McDowell. Oxford: Clarendon Press, 1986.

Kitcher, Philip. *Abusing Science: The Case Against Creationism.* Cambridge, Mass.: MIT Press, 1982.

Lackner, J. and M. Garrett. "Resolving Ambiguity: Effects of Biasing Context in the Unattended Ear". *Cognition*, 1 (1972): 359-72.

Lewis, D. "Causal Explanation". In *Philosophical Papers*: Volume II, New York: Oxford University Press, 1986.

Marr, D. *Vision.* San Francisco: Freeman, 1982.

Marslen–Wilson, W. and L. Tyler. "Against Modularity". In *Modularity in Knowledge Representation and Natural–Language Understanding*, ed. J. Garfield. Cambridge, Mass.: MIT, 1987.

McDowell, J. "Functionalism and Anomalous Monism". In *Actions and Events: Perspectives on the Philosophy of Donald Davidson*, ed. E. LePore and B. McLaughlin. Oxford: Blackwell, 1986.

Peacocke, C. "Demonstrative Thought and Psychological Explanation". *Synthese* 49 (1981): 187-217.

Peacocke, C. *Sense and Content: Experience, Thought and their Relations.* Oxford: Oxford University Press, 1983.

Peacocke, C. "Externalist Explanation". *Proceedings of the Aristotelian Society* XCIII (1993): 203-230.

Putnam, H. "The Meaning of 'Meaning'". In *Mind, Language and Reality*, Cambridge: Cambridge University Press, 1975.

Ruben, D. "Introduction". In *Explanation*, ed. D. Ruben. Oxford: Oxford University Press, 1993.

Stich, S. "Paying the Price for Methodological Solipsism". In *The Nature of Mind*, ed. D. Rosenthal. 499-500. New York: Oxford University Press, 1991.

PHILOSOPHICAL ISSUES, 6
Content, 1995

Can We Knock Off the Shackles of Syntax?

Daniel Andler

Professor Peacocke's clearly stated goal is to reconcile two perspectives on cognition which seem to contradict one another, while both exerting considerable appeal on many philosophers of mind. One is 'solipsism' in the sense made current by Fodor's 1978 paper: a computational system cannot be sensitive to the content of what it computes, hence if the mind is such a system, then the basic workings of the mind are content–free. The other view is what the most casual inspection of the products of today's 'computational' cognitive psychology reveals, viz. that content enters practically everywhere in what passes as explanations —or models— of cognitive processes, whether 'personal' or 'subpersonal'.

I should say at the outset that I am in basic agreement with the general direction of the paper. Yet there is something disarming about this way of putting the problem. It seems to desecrate what we, who have chosen to play a certain kind of philosophical game, have been taught to regard as the Great Problem. We have been led through the often painful process of convincing ourselves that —roughly put— the mind is (according to the working hypothesis which in accepting to play the game we chose to take seriously) a

'syntactic' machine which 'mimics' a 'semantic' one; and that therefore what cognitive scientists and philosophers are paid to do is to find a way, or ways, of making sense of the 'mimicking' involved and decide whether among the conceptual possibilities, one happens to have empirical validity. We have waivered between Stich's refusal to accept any form of semantics (within the computational framework) and Searle's *modus tollens* on the Stichian syllogism. We have read hundreds of pages full of imagination and subtlety purporting to show the tenability of some third way, have let Fodor *guide* us through the maze, and Dennett *contrast* the main positions. We have listened to patient pedagogues, from Haugeland to Cummins, from Simon & Newell to Johnson–Laird, explaining how it's all supposed to hang together, and how difficult it is to find arguments, let alone decisive ones, bearing on the feasibility of the project.

Now Christopher Peacocke comes along and seems to be telling us that our perplexity's all in vain, that we need not worry: we can have computation *and* content; all we need to do is relax our misguided anti–content puritanism regarding computation, open our eyes and *see* that a content–involving computation is a coherent notion. There is a Gordian knot feeling to the proposal. All the more so as Peacocke goes to some trouble to rid us for sure of any suspicion (in the unlikely case that we should entertain some) that he is naive and wants to do what others have shown can't be done —'[his] view [...] does not in any case involve anything's executing the impossible task of getting semantics from syntax', he assures us for example (§6); 'that we should aim to identify the content–free mechanisms which make cognition possible is certainly something which [he] is not doubting' (§5), etc.

Is Peacocke then simply denying that psychosemantics raise hard problems, that content is virtually assimilated to the furniture of the natural world, yet retains some specific quality which sets it apart from the sort of thing that might constitute, or contribute to, the syntax of some natural system? To someone who hadn't read the paper with care, that would be one way of attempting to make sense of the enterprise: computational systems manipulate natural entities, but they certainly can distinguish types of information, just as they distinguish, for example, linguistic from visual input. Content would be some special modality of material stuff. But this is absurd on at least two grounds: by definition, anything that a computational system manipulates in the usual sense *is* syntax, hence analytically *not* content; and in virtue of the author's central position in the discussion of contemporary issues in philosophy of mind,

he is unlikely to underestimate the importance and difficulty of 'naturalizing' content.

A less implausible way of 'co–opting' the proposal without changing one's adherence to the standard view of the computational enterprise might go as follows: solipsistic accounts are not enough; whatever sense one can make of Fodor's judgment that 'semantics does not belong to the domain of psychology', psychologists and other cognitive scientists do see content as falling within their purview, both in explanda and in explanantia; and the way to account for that brute fact is to construe the scientist's contentful talk as compact notation for two–factor processes, involving syntactic operations supplemented by environmental constraints —the factoring out being in many cases a task which remains on the agenda. But Peacocke explicitly rejects such a theory, (end of §1), so he can hardly be thought to be meaning to defend it in his own sophisticated way.

Better luck awaits the reader who focuses on the numerous examples which are helpfully laid out all along the way, and on the arguments. For the sake of brevity, I will single out what strikes me as the main strands of the text.

a. Much of it is devoted to what amounts to no more (and no less) than a sophisticated defense of functionalism: Peacocke again and again aims at invisible or visible eliminativists and type–type identity–theorists, as when he (with good reason) insists that 'syntactic' accounts fail on counterfactual tests, or that different 'syntactic' events can underwrite the same contentful event. Although (surprisingly I should think) he never uses the word, he convincingly pleads for the autonomy of information.

b. The examples all go to show that explanatory accounts in psychology tend to invoke content: scientists are concerned with explaining gestures and sayings rather than mere bodily movements and mouthings, and they do so typically by invoking beliefs, inferences, desires, measured magnitudes, perceptions, etc. The striking thing about these examples is that, while quite reasonable taken separately, taken together they constitute a rather incongruous set. My intuition is that the paper basically stands or falls according to whether we are able or not, all things considered, to overcome the feeling of strangeness and see the important commonality which had escaped us till now. Let me first state the incongruity and then try and say a few words about its possible fecundity.

Peacocke's examples fall into two broad categories: those which involve an unproblematic sort of content (type A), and those which involve a problematic sort (type B). Visuo–motor neurophysiology provides a host of type A examples; neurophysiologists experience no

conceptual qualms when they present an account of spatial orientation resting on complex computations over data provided by captors in the inner ear and proprioceptive information stemming from the eyeball muscles; the contents of the various states involved, whether directly or indirectly linked to the stimuli, would seem to raise no difficult, at least no difficulty which would not equally beset for example the explanation of the functioning of a mobile robot. The same goes for research in psycholinguistics bearing on phonological aspects of speech perception. Type B examples, on the other hand, typically involve the grasping of meaning, as in linguistic theory and empirical psycholinguistics involving speech understanding, or in the psychology of reasoning. Whether this —completely obvious— distinction corresponds roughly or closely to the personal/subpersonal distinction, or again to the input systems/central systems one, is not an issue of immediate concern. Rather, it would seem that by refusing to take it into account, one is shutting oneself into the Fodorian jail one was seeking to escape from. At any rate, it simply goes counter to fact to take scientists working on B–type problems to think of themselves as both adhering to the classical computational paradigm and helping themselves to a free notion of content. On the contrary, they subscribe to Dennett's image of a cascade of loans which eventually must be paid back. They may have to content themselves for the time being to show that, on some temporarily acceptable notion of mechanism, the processing which takes place results from the application of a mechanism to some contentful input; but the ultimate goal they conceive as replacing the temporary notion by a well–founded, literally (narrowly) computational one.

To take an extreme example: Peacocke writes that because sentences enter psychological explanations of thought and action mainly as meaning–bearers, it cannot be the case that what these explanations seek is to give a syntactic account; such a claim seems to rest on a confusion between two meanings of semantics (say, between natural–language and language of thought semantics): formal semantics is formal, in the exact same sense as syntax is, *and* it does raise a problem, which is to account for the graspings by the agent with the sole help of syntactic happenings within the agent (plus 'hook–up' conditions). *Of course* any explanation *whatsoever* of the grasping of the meaning of a sentence must appeal to the meaning of the parts! It is equally obvious —in fact, definitional— that any causal explanation will appeal to instantiated computational processes, and that the'syntax' appealed to in functional explanations covers, under a suitably abstract description, all and every relevant (*i.e.* causal) aspect of the physical make up of the system.

But just repeating the old stuff may look silly, and perhaps it is of no help here: agreed, fighting about *definitions* is useless, and after all, isn't it Peacocke's aim to show that computationalism need not stick to its 'puritanical' roots, and that the cognitive scientist who sees that he uses content but thinks that he shouldn't, at least in the long run, is unnecessarily hard on himself? Isn't the point precisely that such a scientist is misconceiving his real task? (And certainly this would not be the first time in the history of science that an essentialist illusion is finally put to rest.) But at least the contrast with type A situations must be acknowledged: for there no problem whatsoever presents itself. There is no loan, there is no need to worry about reducing semantics to syntax: the reduction is performed and operates unmysteriously.

c. There is however a very different perspective on the matter to be gained by considering an even broader class of situations, one which includes both types A and B, and much more. This class may indeed be a theoretical natural kind. It is constituted by processes whose very individuation relies on relational properties. By characterizing these situations and showing that the sort of explanations they give rise to is distinctive and nonetheless quite legitimate, Peacocke is launching a wide–range theoretical entreprise in the philosophy of science which, as far as I understand it, can only be met with great interest and raises no objection originating specifically in the philosophy of cognitive science. Quite the contrary, fitting the kind of explanation forthcoming from cognitive science within the broad category of 'externalist explanations' holds the promise of easing, quite as Peacocke says, a definite tension between two views of the explanatory goals of cognitive science.

The tension arises, quite as Peacocke says, due to a discrepancy between some philosophers' view of the nature of cognitive–scientific explanation and most cognitive scientists' explanatory practice. However as I see it, the bone of contention is not constituted by a definite view held by one party and contested by the other. Rather, it stems from the discrepancy between a definite view leaving no essential further choices to be made, on one hand, and on the other a deliberately incomplete (as well, of course, as very often mostly implicit) theory. In other words, I believe that most cognitive scientists are agnostic or undecided on the way in which their present results will eventually find their way into a complete or near–complete framework. Some are prepared to place a bet on Turing functionalism in the strict sense, some are eliminativists at heart, some may be willing to explore Peacocke's suggestion, some are seeking more radical alternatives. But most, I submit, are unwilling to speculate.

Let me at this point briefly contrast a weak and a strong reading of Peacocke's proposal. On the weak reading, emphasis is laid on easing the pressure exerted on cognitive psychologists by the set and stringent views of (some) philosophers: though admittedly incomplete, their work is not threatened with illegitimacy should they fail to find a way, before Doomsday, of fitting it in the framework of classical computationalism. Nor are they forced to take a stand on that sort of completion being *im*possible: Peacocke, thus read, yields an elegant way of presenting the 'recursive loan' paradigm which Dennett defends in the 1978 paper which he quotes. Psychologists, on this view, may continue —with a clear conscience— to seek mechanisms operating directly on the level of content, relying on an *informal* sense of mechanism: something which does not seem to call on mysterious capacities, and which may well be executable by a rather silly agent or system (man–made, for instance). In other words, under this interpretation the notion of 'content–involving algorithm' is not taken too seriously; it entails no obligation to scrutinize the notion and discover some natural kind of processes parallel to the kind 'algorithm' in the Turing sense.

The strong reading, by contrast, requires us to take the notion very seriously indeed, and is an appeal to reform the methodological code of cognitive science. Ontologically, Peacocke remains classical: he holds on to the syntax/semantics parallelism (construed in such a way as to make room for his externalist intuitions), but he wants to reorganize the explanatory structure of the field.

This is where I am tempted to part company with him (assuming of course that I have grasped the idea correctly). Siding with such authors as Perry and Barwise, I tend to favor the idea that cognitive science deals with a stuff called information without having much of a clue about what information is; content–involving explanations are in this sense informational explanations and insofar as they are produced in a spirit of informal rigor and conceptual sparseness, they form a legitimate body of scientific work. Although this view is logically compatible with classical computationalism, its thrust is directed against such a conservative stand. Information fits uncomfortably, as it turns out, in a world centered on the syntax/semantics divide, classically construed.

But backing out of a premature commitment to classical computationalism does not force us to face total reconstruction of the work accomplished under the notional guidance of that doctrine: this is the beauty of a stance such as Perry's and Barwise's. And, as I would prefer to interpret it, such as Peacocke's.

PHILOSOPHICAL ISSUES, 6
Content, 1995

Content Preservation*†

Tyler Burge

Near the beginning of *Rules for the Direction of the Mind* Descartes holds that some things known "with certainty" and "by deduction" are not evident. He notes that in long deductions, we may know that "the last link is connected with the first, even though we do not take in by means of one and the same act of vision all the intermediate links on which that connection depends, but only remember that we have taken them successively under review...".[1] Though he acknowledges that such knowledge is not evident or purely intuitive, and that long deductions are more subject to error than is intuitive knowledge, Descartes thinks that if the knowledge is deduced from evident mathematical premises, it is certain and demonstrative. Presumably he would not doubt that it is apriori. I lay aside certainty.

---

*I am indebted to Tony Anderson, Hilary Bok, Larry Bonjour, Robert Brandom, Michael Bratman, Robin Jeshion, Bill Hart, Bernie Kobes, Ruth Marcus, Stanley Munsat, Christopher Peacocke, W. V. Quine, Corliss Swain, Fred Stoutland, and the editors for valuable remarks.

†Reprinted by kind permission of *The Philosophical Review.*

[1]Descartes, *Philosophical Works*, ed. Haldane and Ross (New York: Dover, 1955), vol. 1, p. 8. Locke, in *Essay Concerning Human Understanding*, bk. 4, chap. 2, sec. 7, notes that such knowledge is "less perfect" in the sense of more subject to error than intuitive knowledge.

But the view that the knowledge is demonstrative and apriori seems to me true.

Roderick Chisholm sees matters differently. He defines 'apriori' in such a way that a proposition is apriori (and known apriori) only if it is either evident or follows directly by evident entailment from something that is evident. He explicitly rules out the results of multistepped deductions:

> What if S derives a proposition from a set of axioms, not by means of one or two simple steps, but as a result of a complex proof, involving a series of interrelated steps? If the proof is formally valid, then shouldn't we say that S knows the proposition a priori? I think that the answer is no.
>
> [I]f, in the course of a demonstration, we must rely upon memory at various stages, thus using as premisses contingent propositions about what we happen to remember, then, although we might be said to have "demonstrative knowledge" of our conclusion, in a somewhat broad sense of the expression "demonstrative knowledge", we cannot be said to have an a priori demonstration of the conclusion.[2]

Some of the difference between us derives from different conceptions of apriority. There are many such conceptions. I will be explicit about mine. I understand *'apriori'* to apply to a person's knowledge when that knowledge is underwritten by an apriori justification or entitlement that needs no further justification or entitlement to make it knowledge. A justification or entitlement is *apriori* if its justificational force is in no way constituted or enhanced by reference to or reliance on the specifics of some range of sense experiences or perceptual beliefs.

I take 'apriori' to apply primarily to justifications or entitlements, rather than to truths. There are, of course, conceptual relations between these notions. Justification or entitlement aims at truth since it rationally supports belief. Moreover, the notion of apriori truth is important, though it should probably be explicated in terms of possible apriori knowledge. But in this account, justification and entitlement are fundamental.

The distinction between justification and entitlement is this: Although both have positive force in rationally supporting a propositional attitude or cognitive practice, and in constituting an epistemic right to it, entitlements are epistemic rights or warrants that need

---

[2]Roderick M. Chisholm, "The Truths of Reason", in *Theory of Knowledge*, 2d ed. (Englewood Cliffs, NJ.: Prentice Hall, 1977), reprinted in *A Priori Knowledge*, ed. Paul K. Moser (Oxford: Oxford University Press, 1987).

not be understood by or even accessible to the subject. We are entitled to rely, other things equal, on perception, memory, deductive and inductive reasoning, and on —I will claim— the word of others. The unsophisticated are entitled to rely on their perceptual beliefs. Philosophers may articulate these entitlements. But being entitled does not require being able to justify reliance on these resources, or even to conceive such a justification. Justifications, in the narrow sense, involve reasons that people have and have access to. These may include self–sufficient premises or more discursive justifications. But they must be available in the cognitive repertoire of the subject. The border between the notions of entitlement and justification may be fuzzy. I shall sometimes use 'justified' and 'justification' broadly, to cover both cases.

A person's knowledge of a proposition might be adequately supported both by an apriori body and by an empirical body of justification or entitlement. Then the person's knowledge would be heterogeneously over determined. The person would have both apriori and empirical knowledge of the proposition. To be apriori, the knowledge must be underwritten by an apriori justification or entitlement that needs no further justificatory help, in order for the person to have that knowledge. To be apriori, a person's justification or entitlement must retain its justificational force even if whatever empirical justifications or entitlements the person also has to believe the relevant proposition are ignored.

In holding that the justificational force of an apriori justification or entitlement is in no way constituted or enhanced by reliance on the specifics of some range of sense experiences or perceptual beliefs, I do not require that an apriori justification rely on reason or understanding alone —as pre–Kantian rationalists required. A justification or entitlement would count as apriori if it did not rely for its justificational force on sense experience or perceptual belief at all. But it might also count if it depended on entirely general aspects of sense experience or perceptual belief, or on aspects of the structure of the subject's sense capacities and on their function in yielding categories of information.[3]

---

[3]Kant thought that all synthetic apriori judgments, except those in his practical philosophy —and perhaps in the critical philosophy as a whole— rested on general ("pure") aspects of the structure or function of sense experience. In fact, he believed that the justificational force of all such judgments depended on one s actually having had sense experiences. My conception of apriori knowledge makes room for Kant's conception. I do not, however, agree with Kant that those apriori justifications whose justificational force is not enhanced at all by sense ex-

An individual need not make reference to sense experiences for his justification or entitlement to be empirical. My term 'reliance on', in the explication of apriority, is meant to acknowledge that most perceptual beliefs about physical objects or properties do not refer to sense experiences or their perceptual content. Such beliefs make reference only to physical objects or properties. But the individual is empirically entitled to these perceptual beliefs. The justificational force of the entitlement backing such beliefs partly consists in the individual's having certain sense experiences, or at any rate in the individual's perceptual beliefs' being perceptual.

An apriori justification (entitlement) cannot rely on the specifics of sense experiences or perceptual beliefs for *its justificational force.* An apriori justification will usually depend on sense experiences or perceptual beliefs in some way. They are typically necessary for the acquisition of understanding or belief. But such dependence is not relevant to apriority unless it is essential to justificational force. Distinguishing the genesis of understanding and belief from the rational or normative force behind beliefs is fundamental to any view that takes apriori justification seriously.[4]

---

perience are vacuous, or analytic in the sense of being true independently of any relation to a subject matter. I he distinction between reliance on the specifics of a range of sense experiences, or perceptual beliefs, and reliance on the structure or function of one s sense capacities in obtaining categories of information is not sharp. I think it may remain useful.

[4]This explication of apriority applies to justification of cogito–type thoughts like *I am thinking*, and of other judgments about intellection. (It does not apply to *I am having an afteritmage.*) These thoughts' justification is grounded on understanding, not on sense experience or perceptual belief. I am aware that some traditional conceptions of apriority would exclude *cogito* cases. Some of these conceptions emphasize not justificational independence of sense experience, but justificational independence of any "experience" at all, including intellectual "experience". (I leave open here whether this use of 'experience' is appropriate.) This is one of Leibniz's conceptions (see *New Essays*, IV, *ix*). Of course Leibniz centered on apriori truth rather than on an individual's justification. Frege's conception features justificational independence of any relation to particular events or facts in time (see Gottlob Frege, *The Foundations of Arithmetic*, sec. 3). On his conception, only general truths and truths derivable from general truths could be known apriori.

The terminological issues here are complex; but this difference with traditional explications will not affect my argument with Chisholm, which goes through on any of these conceptions. Moreover, the broader argument of the paper does not depend on how one uses the term 'apriori'. I am less interested in the term than in the conception I associate with it. The argument of the paper hinges on the role of perception in justification or entitlement. I do think that there are significant substantive and historical issues regarding these different notions associated with the term 'apriori' that bear on the way the issue between empiricism and rational-

No serious conception of apriority has held that all justifications held to be apriori are unrevisable or infallible. Traditionally, the deepest apriori justifications were seen to be hard to come by. Putative apriori justifications were traditionally held to be revisable because one could fail to understand in sufficient depth the relevant propositions, or make errors of reasoning or analysis.

Traditional views did tend to overrate the tightness of connection between genuine (as opposed to putative) apriori justifications and truth. First, apriori justification (entitlement) can be nondemonstrative: an apriori justification can be outweighed without being shown to be rationally deficient or based on misunderstanding — without being shown not to have justificational force (not to be a justification). Some mathematical arguments are nondemonstrative, even broadly inductive, yet apriori in my sense. If a principle is accepted because its truth would explain or derive a variety of other accepted mathematical principles, the justification for accepting the principle is nondemonstrative; but it may not derive any of its force from perceptual beliefs. Second, although some apriori justifications or entitlements may be invulnerable to empirical counterconsiderations, such invulnerability does not follow from the notion of apriority. As will emerge, I think that some beliefs with genuine apriori justifications or entitlements are vulnerable to empirical overthrow.

In both ways, a belief's being apriori justified, for a person at a time, does not entail that it is true. There are, I think, some apriori justifications or entitlements that are demonstrative and do entail truth. But they do not do so purely by being apriori. The present conception of apriority fixes on the nature of the positive rational support for a belief. It says nothing about ways in which a belief may be vulnerable to counterconsiderations.

Thus apriori justification may be unevident, fallible, nondemonstrative, and not "certain". Beliefs thought to be apriori, and even actually justified apriori, are subject to revision. In these ways, my conception of apriority differs from Chisholm's.

Our differences are not primarily verbal, however. Chisholm regards long deductions as importing memory of particular past mental events into the justificational of the deduction.[5] If such memories

---

ism has come to be understood since the work of Kant, Mill, and the positivists. For now, it is enough that the present explication signals my interest in justifications or entitlements whose force is grounded in intellection, reason, or reflection, as distinguished from perception, understood broadly to include feeling.

[5]Descartes's own remark that in deductions we must remember that we have taken the links of the deduction "successively under review" may suggest this view. I find it unclear how he intended the remark.

are a necessary part of the justification of the deduction, then —at least where they include memories of empirical beliefs or experiences (memories of reading symbols carefully, for example)— such deductions are not apriori, even on my conception of apriority.

But Chisholm's conception of the *role* of memory in demonstrative reasoning seems to me off the mark. If memory supplied, as part of the demonstration, "contingent propositions about what we happen to remember", the demonstration could not be purely logical or mathematical. But the normal role of memory in demonstrative reasoning is, I think, different. Memory does not supply for the demonstration propositions about memory, the reasoner, or past events.[6] It supplies the propositions that serve as links in the demonstration itself. Or rather, it *preserves* them, together with their judgmental force, and makes them available for use at later times. Normally, the content of the knowledge of a longer demonstration is no more about memory, the reasoner, or contingent events than that of a shorter demonstration. One does not justify the demonstration by appeals to memory. One justifies it by appeals to the steps and the inferential transitions of the demonstration.

Why did Chisholm think otherwise? Long demonstrations are more fallible, and fallible in different ways, than short ones are. As he notes, people make mistakes of haste or incomplete understanding in judgments about relatively obvious propositions. But in longer demonstrations there are not only more opportunities to make these mistakes. One may suffer memory slips, even if one is careful and fully understands each proposition in the deduction. Traditionally, belief that appealed to apriori justification was held to be subject to error. But the sources of error limited to failures of understanding and reason. It may seem that failure of memory, is a source of error not easily accommodated by the traditional conception.

But relevant differences between short and long demonstrations are at most those between short–term and long–term memory. Even onestep demonstrations could go bad if the reasoner's short–term memory were defective enough. So if we take vulnerability to memory failure as a sign that a justification of reasoning must *make reference to* memory, no reasoning at all will be independent of premises about memory. This is unacceptable. It is one thing to rely on memory in a demonstration, and another to use premises about memory.

---

[6]Chisholm's "thus", in the quoted passage, is clearly a mistake. It does not follow from a deduction's reliance on memory that it, or any justification associated with it, uses "contingent propositions" about memory as premlses.

Any reasoning in time must rely on memory. But not all reasoning must use premises about memory or the past.

Here as elsewhere, to be justified in a cognitive process, one need not include premises in the justification that rule out all possible sources of error. This is a widely accepted point about perceptual justification. To be entitled to a perceptual belief that there is a bird there, one need not rule out all ways that one could be fooled. The same point applies to reasoning. To be justified in deductive reasoning, one need not include in one's justification propositions that guard against memory lapses, short or long term. Reliance on memory does not even add to the justificational force of the deductive justification.

If a justification depends on valid deductive reasoning from (let us presume) premises that are known apriori, then one's being justified by the justification depends only on one's *actually* understanding the reasoning sufficiently, and on one's reasoning processes' *actually* working properly. The justification does not depend on a premise that says that these conditions obtain, a premise that would itself require further justification. (I think that such dependence would involve a vicious regress.) One can presume that they obtain, without needing justification for the presumption, except in special situations in which these presumptions are called reasonably —and perhaps even correctly— into question.

In a deduction, reasoning processes' working properly depends on memory's preserving the results of previous reasoning. But memory's preserving such results does not add to the justificational force of the reasoning. It is rather a background condition for the reasoning's success. Memory is no more intrinsically an empirical faculty than it is a rational faculty. Its function in deductive reasoning is preservative. Its role in justification derives from what it preserves. Our entitlement to rely on memory in long deductions derives from our entitlement to rely on reasoning to carry out its functions. Memory failures that cause demonstrations to fail are failures of background conditions necessary to the proper function of reasoning. Hence the fallibility of memory in deductive reasoning is a source of error that can be countenanced by the traditional conception of apriority —and our conception as well.

Even in empirical reasoning, memory has a purely preservative function that does not contribute to the force of the justification, but simply helps assure the proper working of the capacities over time. When we perceive events and infer an explanation, memory preserves the perceptual beliefs as we carry out the explanation. But this preservation is not part of the justification of the explanation,

nor does it add to it —even though if it were to fail, the explanation would be jeopardized. Rather, memory just holds the results of the perception intact long enough for explanation to be carried through.

Of course, memory sometimes is not purely preservative, but is an independent element in justification. Memory of events, objects, experiences, or attitudes may form a premise in a justification of an empirical belief. The beliefs that such memories support are justified partly by *reference to* the memory. Or else they may partly rely for their entitlement on memory.

Substantive memories of specific events, objects, experiences, or attitudes may play a role in deductive reasoning. They may aid reasoning without being elements in the justification they aid. So, for example, we may draw pictures in a proof, or make use of mnemonic devices to aid understanding and facilitate reasoning, without relying on them to enhance the mathematical justification. Alternatively, substantive memories may be part of an auxiliary, double–checking justification. In such cases, they may play a justificational role, yet be justificationally dispensable.

Substantive memory can even be needed to shore up gaps in a person's deductive reasoning. When a purely preservative instance is reasonably challenged, because memory has proved unreliable, one may have to rely on substantive memory. For example, if one knows one's memory has been slipping, one might have to resort to remembering counting the number of implication signs in a pair of formulas to support one's presumption that one's inference was based on correct memory. In such a case, reliance on the mnemonic devices may be indispensable to the person's justification —not merely a part of an auxiliary doublechecking procedure. For the person is no longer entitled to the presumption that memory can be relied upon. I think, however, that the need to make reference to memory in deductions in order to be justified by the deductions is uncommon. In certain cases one might reasonably doubt that one is entitled to rely on one's memory, but be wrong to doubt it.

But the fact that memory can play substantive roles in justification or entitlement should not obscure the distinction between substantive and purely preservative memory. Let me summarize the distinction. Substantive memory is an element in a justification; it imports subject matter or objects into reasoning. Purely preservative memory introduces no subject matter, constitutes no element in a justification, and adds no force to a justification or entitlement. It simply maintains in justificational space a cognitive content with its judgmental force. Like inference, it makes transitions of reason possible, but contributes no propositional content. Unlike inference,

it is not a transition or move —so it is not an element in a justification. Hence in deductions, neither reliance on it nor susceptibility to errors that arise from its malfunction prevents the justification associated with the deduction from being apriori.[7]

My discussion of memory is pointed toward exploring analogies between memory and acceptance of the word of others. What is the role of interlocution in the justification of our beliefs?

Relying on others is perhaps not metaphysically necessary for any possible rational being. But it is cognitively fundamental to beings at all like us. Though ontogenetically later than perception and memory, reliance on others for learning language and acquiring beliefs is deeply ingrained in our evolutionary history. Acquiring beliefs from others seems not only psychologically fundamental, but epistemically justified. We do not as individuals justify this reliance empirically, any more than we justify our use of perception empirically. But we seem entitled to such reliance. Most of the information that we have, and many of the methods we have for evaluating it, depend on interlocution. If we did not acquire a massive number of beliefs from others, our cognitive lives would be little different from the animals'.

What is the epistemic status of beliefs based on interlocution? I will state my view broadly before qualifying and supporting it. The use of perception is a background condition necessary for the acquisition of belief from others. But in many instances, perception and perceptual belief are not indispensable elements in the justification of such beliefs, or in the justificational force of entitlements underwriting such beliefs. The function of perception is often analogous to the function of purely preservative memory in reasoning. Without perception, one could not acquire beliefs from others. But percep-

---

[7]The distinction between substantive memory and purely preservative memory roughly parallels a distinction in psychology between "episodic memory" and "semantic memory". There is evidence that these sorts of memory function differently in our psychologies. See E. Tulving, "Episodic and Semantic Memory", in *Organization of Memory*, ed. Tulving and Donaldson (New York: Academic Press, 1972).

Another difference between the two types of memory is that purely preservative memory necessarily plays a role in any reasoning in time. The extent to which substantive memory enters into reasoning depends on the psychology of the reasoner, the subject of the argument, and so on. One should not underestimate, however, our dependence on the use of symbols in reasoning. The role of symbols is partly that of providing perceptual objects. Explicating this sort of dependence is a difficult and important matter. Doing so may complicate or blur the distinction between the sometime dependence on substantive memory and the more general rational necessity of depending on purely preservative memory. But I think that the distinction will remain valuable.

tion plays a triggering and preservative role, in many cases, not a justificatory one. Sometimes, the epistemic status of beliefs acquired from others *is not empirical.* In particular, it is not empirical just by virtue of the fact that the beliefs are acquired from others.[8] Such beliefs are sometimes apriori justified in the sense that they need not rely for justificational force on the specifics of some range of sense experiences or perceptual beliefs.

Thomas Reid insightfully compares acquisition of belief from others to perception as a basic "channel to the mind", with its own functions in acquiring knowledge. Reid also claims that the tendency to rely on others for acquiring beliefs is innate:

> The wise and beneficent Author of nature, who intended that we should be social creatures, and that we should receive the greatest and most important part of our knowledge by the information of others, hath, for these purposes implanted in our natures two principles that tally with each other. The first of these principles is a propensity to speak truth... [the second] is a disposition to confide in the veracity of others, and to believe what they tell us.[9]

Reid notes that credulity, unlike reasoning and experience, is "strongest in childhood, and limited and restrained by experience". We restrain credulity by weighing the character and disinterestedness of witnesses, the possibility of collusion, the antecedent likelihood of information. Moreover, our reliance on others is more fallible than our reliance on perception —as Reid also notes. We make perceptual errors, but the errors derive from illusions that often can be explained by reference to natural law. We are led into mistakes by others through lies and emotional interferences that are

---

[8]Contrast Chisholm, "The Truths of Reason", sec. 5, and James F. Ross, "Testimonial Evidence", in *Analysis and Metaphysics*, ed. Keith Lehrer (Dordrecht: D. Reidel, 1975). They assume that belief based on testimony cannot be justified apriori and, if it is knowledge at all, must be empirical.

I think that some of what I am saying here bears on the common assumption that knowledge based on the output of proofs by computers cannot be apriori. Cf. Kripke, *Naming and Necessity* (Cambridge: Harvard University Press, 1980), 35; also Thomas Tymoczko, "The Four–Color Problem and its Philosophical Significance", *Journal of Philosophy* 76 (1979): 57-83. Kripke says that such knowledge is based on the laws of physics. Although such knowledge depends on the functioning of a machine according to the laws of physics, it is not obvious that knowledge of the laws of physics is an indispensable part of our justification for believing in the results of such output. I discuss this issue in "Computer Proof and Apriori Knowledge" (in preparation).

[9]Thomas Reid, *An Inquiry into the Human Mind* (Chicago: University of Chicago Press, 1970), chap. 6, sec. 24.

capricious in comparison to the patterns of nature. Why do these considerations not show that acquisition of beliefs from others is not only necessarily empirical but far more in need of empirical expertise than ordinary perception for its justification?

Justification in acquiring beliefs from others may be glossed, to a first approximation, by this principle: A *person is entitled to accept as true something that is presented as true and that is intelligible to him, unless there are stronger reasons not to do so.* Call this the *Acceptance Principle.* As children and often as adults, we lack reasons not to accept what we are told. We are entitled to acquire information according to the principle —without *using* it as justification— accepting the information instinctively. The justification I develop below is a reflective philosophical account of an epistemic entitlement that comes with being a rational agent.

Justified (entitled) acceptance is the epistemic "default" position. We can strengthen this position with empirical reasons: "she is a famous mathematician". We can acquire empirical reasons not to accept what we are told: "he has every reason to lie". But to be entitled, we do not have to have reasons that support the default position, if there is no reasonable ground for doubt. Truth telling is a norm that can be reasonably presumed in the absence of reasons to attribute violations.

It is usually said that to be justified in accepting information from someone else, one must be justified in believing that the source believes the information and is justified in believing it. I think this misleading. A presupposition of the Acceptance Principle is that one is entitled not to bring one's source's sincerity or justification into question, in the absence of reasons to the contrary. This too is an epistemic default position.

The Acceptance Principle is not a statistical point about people's tending to tell the truth more often than not. Falsehoods might conceivably outnumber truths in a society. The principle is also not a point about innateness, though Reid's claim that a disposition to acceptance is innate seems to me correct. The principle is about entitlement, not psychological origin.

The epistemic default position articulated by the Acceptance Principle applies at an extremely high level of idealization in most actual communication, especially between sophisticated interlocutors. Social, political, or intellectual context often provides "stronger reasons" that counsel against immediately accepting what one is told. Given life's complexities, this default position is often left far behind in reasoning about whether to rely on a source. One might wonder, with some hyperbole, whether it can ever be the last word in the

epistemology of acceptance for anyone over the age of eleven. The primary point —that it is a starting point for reason— would not be undermined if its purest applications were relatively rare. But I think that it has broader application than the hyperbolic conjecture suggests.

Acceptance underlies language acquisition. Lacking language, one could not engage in rational, deliberative activity, much less the primary forms of human social cooperation. (Indeed, I suggests the line of justification for the principle that I shall begin to develop below.) But unquestioned reliance is also common in adult life. When we ask someone on the street the time, or the direction of some landmark, or when we ask someone to do a simple sum, we rely on the answer. We make use of a presumption of credibility when we read books, signs, or newspapers, or talk to strangers on unloaded topics. We need not engage in reasoning about the person's qualifications to be rational in accepting what he or she says, in the absence of grounds for doubt. Grounds for doubt are absent a lot of the time.

The primary default position, the Acceptance Principle, is not an empirical principle. The general form of justification associated with the principle is: *A person is apriori entitled to accept a proposition that is presented as true and that is intelligible to him, unless there are stronger reasons not to do so, because it is prima facie preserved (received) from a rational source, or resource for reason; reliance on rational sources —or resources for reason— is, other things equal, necessary to the function of reason.* The justificational force of the entitlement described by this justification is not constituted or enhanced by sense experiences or perceptual beliefs.[10] Before filling in this form of justification, I want to make some preliminary points.

I think that I need not show that other rational beings are necessary to the function of one's reason in order for one to have these entitlements. One has a general entitlement to rely on the rationality of rational beings. The Acceptance Principle can be apriori instan-

---

[10]Principles narrower than the Acceptance Principle could with luck and context achieve the same utility: rely on the first person one comes across and no one afterward. Such principles are not rational starting points. We are entitled to something more general. In learning a language, one usually need not know the credentials of one's source —beyond the fact that the source is intelligible. Having an apriori entitlement based on the Acceptance Principle is compatible with also having empirical justifications of prima facie acceptance —or of narrower principles, such as nonaggressive care givers are more trustworthy than strangers who threaten one". I think that one does not have to have these empirical justifications to be entitled to accept what one is told in particular cases (even though people do have such empirical justifications).

tiated where one has apriori, undefeated, prima facie entitlement to construe something prima facie intelligible as having a rational source. So I think that to maintain that one is apriori entitled to rely upon rational interlocutors, I need not show that a solitary reasoner is impossible.

Our account distinguishes rational sources and resources reason. Resources for reason —memory and perception, for example— need not themselves be rational beings or capacities to reason. In these senses they need not themselves be rational. Yet they may provide material and services that a rational being is apriori entitled to rely upon. Rational sources are sources that themselves are a capacity to reason or are rational beings.

As with rational sources, I think that to show that we are apriori entitled to rely upon a given resource for reason, I need not show that such a resource is necessary to any possible reasoning. One is entitled to rely upon resources for reason in general —other things equal— even if some particular resource for reason is not indispensable to the function of reason. Such resources may enrich reason without being necessary to every rational activity. This view puts pressure on explicating the notion of a resource for reason. This matter can be postponed, for it is relevant to interlocution only in special cases.

There are deeper questions about rational entitlement that I cannot pursue in depth here. One can ask why one is entitled to rely on rational sources (or resources for reason), in view of the fact that they can be mistaken or misleading. This is tantamount to a traditional skeptical question about how putative rationality or justification is associated with truth. One can apparently imagine systematic misconnections between being justified (entitled), according to ordinary canons, and having true belief. Why then should one ever think that ordinary canons provide ground for belief? I will not take on skepticism here. I will assume that we are rationally entitled to rely on reason, memory, and perception. The Acceptance Principle is an extension of this assumption: we are rationally entitled to rely on interlocution because we may presume that it has a rational source.

Now I turn to filling in the justification for the Acceptance Principle. First, if something is a rational source, it is a prima facie source of truth. For a condition on reasons, rationality, and reason is that they be guides to truth. Explicating this idea is notoriously difficult; but I do not apologize for it. An epistemic reason for believing something would not count as such if it did not provide some reasonable support for accepting it as true. The same point applies to rational entitlements for belief. The entitlements that I am discussing are epistemic, not matters of politesse. If one has a reason or entitlement

to accept something because it is, prima facie, rationally supported, one has a reason or entitlement to accept it as true. A source is a guide to truth *in* being rational. Rational mistakes are possible. But if there is no reason to think that they are occurring, it is rational to accept the affirmed deliverances of a rational source. For other things equal, reason can be reasonably followed in seeking truth.

It is not just the rationality of a source that marks an apriori prima facie connection to truth. The very content of an intelligible message presented as true does so as well. For content is constitutively dependent, in the first instance, on patterned connections to a subject matter, connections that insure in normal circumstances a baseline of true thought presentations. So presentations' having content must have an origin in getting things right. The prima facie rationality of the source intensifies a prima facie connection to truth already present in the prima facie existence of presented content.

The remaining main step in justifying the Acceptance Principle lies in the presumption that the source of a message is a rational source, or a resource for reason. I think that one is apriori prima facie entitled to presume that the interlocutor is a rational source or resource for reason —simply by virtue of the prima facie intelligibility of the message conveyed. That is enough to presume that the interlocutor is rational, or at least a source of information that is rationally underwritten.

The idea is not that we reason thus: "If it looks like a human and makes sounds like a language, it is rational; on inspection it looks human and sounds linguistic; so it is rational". Rather, in understanding language we are entitled to presume what we instinctively do presume about our source's being a source of rationality or reason. We are so entitled because intelligibility is an apriori prima facie sign of rationality.

If something is prima facie intelligible, one is prima facie entitled to rely on one's understanding of it as intelligible. One is entitled to begin with what putative understanding one has. But anything that can intelligibly present something as true can be presumed, prima facie, to be either rational or made according to a rational plan to mimic aspects of rationality. Presentation of propositional content presupposes at least a derivative connection to a system of perceptual, cognitive, and practical interactions with a world, involving beliefs and intentional activity.[11] Belief and intention in turn

---

[11] The expression may be derivative in that a nonrational machine might express linguistic content. But such machines are ultimately made by beings who have propositional attitudes.

presuppose operation under norms of reason or rationality —norms governing information acquisition, inference, and practical activity. For propositional attitudes, especially those complex enough to yield articulated presentations of content, are necessarily associated with certain cognitive and practical practices. To be what they are, such practices must —with allowances for some failures— accord with norms of reason or rationality.

To summarize: We are apriori prima facie entitled to accept something that is prima facie intelligible and presented as true. For prima facie intelligible propositional contents prima facie presented as true bear an apriori prima facie conceptual relation to a rational source of true presentations–as–true: Intelligible propositional expressions presuppose rational abilities and entitlements; so intelligible presentations–as–true come prima facie backed by a rational source or resource for reason; and both the content of intelligible propositional presentations–as–true and the prima facie rationality of their source indicate a prima facie source of truth.[12] Intelligible affirmation is the face of reason; reason is a guide to truth. We are apriori prima facie entitled to take intelligible affirmation at face value.

We could be apriori entitled to false beliefs. Sounds or shapes could have no source in rationality but seem intelligible. A quantum

---

[12]I think that the distinction between merely having attitudes with intentional content and being able to understand and present them is deeply significant, and marks a deeper level of rationality than that associated with merely having propositional attitudes and inferential abilities. But I need not explore this point here.

I have not here argued in depth for the connections between content, propositional attitudes, and rationality because they are a widely accepted theme in much contemporary work. The idea that language is inseparable from propositional attitudes, which are inseparable from assumptions about rationality is present, for example, in the work of Paul Grice, *Studies in the Way of Words* (Cambridge: Harvard University Press, 1989), and Donald Davidson, *Essays on Actions and Events* (Oxford: Oxford University Press, Clarendon Press, 1980) and *Inquiries into Truth and Interpretation* (Oxford: Oxford University Press, Clarendon Press, 1984). Elsewhere I have sought to show how having linguistic and propositional content is necessarily associated with individuals' having *de re* propositional attitudes to objects of reference and with their interacting practically and perceptually with such objects. See my "Belief De Re", *Journal of Philosophy* 74 (1977): 338-63, and "Other Bodies", in *Thought and Object*, ed. Woodfield (Oxford: Oxford University Press, 1982). The main novelty of the above argument lies in its first step —the claim that we are apriori entitled to rely on our understanding and acceptance of something that is prima facie intelligible— and in its drawing an epistemic consequence from the constitutive, conceptual relations between content and rationality that others have long explored and elaborated.

accidental sequence of sounds could correspond to those of Hamlet's most famous speech'.[13] But the fact that we could be mistaken in thinking that something is a message, or in understanding a message conveyed, is compatible with our having an apriori prima facie rational right to rely on our construal of an event as having a certain meaning or intentional content. And where a message has meaning or intentional content, we are entitled to presume apriori that it has a rational source, or is a resource for reason.

Just as the Acceptance Principle does not assume that truth is in a statistical majority, the justification of the Principle does not assume that most people are rational. We could learn empirically that most people are crazy or that all people have deeply irrational tendencies —not just in their performance but in their basic capacities. Human beings clearly do have some rational entitlements and competencies, even though we have found that they are surprisingly irrational in certain tasks. The justification presupposes that there is a conceptual relation between intelligibility and rational entitlement or justification, between having and articulating propositional attitudes and having rational competencies.

Rational backing is, other things equal, a ground for acceptance of something as true. But in dealing with others, one must often take account of their lies. Why is one *apriori* entitled, except when reasonable doubt arises, to abstract from the possibility that it may be in the interlocutor's rational interest to lie?

This issue is more complex than I can see through now. I will make some general observations, and then sketch one line of reply. (I think

---

[13]In *Dialogues Concerning Natural Religion*, part 3, Hume imagines hearing an "articulate voice" from the clouds and asks whether one can avoid attributing to it some design or purpose. He never objects to this inference, though he objects to much of the theological purposes it was put to. He would, however, regard it as a nonapriori causal inference. One of the reasons that he would invoke for thinking that the presumption of a rational source could not be based apriori on prima facie intelligibility is that one could learn empirically that the "voice" was meaningless. This reason is powerless against my conception of the presumption, for I agree that the presumption is empirically defeasible. Apriority has to do with the source of epistemic right; defeasibility is a further matter. For recent criticisms of Hume's view, see A. J. Coady, "Testimony and Observation", *American Philosophical Quarterly* 10 (1973): 149-55; Frederick F. Schmitt, "Justification, Sociality, and Autonomy", *Synthese* 73 (1987): 43-85. I think that empiricism cannot possibly explain all our justified acceptance of what we read or hear. The idea that we should remain neutral or skeptical of information unless we have empirical grounds for thinking it trustworthy is, I think, a wild revisionary proposal. I also think that empiricism cannot account for norms for children's relying on others in the acquisition of language or knowledge.

there are others.) The Acceptance Principle and its justification are formulated so as to be neutral on whether what is "presented as true" comes from another person. Its application does not depend on an assumption that the source is outside oneself (although further articulation will, I think, give this source a place in the account). Many of the differences between content passing between minds and content processed by a single mind derive from differences in modes of acquisition and in necessary background conditions, that do not enter into the justificational force underwriting an entitlement.

An account of an entitlement that includes, as special case, relying on the word of others must, however, acknowledge the following issue: The straightline route from the prima facie intelligibility of a presentation–as–true to prima facie rational characteristics of the source to prima facie acceptability (truth) of the presentation, is threatened by the fact that certain aspects of *rationality* (rational lying) may go *counter* to true presentations. So why should rationality, especially in another person, be a sign of truth? One can have empirical reasons to think someone is no lying. One could have nonrational tendencies to believe, which with luck might get one by. But can one have apriori prima facie rational entitlement to accept what one is told, without considering whether the interlocutor is lying —lacking special reasons to think he is?

Apart from special information about the context or one's interlocutor, neutrality (as well as doubt) is, I think, a rationally unnatural attitude toward an interlocutor's presentation of something as true. (Compare: lying for the fun of it is a form of craziness.) Explaining why, in depth, would involve wrestling with some of the most difficult issues about the relation between "practical" reason and reason. I will broach one line of explanation.

Reason necessarily has a teleological aspect, which can be understood through reflection on rational practice. Understanding the notion of reason in sufficient depth requires understanding its primary functions. One of reason's primary functions is that of presenting truth, independently of special personal interests. Lying is sometimes rational in the sense that it is in the liar's best interests. But lying occasions a disunity among functions of reason. It conflicts with one's reason's transpersonal function of presenting the truth, independently of special personal interests'.[14]

---

[14] Although I think that my claim about this constitutive reason is apriori, I do not maintain that it is self–evident. It has been coherently questioned, as I will note. But the claim has initial plausibility, and I believe that this plausibility is deepened through reflection, including reflection on challenges to it.

The Humean reply that reason functions *only* to serve individual passions or interests is unconvincing. Reason has a function in providing guidance to truth, in presenting and promoting truth without regard to individual interest. This is why epistemic reasons are not relativized to a person or to a desire. It is why someone whose reasoning is distorted by self deception is in a significant way irrational —even when the self deception serves the individual's interests. It is why one is rationally entitled to rely on deductive reasoning or memory, in the absence of counter reasons, even if it conflicts with one's interests. One can presume that a presentation of something as true by a rational being —whether in oneself or by another— has, prima facie, something rationally to be said for it. Unless there is reason to think that a rational source is rationally disunified —in the sense that individual interest is occasioning conflict with the transpersonal function of reason— one is rationally entitled to abstract from individual interest in receiving something presented as true by such a source.

Another consideration pointing in the same direction is this. A condition on an individual's having propositional attitudes is that the content of those attitudes be systematically associated with veridical perceptions and true beliefs:[15] true contents must be presented and accepted as true within some individual; indeed, the very practice of communication depends on preservation of truth. If a rational interlocutor presents intelligible contents as true, one can rationally presume that the contents are associated with a practice of successfully aiming at and presenting truth. Now an inertial principle appears applicable: since the intelligibility of a presentation as true indicates a source of both rational and true content presentations, one needs special reason to think there has been deviation from rationally based, true truth presentation. Other things equal, one can rationally abstract from issues of sincerity or insincerity.

The apriori entitlement described by the Acceptance Principle is, of course, no guarantee of truth. It is often a much weaker sign of truth, from the point of view of certainty, than empirically justified beliefs about the interlocutor. The lines of reasoning I have proposed justify a prima facie rational presumption, a position of nonneutrality —not some source of certainty.

Even if the Acceptance Principle is not an empirical principle, it may seem that particular entitlements sanctioned by it, "applica-

---

[15]These true beliefs could fail to be the individual's own, but they must occur somewhere in the development of the content —for example, in the evolution of the cognitive apparatus.

tions", must inevitably be empirical. To know what one is being told, one must use perception. One must perceive words as expressing content presented as true. In interlocution, perception does inevitably figure in acquisition of understanding and belief. Perception is necessary to minimal understanding; and minimal understanding is essential to belief and justification. But our question concerns perception's role in justification or entitlement. I will first consider its role in justification in our narrow sense, and then turn to its role in entitlement.

One might reason that since the Acceptance Principle counts it. rational for a person to accept what is presented as true, and since one can know what is presented as true by another person only through perceiving an event in time, a person must rely for justificational force on perception of particular events to apply the principle.

This reasoning rests on a confusion about the status of the Acceptance Principle and its justification. The Acceptance Principle is not a premise in an argument applied by recipients of information. It is a description of a norm that indicates that recipients are sometimes entitled to accept information from others *immediately* without argument. The justification of the principle is not an argument that need be used by interlocutors, but an account of why the practice of acquiring information from others is rationally justified.[16] It is well known that we do not store the physical properties of sentences we hear or read.[17] The content of the linguistic forms is what is important. We seem normally to understand content in a way whose unconscious details (inferential or otherwise) are not accessible via

---

[16]Here is a more sophisticated objection along the same line. Suppose that a belief acquired from others may count as knowledge, though one often lacks sufficient grounds, on one's own, to underwrite the belief as knowledge. Suppose that one hIows one lacks autonomous grounds for such a belief. Then one's knowledge that the belief was acquired from others would have to be used to enable one's belief to count as knowledge in view of the known fact that unless the belief had been acquired from others, one's lack of autonomous justification would be insufficient for knowledge. (It is assumed that knowledge that a belief was acquired from others must be empirical. Let us grant the assumption for now.)

This reasoning again rests on a level confusion. If one has acquired one's belief from others in a normal way, and if the others know the proposition, one acquires knowledge. No further reasoning about the practice is needed for the knowledge. No reasoning that does not show that the entitlement has lapsed can undermine the entitlement (though it might mistakenly undermine one's belief that one was entitled).

[17]Kenneth I. Forster, "Lexical Processing", in *An Invitation to Cognitive Psychology*, vol. I, ed. Osherson and Lasnik (Cambridge: MIT Press, 1990).

ordinary reflection. To be entitled to believe what one is told, one need not understand or be able to justify any transition from perceptual beliefs about words to understanding of and belief in the words' content. One can, of course, come to understand certain inferences from words to contents. Such empirical metaskills do enrich communication. But they are not indispensable to it. To be justified in understanding, we have to reason empirically about what we perceive only when communication runs into trouble, or when special, contextual, nonliteral expressive devices are used (see note 21). Other things equal, we are entitled to presume that what seems intelligible is understood. Justification in the narrow sense is not basic to the epistemology of interlocution.

But the question of entitlement is more subtle. In ordinary perception of physical objects and properties we have sense experiences that are not ordinarily the objects of reference or the basis of a justifying inference to perceptual beliefs to which we are entitled. Yet having such experiences, or having perceptual beliefs, contributes to the justificational force of our empirical beliefs:[18] A perceptual belief's being perceptual is all element in its justificational power. The belief's being causally or constitutively associated with sense perception is part of the force of our entitlement to the belief.

In interlocution, we are also causally dependent on perception. Our entitlements are thus dependent on perception. But in my view, perception contributes nothing to the epistemic force of the fundamental "default" entitlement.

Perceptions or perceptual beliefs about physical objects are constitutively dependent on bearing natural lawlike causal relations to objects of perception —to their subject matter, physical objects.

---

[18]Davidson and Sellars deny that having sensations plays a role in justifying perceptual beliefs. I am not convinced by their reasons as applied to entitlements to perceptual belief. See Donald Davidson, "A Coherence Theory of Truth and Knowledge", in *Truth and Interpretation*, ed. Lepore (Oxford: Basil Blackwell, 1986), 311; and Wilfrid Sellars "Empiricism and the Philosophy of Mind", in *Science, Perception, and Reality* (London: Routledge and Kegan Paul, 1963), 164ff. For an alternative to their views, see Steven L. Reynolds, "Knowing How to Believe with Justification", *Philosophical Studies* 64 (1991): 273-92. My view here does not, however, rest on giving sensations (particularly seen as nonintentional) a role in perceptual entitlement. One need not think of sensations as entities, though I do. It is enough that the perceptual character of perceptual belief contribute to the force of the entitlement. Moreover, I am not convinced that there is an epistemic *transition* from perceptual experience to perceptual belief in the ordinary case. One can, of course, learn to suspend such belief. But perceptual experience seems a constituent element in perceptual belief; and perceptual belief seems to be a default position.

The contents of the beliefs and perceptions are what they are partly because of these relations to specific physical objects or properties. Our entitlement to rely on perception and perceptual beliefs is partly grounded in this causally patterned, content–giving relation which is partly constitutive of perception.

When we receive communication, the situation is different. The objects of cognitive interest —the contents and their subject matters— are not the objects of perception. We do not perceive the contents of attitudes that are conveyed to us; we understand them. We perceive and have perceptual beliefs about word occurrences. We may perceive them as having a certain content and subject matter, but the content is understood, not perceived. The subject matter, word occurrences, of our perceptual experiences and beliefs bears a nonconstitutive (quasi–conventional) relation to the content and subject matter of the beliefs to which we are entitled as a result of communication. So the accounts of our noninferential entitlements to perception and to interlocution must be different.

One might note that the relation between perceived words and their contents or subject matters must involve some sort of explanatory relation. So one might be tempted to think that although one does not typically infer the content from the words explicitly and consciously, the entitlement must somehow be based on this explanatory relation. But it would be a mistake to embrace this temptation without reflecting carefully on the special character of the relation as it occurs in interlocution. The relation between words and their subject matter and content is not an *ordinary*, natural, lawlike causal explanatory relation. Crudely speaking, it involves a mind.

There are, of course, complex causal–explanatory relations that may be used to infer the content or subject matter of an interlocutor's speech from perceived word occurrences. One could give an account of entitlement centered on possible inferential interpretations, or on reason–giving explanatory connections between words and content. The interpretation might not be accessible to the recipient, but it could represent a reasonable route from the received message to a putative truth. Such an account —broadly familiar in current discussion—would make the entitlement empirical, because it would appeal in the account of justificational force to an inductive connection to perceived word occurrences.

I do *not* doubt that such accounts are true. I doubt that they are fundamental. I think that what is fundamental is not a metalinguistic connection between word occurrences, taken as objects of perception, and their contents or subject matters. What is fundamental is an apriori prima facie entitlement to rely upon putative under-

standing, and an apriori prima facie connection between putatively understood contents and rational sources of truths. Understanding is epistemically basic. Traditionally, a justification or entitlement was apriori if it could be derived from conceptual understanding — however experientially dependent the understanding might be. The issue over apriority begins with conceptual understanding and asks whether perceptual experience is needed to supplement the understanding for one to be justified or entitled to one's belief.

The epistemic status of perception in normal communication is like the status it was traditionally thought to have when a diagram is presented that triggers realization of the meaning and truth of a claim of pure geometry or logic. Perception of physical properties triggers realization of something abstract, an intentional content, expressed by the sentence, and (often) already mastered by the recipient. Its role is to call up and facilitate mobilization of conceptual resources that are already in place. It is probably *necessary* that one perceive symbolic expressions to accept logical axioms —just as it is necessary to perceive words in interlocution. But perception of expressions is not part of the justificational force for accepting the contents. In both cases, no reference to a possible metainference from expressions to contents is needed in an account of justificational force. The primary entitlement in interlocution derives from prima facie understanding of the messages, and from a presumption about the rational nature of their source —not from the role of perception, however necessary, in the process.[19]

---

[19]The analogy goes with certain disanalogies. Understanding a simple logical truth yields a justification; understanding a communicated message yields an entitlement. This is because in the logic case justificational force derives from the content itself, whereas in interlocution justificational force derives from one's right to putative understanding and from the presumed status of the source of the message, not (typically) from the content itself. A corollary is that knowledge of a simple logical truth does not depend on anything further than understanding and believing it, whereas knowledge based on interlocution depends on there being knowledge in the chain of sources beyond the recipient. In neither case is correct perception of words or correct understanding of what they express necessary to the justification (or entitlement). In neither case is correct perception of words necessary even for knowledge. But in the interlocution case (because knowledge depends on inheriting knowledge from a source), correct understanding of what the interlocutor conveys by the words is necessary for knowledge based on interlocution. (Correct understanding of words or interlocutor is not necessary for knowing whatever logical truth one happens to associate with them, if one understands the logical truth sufficiently.) The important analogy between the logic and interlocution cases is that perception of words makes understanding possible, but justificational force can be derived from the individual's understanding with-

In interlocution, perception of utterances makes possible the passage to propositional content from one mind to another rather as purely preservative memory makes possible the preservation of propositional content from one time to another. Memory and perception of utterances function similarly, in reasoning and communication respectively. Their correct functioning is necessary for the enterprises they serve. Their failure could undermine those enterprises. They preserve the content of events (past thoughts in proof, word utterances in interlocution) —events that *can* become objects known empirically. But the basic epistemic role of memory and perception in these enterprises is not to present objects of knowledge. They function to preserve and enable —not to justify.

In interlocution, the individual's basic default entitlement normally derives from the presumptive intelligibility of a message understood, not from anything specific in the words perceived. Unless reasonable doubt arises about the reliability or interpretation of the source, the specific perceptions of utterances need not be relied upon in contributing force to the receiver's entitlement to his understanding of or belief in what is communicated.

Perception might be thought part of the justificational force of our entitlement in another way. The justification of the Acceptance Principle says that one is entitled to accept intelligible contents "presented as true". We must perceive a speech act as involving a presentation–as–true to be justified under the principle. Why does it not follow that our entitlement to accept what we are told in particular cases relies for its force on perceptual beliefs?

The issues here are again very complex. But the short to the question is that one's intellectually grounded entitlement to one's understanding of content includes an entitlement to understand presentations–as–true. Understanding content presupposes and is interdependent with understanding the force of presentations of content. So entitlement to the former must presuppose entitlement to the latter. In many normal cases the epistemology of our entitlement to understanding assertive force has a default status that is parallel to that of our entitlement to understanding content. Perception is no more basic to understanding assertive force than it is to understanding conceptual content. The default position is that presumed

---

out supplementary appeal to perception. I am abstracting, in this discussion of applications, from cases where understanding a particular content itself involves perceiving —for example, perceiving the referents of demonstratives. Such understanding is not purely conceptual; and as a consequence, the relevant entitlement to the particular belief is partly perceptual.

understanding of both content and force is epistemically fundamental. Empirical justification for an interpretation of content or force is demanded only when elements in the context demand reconsideration or supplementation of the default understanding. I find the parallel compelling. But I will sketch in two steps a picture of how default understanding of a presentation–as–true can sometimes be derived from no more than default understanding of propositional content. This picture is not needed, but it may enrich the account.

First, entitlement to one's understanding of a message's content carries with it, indeed rests on, an entitlement to understanding intentional *events* as having specific content. Understanding speech acts or thoughts as they occur is the root of understanding content types. The necessary role of perception in enabling one to follow another's speaking or thinking is not fundamentally different from its role in enabling one to grasp the abstract content of another's sentence. All that I have argued on the latter score applies to the former. Perception's basic role is to make understanding possible and to trigger it on particular occasions. But the justificational force of one's basic default entitlement to understand something as an event with a specific content is not perceptual. It is intellectual in that it resides in one's putative understanding of conceptual content in application or use, in one's ability to think–with.

Second, understanding conceptual content —both abstractly and in contentful events or uses— involves understanding the content's mood. But for contents in the indicative (declarative) mood —as distinguished from interrogative or imperative mood— presentation–as–true is the defeasible default use. The connection between declarative mood and presentations–as–true is conceptual. The justificational force of the entitlement to rely on the connection is correspondingly conceptual, not perceptual.[20]

---

[20]Donald Davidson has argued that there is no conventional connection between indicative sentences and assertive use. See "Moods and Performances" and "Communication and Convention" in *Inquiries into Truth and Interpretation.* His reason is that one can always use indicative utterances for nonassertive purposes. I find the argument unconvincing. A conventional connection between indicative mood and assertive use could be flouted. I believe that the connection between assertive use and indicative (declarative) mood is deeper and firmer than merely conventional. But it is a contextually defeasible connection.

I use the term 'presentation as true' to cover more than assertions and judgments. Obvious presuppositions, or conventional implicatures, are examples. When someone says to kill the shortest spy, he or she presents it as true that there is a shortest spy. In such cases, as well as the indicative cases, the entitlement to accept what is presented as true can be independent for its justificational force of perceptual connection to context (see note 21).

In the absence of overriding reasons, the default presumption stands. Nonassertive uses (jokes, irony, fiction) that drain declaratives of assertive implications must employ context to make themselves understood. The recipient must infer that the sentence is used nonassertively from empirical information about the context. Although affirmative use of declarative contents must, on occasion, also be inferred from special contextual information, taking a declarative sentence utterance as a presentation–as–true normally requires no such reasoning or empirical interpretation.[21]

Thus in many instances, one's entitlement to take something as a presentation–as–true in interlocution derives from understanding an event's content, and need not rely for its justificational force on perception of word occurrences. What one is entitled to on intellectual grounds is merely, prima facie, that a given content is presented as true. One gets nothing about the time, form, or circumstances of the assertion. All such information is epistemically grounded in perception of aspects of the context. But the fundamental entitlement to accept something as a presentation–as–true derives from understanding. It can even be derived sometimes from understanding of content (its tokening and the relation of its mood to presentations–as–true). The justificational force of the derivation does not depend on any supplementation from perception. Perception plays its role in making understanding possible and in justifying supplemental information about the form, existence, and context of the assertion.

In appreciating these points, one must distinguish between knowing about the assertion as part of a pattern for explaining the psychology and behavior of the asserter, and using the interlocutor as a source of information. In the former enterprise, perception of an assertion as an action by a particular individual is commonly taken as an element in the justification of an explanation, or an object of interpretation. But in interlocution, perception need not play this role

---

[21]This point allies with Grice's distinction between conventional and conversational implicature. See Paul Grice, *Studies in the Way of Words*, 28-31. Grice requires that to be "conversational", an implicature must be capable of being "worked out' from considerations of the conversational context. Conventional implicatures may be inferred "intuitively" from the meaning of the words. I think that understanding based on conversational implicatures *must* be justified, usually empirically, whereas understanding based on conventional implicatures can rest on apriori entitlement. Analogously, I think that a construal of a sentence or content as ironic *must* be justified, usually empirically, whereas a construal of a sentence as asserted can rest on an apriori entitlement. A parallel story needs to be told about ambiguities. Our ability to understand many ambiguous sentences as they are meant, even apart from context, indicates that certain readings are default readings.

unless some reasonable doubt arises about the informant's message or the recipient's understanding (see note 25).

One can know through memory the events that help recall the previous step in a proof, thereby making those events objects of knowledge. One can know on the basis of perception that a particular person made an assertion at a given time. One can surely construct an empirical metajustification (or entitlement) for one's belief based on interlocution: "She asserted that p (known empirically); it is prima facie reasonable to rely on others' assertions; so I should rely on her assertion". Such metajustifications supplement one's epistemic position in interlocution. But they are not, I think, fundamental. Just as remembering events does not enhance the primary object level justification in deductive argument, so relying on perception does not contribute to the justificational force of one's fundamental entitlement to one's understanding of content, or to one's acceptance of what is presented as true.

Let us return from our entitlement to understanding to our entitlement to believe what we hear, given that we understand it. When we receive a message, we often know a lot about the context of the reception, the biography of the source, the antecedent empirical plausibility of the information. This knowledge is inevitably perceptually grounded. Does this fact make our entitlement to believe what we receive from others inevitably perceptual? I do not think so. Our initial entitlement does not depend on this knowledge for its justificational force.

In areas like politics, where cooperation is not the rule and truth is of little consequence, or philosophy, where questioning is as much at issue as belief, we engage in complex reasoning about whether to accept what we hear or read. Reasonable doubt becomes a norm. But these situations are not paradigmatic. They are parasitic on more ordinary situations where acceptance is a norm.

The default position is justified acceptance. Often we need empirical reasons to defeat reasonable doubts that threaten our right to acceptance. But sometimes empirical reasons simply reinforce an(l over(determine the (default entitlement. Our being justified does not then rest indispensably on empirical background information.[22]

---

[22]The scope for intellection–based justification in interlocution is wider than these remarks may suggest. I think that in certain cases special confidence in an interlocutor can be justified on grounds that are inductive but, with subtle qualifications, intellectual. I discuss these matters further in "Computer Proof and Apriori Knowledge".

I turn now from our entitlement to applications of the Acceptance Principle to the role of interlocution in the acquisition of knowledge. In the absence of countervailing considerations, application of the Acceptance Principle often seems to provide sufficient entitlement for knowledge. Most of our knowledge relies essentially on acceptance of beliefs from others —either through talk or through reading. Not only most of our scientific beliefs, but most of our beliefs about history, ourselves, and much of the macro world, would have insufficient justification to count as knowledge if we were somehow to abstract from all elements of their justification, or entitlement. that depended on communication.

Our entitlement to ordinary perceptual belief is usually sufficient knowledge. It is usually sufficient even though we may be unable specifically to rule out various possible defeating conditions. If there is no reason to think that the defeating conditions threaten, one has knowledge despite ignoring them. Something similar holds for acquisition of belief from others. Other things equal, ordinary interlocution suffices for knowledge.[23]

In knowing something through interlocution, the recipient has his own entitlement to accept the word of the interlocutor, together with any supplementary justification the recipient might have that bears on the plausibility of the information. Let this include all the reasons available to the recipient, together with all the entitlements deriving from his own cognitive resources. Call this body (i) the recipient's *own proprietary justification.*

If the recipient depends on interlocution for knowledge, the recipient's knowledge depends on the source's having knowledge as well. For if the source does not believe the proposition, or if the proposition is not true, or if the source is not justified, the recipient cannot know the proposition. The recipient's own proprietary entitlement to rely on interlocution is insufficient by itself to underwrite the knowledge.[24] In particular, the recipient depends on sources'

---

[23]The fact that most of our knowledge is dependent on others and has distinctive epistemic status is increasingly widely recognized. See C.A.J. Coady, "Testimony and Observation"; John Hardwig, "Epistemic Dependence", *Journal of Philosophy* 82 (1985): 335-49; Michael Welbourne, *The Community of Knowledge* (Aberdeen: Aberdeen University Press, 1986). For a wildly implausible, individualistic view of the epistemic status of testimony, see John Locke, *An Essay Concerning Human Understanding* 1.3.24.

[24]Because the interlocutor must have knowledge and because of Gettier cases, the interlocutor must have more than true, justified belief if the recipient is to have knowledge. The recipient s dependence for having knowledge on the interlocutor's having knowledge is itself an instance of the Gettier point. The recipient

proprietary justifications and entitlements (through a possible chain of sources). The recipient depends on at least some part of this body of justification and entitlement in the sense that without it, his belief would not be knowledge. The recipient's own justification is incomplete and implicitly refers back, anaphorically, to fuller justification or entitlement. Call the combination of the recipient's own proprietary justification with the proprietary justifications (including entitlements) in his sources on which the recipient's knowledge depends (ii) *the extended body of justification* that underwrites the recipient's knowledge.

At the outset, I explained apriori knowledge in terms of apriori justification or entitlement. The question arises whether apriori knowledge based on interlocution is underwritten by the individual's proprietary justification or by a justification that must include some nonproprietary part of the extended body of justification.

The extended body of justification —the one that reaches beyond the individual— is the relevant one. If I am apriori entitled to accept an interlocutor's word, but the interlocutor provides me with empirically justified information, it would be wrong to characterize my knowledge of the information as apriori. Similarly, if my source knows a proposition apriori, but I must rely on empirical knowledge to justify my acceptance of the source's word, it would wrong to say that *I* know the proposition apriori —even though I have knowledge that is apriori known by someone. It seems most natural to think that a strand of justification that runs through the extended body into the individual's proprietary body of justification must be apriori for the recipient's knowledge to be apriori. People who depend on interlocution for knowledge of mathematical theorems but do not know the proofs can have apriori knowledge in this sense. The source mathematician knows the theorem apriori and the recipient is entitled apriori to accept the word of the source, in the absence of reasons to doubt. Most of us knew the Pythagorean theorem at some stage in this manner. When apriori knowledge is preserved through reports which the recipient is apriori justified in accepting, the receiver's knowledge is apriori.

The Acceptance Principle is clearly similar to what is widely called a "Principle of Charity" for translating or interpreting others. The

---

could have true justified belief, but lack knowledge because the interlocutor lacked knowledge.

In requiring that the source have knowledge if the recipient is to have knowledge based on interlocution, I oversimplify. Some chains with more than two links seem to violate this condition. But there must be knowledge in the chain if the recipient is to have knowledge based on interlocution.

most obvious difference is that the former applies to situations in which one is not taking another as an object of interpretation, but rather as a source of information presumed to be understood without interpretation. This situation is basic for communication.[25] Radical interpretation is not, I think, the paradigmatic situation for theorizing about linguistic interchange.

We rely on being so formed that we take in information from others without interpretation. Unlike the Principle of Charity, the Acceptance Principle presumes not only that we are like others in being rational. It presumes that we preserve content, other things equal. This presumption works because we share with others around us our cognitive tendencies and means of expressing them, and a common environment. But we do not have to *justify* a claim that these conditions for success are in place to be entitled to rely upon our understanding. (Analogously, we do not have to justify a claim that the environment is normal and we are adapted to it in order to be entitled to rely on perception.) It is enough if we learn how to understand. Once we are in a position to understand, we are entitled to the following presumption apriori, other things equal: We understand what we seem to understand. Or rather, other things equal, we need not use a distinction between understanding and seeming to understand. We need not take what we hear as an *object* of interpretation, unless grounds for doubt arise. Only then do we shift from content preservation to interpretation.

The Acceptance Principle entails a presumption that others' beliefs are justified, that others are sources of rationality or reason. The view that others' beliefs can be presumed to be true is familiar from the Principle of Charity. The presumption that others are reliable indices of truth rests on a presumption that they are rational sources. Their reliability is not some brute correlation between belief and

---

[25]The principle of charity is illuminatingly used by W. V. Quine, in *Word and Object* (Cambridge: MIT Press, 1960), chap. 2; and Donald Davidson, in "Radical Interpretation" (1973), in *Inquiries into Truth and Interpretation.* In holding that interpretation is the basic situation for understanding linguistic interchange, Davidson writes, "The problem of interpretation is domestic as well as foreign: it surfaces for speakers of the same language in the form of the question, how can it be determined that the language is the same?" (Similar passages can be found in Quine.) Davidson presupposes that determining whether we are communicating successfully when we appear to be is a question in place from the beginning. This seems to me mistaken. Such a question arises only when there is some reason to doubt that we are sharing information and preserving content. The default position is that understanding can be presumed until something goes wrong. Incidentally, I do not assume that anything as global as a communal language need be thought of as fundamental. That is a further issue.

world. We are entitled to treat others as reliable partly *because* we are entitled to presume that they are rationally justified or rationally entitled to their beliefs. We are entitled, most fundamentally, to think of others as sources of rationality or reason not because we take them as objects of interpretation and explanation, but because prima facie intelligibility is an apriori prima facie sign of rationality.

This focus on others is articulated from a first person point of view. Each of *us* is justified in presuming that others are justified. But we are possible interlocutors too. The idea that others are prima facie justified in their beliefs makes general sense only if we presume generally: people, including each of us, are reliable rational sources of true justified beliefs. Obviously the conclusion requires qualification and elaboration. But the route to it is, I think, of interest. I arrived at it by arguing that we have intellection grounded prima facie entitlements to applications of the Acceptance Principle, though they are empirically defeasible. I think that this approach to epistemology may help with some of the traditional problems of philosophy.

PHILOSOPHICAL ISSUES, 6
Content, 1995

Testimony and *a priori* Knowledge

J. Biro

It is natural to think that if I learn something from someone else, my knowledge must be *a posteriori*. After all, my being told is a bit of experience in the absence of which I would not have the knowledge I do have.[1] It is this orthodoxy that Tyler Burge's complex and provocative paper aims to challenge. Burge argues that if my informant's knowledge is *a priori*, and I come to acquire that knowledge solely through his testimony, then so is mine. This is because I have an *a priori* (though defeasible) entitlement to accept such testimony. If both my informant's knowledge and my entitlement to accept what he says as true are *a priori*, the fact that the transmission of the information involves empirical matters does not compromise the *a priori* character of the knowledge I acquire.

I want to raise two questions about Burge's argument. The first is whether the *a priori* principle Burge identifies, one that governs the status of testimony in general, is sufficient to guarantee in a

---

[1] Assume here that what I have *is* knowledge, that is, that what I am told is known by my informant. We can, for present purposes, leave open exactly what that requires of him.

particular case that something an informant knows *a priori* will be similarly *a priori* for those who hear it from him. I shall suggest some reasons for thinking that it is not. The second question concerns Burge's arguments for the general principle and its *a priori* status. While I think that something like this principle may be true, I shall express some doubts about Burge's way of justifying it.

Burge calls the principle in question the "Acceptance Principle":

> AP: A person is entitled to accept as true something that is presented as true and that is intelligible to him, unless there are stronger reasons not to do so. (281)

I. I begin with a summary of how I understand those parts of Burge's discussion to which I want to address these questions. It may be that my understanding of these is flawed: if so, it is better that this should be evident at the start.

What things can be said to be *a priori* or, its opposite, *a posteriori* (empirical)? The traditional answer has been that it is propositions, truths, and our knowledge of these.[2] Burge expands the class of things to which he is willing to attribute these properties to include justifications and entitlements. This enables him to argue that knowledge is *a priori* when it is supported by or rests upon *a priori* justification or entitlement and, especially interestingly and importantly, that this can be the case even when empirical evidence enters into, or empirical conditions need to be satisfied for, its actual acquisition.

The key to the initially perhaps surprising claim that the involvement of empirical factors need not render a justification or an entitlement, and thus a piece of knowledge supported by these, *a posteriori* lies in recognizing that such factors can play more than one kind of role. In only one of these roles, that of contributing to what Burge calls "the justificational force" in an essential way, would the presence of empirical elements render the resulting knowledge itself empirical. But, Burge argues convincingly, the empirical factors that

---

[2]Kant —who is much in the background of Burge's paper— talked of *judgments* as being *a priori* or *a posteriori*; these are, I think, best understood as claims or potential claims to knowledge in the way that some statements of belief are. Not all, of course: we often say that we (merely) believe when not ready to claim knowledge. But Kant has in mind that other —perhaps more fundamental— sense of 'judgment' in which to judge is to lay claim to knowing and, correspondingly, saying that one has judged is giving more than a psychological report about oneself. We also talk sometimes of a method as being *a priori* or empirical; Burge's talk of a justification's being so is clearly related to this.

are involved in the way in which we arrive at knowledge, or which have to be present for such knowledge to be possible, often do not play such a role. For example, our necessarily having to rely on memory in carrying out a demonstration does not —*contra* Chisholm and others— compromise the *a priori* character of the demonstration or of its conclusion. Since a chain of reasoning takes time, obviously the reasoner must rely on his memory to somehow "keep before him" the various propositions among which the logical relations enabling the chain to constitute a demonstration (are supposed to) hold.[3] But this does not require any proposition *about* memory to play a substantive role in the demonstration, so it does not enter into its justificational force and, for that reason, does not compromise its *a priori* character.

The distinction between what we might call a substantive or internal role and a merely sustaining or external one applies not only to memory but, more generally, to all propositions or principles that involve empirical matters. What determines whether a piece of knowledge is *a priori* is whether in gaining or justifying it we (must) rely on such a proposition or principle playing the first, substantive, kind of role.[4] This will be the case generally (perhaps only) when the proposition relied upon concerns a particular piece of sense experience or "the specifics of some range of sense experiences or perceptual beliefs". (281)

Burge glosses the non–substantive or external role of memory as "content preservation". His idea is that there are principles that

---

[3]Kant's "synthesis of reproduction" —the second component of the synthesis of experience as described in the **A** deduction (**A**100-2) and Hume's comments on the role of memory in hearing a tune (**T**36-7) are close cousins of this insight. (For detailed discussion of these historical precedents, see J. Biro, "What —if anything— is transcendental synthesis", in *Doing Philosophy Historically*, Peter Hare, ed. (Prometheus Books, 1988) and J. Biro, "Hume's new science", in *The Cambridge Hume Companion*, David Fate Norton, ed. (CUP, 1993.))

[4]On this picture, there are propositions that may be known *a priori* by one person and *a posteriori* by another, or even in both ways by the same person. (273) In this way, the characterization of such a proposition's epistemic status (in this dimension) is relative both to method of actual justification and to epistemic agents. Other propositions can be known only *a posteriori*, which means that one *must* appeal to empirical facts in justifying them. Yet others —presumably logical and mathematical ones among them— may be thought to be knowable only *a priori*. Traditionally, philosophers have focused on the last group in their discussions of the *a priori*, giving the misleading impression that only these can be known non–empirically. One of Burge's main aims is to develop a more generous conception of the *a priori*, one on which more kinds of knowledge, including some acquired through perception, can count as *a priori*, as long as its justification is.

back up our ability to acquire knowledge from others whose status is analogous to those allowing memory to play the merely content–preserving role it does in a demonstration. As with memory, these principles need enter neither into the process through which the knowledge is acquired nor into any justification of the claim that what has been acquired *is* knowledge; nonetheless, they function to underwrite that claim. A principle that plays this role is, for Burge, *a priori* in the most fundamental sense, in that it expresses a justification or entitlement which are the things to which '*a priori*' "primarily" applies. (272)

II. Let me grant, at least for the sake of the argument, that AP is true and, furthermore, that it is knowable *a priori*. Would its truth, along with the fact that in a particular instance my informant's knowledge is *a priori*, suffice to render the knowledge I acquire through his testimony *a priori*? Even to guarantee that I have knowledge, *simpliciter*, would depend on whether the condition in its defeasibility clause is met. That is, it would have to be the case that there *are* no stronger reasons for me not to accept the testimony; only then would I actually have the entitlement described. This much is obvious. But let us suppose that the condition *is* satisfied. And suppose, for the sake of the argument something that is actually arguable, namely, that there being no defeaters, as distinct from my correctly judging that there are none, is sufficient for my having knowledge. We still do not have an answer to whether the knowledge I have have is *a priori* or not. It would seem that the answer to *that* question must require that I correctly *judge* that there are no defeating reasons. Only then will the principle kick in to do the work required here, namely, to serve as (part of) my justification for accepting what I am told as true. If the traditional view about what makes a piece of knowledge *a priori*, namely, that its possessor is able to justify it without relying on empirical propositions, is at all plausible, having a purely external entitlement cannot by itself suffice (even if it did with respect to empirical knowledge). But now we can see the problem: a judgment that there are no defeaters in a particular situation looks to be thoroughly *a posteriori*. If it is, then empirical considerations enter essentially into the justification of my knowledge, rendering it, on Burge's own account, *a posteriori*.[5]

---

[5]There is another reason why one might think that knowledge "underwritten" by (AP) must be *a posteriori*: any instantiation of the principle, even in its conditional form, presupposes the existence of individuals. But I leave this complication aside here.

Having the sort of entitlement expressed by AP is enough, according to Burge, to make the agent's epistemic state one of knowledge. He may be right in this. As externalists have emphasized, the ability to justify one's belief is neither necessary nor sufficient for knowledge. Typically, such externalism focuses on the satisfaction of certain causal and reliability conditions, but AP can be seen as expressing another kind of external condition, one special to knowledge acquired from others.[6]

However, there is also a flavour in some of Burge's discussions of the entitlement in question of something more internal. While he insists that "[AP] is about entitlement, not psychological origin", he sometimes couples having an entitlement and having a reason in a way that suggests that what he is interested in is not the fact of the entitlement but the agent's reliance on it. He says: "If one has a reason or entitlement to accept something because it is, prima facie, rationally supported one has a reason or entitlement to accept it as true". (283) One natural interpretation of 'having a reason to accept' is as meaning that that reason is a cause of one's acceptance of what one is told.[7] One's belief that AP is satisfied could serve as such a causally efficacious reason for one's acceptance of a piece of testimony. Merely being entitled, in the non–psychological sense, could not, since nothing non–psychological could.

The interpretation of 'having an entitlement' suggested by Burge's remarks about their non–psychological character is that of *being* entitled, whether one knows that one is or not. But "[having] a reason or entitlement to accept something *because* it is, prima facie, rationally supported..." (my emphasis) carries a different suggestion, namely, that the *prima facie* rationality of some claim leads one to take oneself to be entitled to accept it and that so taking oneself is the (causally efficacious) reason for one's accepting the claim. Such a taking oneself to be entitled *is*, of course, a psychological matter.

---

[6]There is an analogous principle we could formulate for perception that would go roughly like this: (*APp*) *A person is entitled to accept as real something he perceives, unless there are reasons not to do so.* I am entitled to assume that the things I perceive exist, as long as I have no reason to think that my condition or the circumstances are abnormal. There is a presumption (defeasible, of course) that normal perception is to be taken at face value.

[7]It is not the only one, of course. One may not know the reason that causes one to do what one does. It can also be that there *is* a reason for doing what one does without that reason's being one's reason, in the sense that the thing *would* and *should* be a reason for an ideally rational agent. Only the last sense carries a normative implication; only the first two are relevant in an explanation of action.

The fact that there are these two ways of seeing what 'entitlement' talk comes to presents what I see as a dilemma for Burge. If we take such talk in the non–psychological, non–explanatory, way he says we should, we do not seem to have anything that can play a justificatory role for the epistemic agent whose knowledge we are interested in characterizing. If, on the other hand, in spite of Burge's official line, we opt for the explanatory, psychological, reading, it seems fairly clear that to come to have this sort of reason for accepting something, namely, the thought that one is entitled (on this occasion) to accept it, one would have to rely on *a posteriori* evidence.

It is for these reasons that I have trouble parsing locutions such as "We are *a priori* prima facie entitled..." (284), on the one hand, or ones like "...the recipient has his own entitlement" (291), on the other, which makes me uncertain about the force of AP *vis–a–vis* the status of beliefs acquired from others. As we have seen, the mere fact that I am, defeasibly, entitled to accept what they say does not seem to me enough to render what I accept *a priori*. As Burge would agree, that depends on how I would/could justify accepting what I do. But when he says "[beliefs acquired from others] are sometimes apriori justified in the sense that they need not rely for justificational force on the specifics of some range of sense experiences or perceptual beliefs". (282), he overlooks the fact that the experience required to judge that one is entitled *does* enter into one's justification for accepting what one does.[8]

It is, I think, plausible to think that there is an analytic connection between the *prima facie* intelligibility of what one hears and one's (defeasible) entitlement to accept it as true, and since (most, if not all) analytic connections are knowable *a priori*, there is a clear sense in which AP would have *a priori* status for anyone who came to believe it on the basis of conceptual analysis. But while in this way AP itself can be said to be *a priori*, it does not follow that the knowledge it can underwrite *in the default situation* is *a priori*. In AP, *we* have an *a priori* justifier for saying that a person relying on it (in the right circumstances) has knowledge. That does not mean that *his* knowledge is *a priori*.[9]

---

[8]This oversight is all the more puzzling in the light of Burge's insistence that whether a belief is *a priori* depends on whether anything empirical enters essentially into the justification the believer has, or would offer, for it. (274) It seems that Burge himself thinks that i) it is beliefs that are *a priori* or empirical and ii) that whatever makes a belief *a priori* or empirical is itself a belief or a set of beliefs.

[9]Burge himself says things that suggest that he has analyticity at least half in mind. He speaks of an *a priori* connection between rationality and truth (283)

As I have been emphasizing, it is one thing to be entitled, another to think oneself entitled. While the nature of one's epistemic state may well depend on the former, what one does to come to be in a state of a certain sort, and how one would defend one's claim to be in a state of that sort, essentially involve the latter. But since whether one's knowledge is *a priori* depends on the answers to the latter questions, it depends on whether empirical considerations are essential for answering these. And it is hard to see how the questions whether one should accept a particular piece of testimony and whether one is justified in thinking oneself entitled to accept it can be anything other than empirical ones, in spite of there being a defeasible general entitlement to accept testimony at face value.

We can agree with Burge that relying on AP is the default position for a rational person and still ask what leads such a person to apply that general reliance in particular cases. How do I come to believe, and what justifies my belief, that I *am* in the default situation? By definition, I am not always in that situation; but then I must have reasons for thinking that I am, when that is what I think. These reasons inevitably involve empirical evidence and empirical judgments. Thus the fact that AP itself is not an empirical principle, as Burge repeatedly insists (e.g., 282), does not mean that the beliefs it underwrites are not empirical. And that is not (just) because hearing others' words requires perception. That consideration Burge dismisses (286) on the grounds that the role of perception in *this* regard is mere "content preservation", and he may be right to do so. But the real reason why second–hand beliefs are empirical is unaffected by such a reply: that reason has to do with having to have empirical reasons for relying on AP in particular cases, something one may have empirical reasons not to do.[10]

That one may have such negative reasons in particular cases is implied by the description of AP as defeasible. But one could have

---

and of "conceptual relations between content and rationality" (284). He also speaks of intelligibility "presupposing" rational abilities and entitlements (285) and of "a conceptual relation between intelligibility and a rational entitlement or justification". (284) The idea of something like AP being presupposed by rational discourse by second–hand knowledge may be thought to be a plausible one, and my present line of criticism is not meant to cast doubt on it. (Note, however, the expression 'rational entitlement', suggestive, again, of an internal, substantive, role.)

[10]These reasons need not, of course, be explicit. That they are typically not is part of what is meant by 'default' here.

reasons for a more wholesale refusal to rely on AP. There is nothing incoherent in the idea that I might find myself among people whose behaviour convinces me that I should be generally more cagey in my dealings with them than AP recommends. They could have shown themselves to be habitual liars, giving me reason not to trust them. (My belief that they lie and thus my belief that I should not rely on AP may, of course, be false. So an argument, if there is one, that being a habitual liar is incompatible with being rational would not affect the present point.) If I had never been among more trustworthy folk, my choice not to rely on AP would not be a conscious, reflective one; I would naturally act in accordance with converse.[11] All of this is a thoroughly empirical matter.

III. At the heart of Burge's justification of AP lie the claims he makes about conceptual connections among three things: the *prima facie* intelligibility of utterances, the *prima facie* rationality of speakers, and the general reliability of *prima facie* rational speakers. I am inclined to agree that there are analytic truths about the relations among these. If there are such truths, they are, of course, discoverable *a priori*. But the ones I see are not sufficient to support AP.

Burge's three main claims are 1) that the *prima facie* intelligibility of an utterance is sufficient for accepting it as intelligible, 2) that the intelligibility of an utterance, along with the rationality of its source, "mark" a *prima facie* connection to truth, and 3) that the source of a *prima facie* intelligible utterance *is* rational. (283) I think 1) is unquestionably true: to be *prima facie* intelligible is to *be* intelligible, as long as we do not confuse being intelligible with having an intelligent source. For 2), Burge offers a version of the argument that for utterances to have content —for there to *be* content at all— many of them must be true. Such a general, though not, of course,

---

[11] In one place Burge explicitly recognizes the difference between being entitled and believing oneself entitled. (285 fn.) There he suggests that the former is (defeasibly) sufficient for the possession of second–hand knowledge. Since AP can be known *a priori*, knowledge it is sufficient to underwrite must itself be *a priori*. But there is a gap in this reasoning. If, as the orthodox analysis has it, having a justification for believing that p is necessary for knowing that p, there must be something *in the believer's mind* that constitutes that justification, something that underwrites his belief in a sense different from the sense in which AP does. As I have argued, being entitled is just not the right sort of thing for such a job; only (rightly) thinking oneself entitled is. But for the latter, one must have empirical reasons, even in the default case.

universal, connection is claimed to be constitutive of intelligibility. While there may be problems this line of thought, I shall let it go unexamined here. The problem I want to highlight concerns 3). Burge apparently assumes that the connection between *prima facie* rationality and rationality is like that between *prima facie* intelligibility and intelligibility. But it is not. The *prima facie* rationality of a source does not guarantee its actual rationality in the way that the *prima facie* intelligibility of a message guarantees its intelligibility. Appearances can be misleading in the former case in a way that they cannot be in the latter. Intelligibility is an intrinsic property of a message, and a message either has it or it does not, regardless of its source. By contrast, coming from a source of a certain sort is not an intrinsic property but a relational one. We do not have to *infer* that a message is intelligible from some other property of it, in the way that we do have to infer that it has a rational source, from, among other things, its intelligibility. In short, intelligibility is conceptually independent of having a particular (kind of) source (think of the monkey typing out all of Shakespeare), whereas the rationality of a message's source is obviously not.

The upshot is that the fact that one is entitled to take one's fellow-speakers' words at face value, in the sense that one is entitled to assign to them the meaning they appear to have, does not entail that one is entitled to take those words to be true. Even if an utterance's being intelligible and actually having a rational source were sufficient warrant for taking it to be true, its being intelligible and its (merely) appearing *prima facie* to come from a rational source is not.

Burge may reply that AP was never intended to guarantee as tight a connection between rationality (never mind *prima facie* rationality!) and truth as this objection assumes: that is the point of its defeasibility clause. But such a reply would be inadequate. One has to have reasons for taking particular utterances as true, given that not all are. As we have seen, merely being entitled to do so in general (in the absence of defeaters) cannot serve as such a reason. While *that* one is generally entitled to believe people may be an *a priori* truth deducible from concepts such as intelligibility, rationality, communication and the like, only one's way of justifying one's belief that one is so entitled in a particular instance is relevant to the status of one's knowledge.[12] And, as I have argued, that belief must be based on, and justified by, empirical evidence; hence any knowledge supported by AP cannot be *a priori*.

---

[12]I am grateful to Kirk Ludwig for many helpful suggestions.

The principles Burge investigates may indeed suffice for the preservation of content and the transmission of knowledge. But even if they do, identity of content is not the same thing as identity of epistemic status. Even if principles constitutive of communication among rational beings guarantee that if, in normal conditions, one of them knows something and tells the other, the latter will thereby come to know it, too, it does not follow that they know that thing in the same way.

PHILOSOPHICAL ISSUES, 6
Content, 1995

Can Peter Be Rational?[1]

Lourdes Valdivia

1 Introduction

The aim of this paper is to discuss a puzzle about belief attribution. My case could be seen as a variation of Kripke's Paderevski–case[2] but differs from his in that the occurrence of two tokens of the same type–name refer to different bearers. Notwithstanding the difference in bearers Peter, the believer in my case, is *prima facie* justified for changing or not his former belief that Pavarotti is the King of the High C's to Pavarotti is not the King of the High C's . I conclude that if we take my case as one of attributting a belief either *de re* or *de dicto*, neither of the two different semantics for proper names, namely (ET) and (IT) here discussed, provide a way out of the puzzle. (ET)

---

[1]Previous versions of this essay were writen while spending my sabatic year as visiting scholar at CUNY. I'm grateful to Jerrold Katz for discussion on his theory as well as to the members of the Department of Philosophy of CUNY. Special thanks are due to Enrique Villanueva for encouraging me to publish this paper and for his thoughtful comments on many many drafts. To James Tomberlin whose clarifying remarks made possible this final version and to Maite Ezcurdia for her discussion on the externalist's arguments.

[2]Saul Kripke, "A Puzzle About Belief", in *Meaning and Use*, A. Margalit (ed.) (Dordrecht: Reidel, 1979); also in *Propositions and Attitudes*, N. Salmon and S. Soames (eds.) (Oxford: UP, 1988) p. 102-148. Quotations are from the latter edition.

troubles with Kripke's cases as well as with mine, while if (IT) gives a solution to my case, then Kripke's Paderevski–case is a problem for (IT).

(ET) is rooted in the Millian tradition which claims that the semantics for proper names is explained by taking into account the bearer of the name. In other words, difference in reference in the terms which aid in the attribution of *de re* belief gives rise to a difference in mental content. (IT) is inspired by the Fregean tradition but contrary to it, claims that the determination of mental content doesn't have to take into account the external circumstances in which the tokening is realized, that is to say, difference in reference in the terms which aid in the attribution of *de dicto* belief does not give rise to a difference in mental content.

Frege's systematic semantics held a definition of *sense* that promised to play two roles: (1) *sense–extensional* role and (2) *sense–intensional* role.

(1) According to *sense–extensional* role, *sense* determines reference, *i.e.* the satisfaction of the properties expressed by the *sense* of a sentence leads to its truth value, if and only if the singular term occuring in the sentence has a reference; the satisfaction of the properties expressed by the *sense* of a proper name, determines the bearer's name; and the satisfaction of the properties of the *sense* of a predicate determines the set of objects which fall under it. For simplicity I will call *extension*(s) to bearers, to sets of objects wich fall under a predicate, and to trurth values. In short, (1) characterizes the standard view in semantics and provides a strong motivation for a theory of meaning to embrace a world–meaning relation, where *intension* determines *extension* whenever and if the expressions and sentences does not occur into opaque and modal contexts.

(2) *Sense–intensional* rol. *Sense* is independent of reference in order to accout for the three following features: (i) sentences are meaningful in spite of the fact that it could be the case that the singular term occurring in the sentence may have no reference; (ii) the cognitive value of '$a = b$' statements differs from that one of '$a = a$' statements, notwithstanding the fact that both are true of the same object; (iii) the substitution of co–referring terms into opaque contexts[3] is not truth–preserving in spite of the fact that the terms are co–referential.

---

[3] I will discuss only belief contexts.

Propositional attitudes like believing that Pavarotti is the King of the High C's generate opaque contexts. Given the assumption that that–clauses refer to *sense*, the idea is that the components of the sentence occurring after a that–clause contribute only their *sense* to determine or individuate the believed proposition.

In particular, were a proper name to occur after a that–clause, it should contribute only its *sense*[4] to that proposition. In other words, mental content referred to by a sentence after a that–clause is "somehow" independent of the world. But how the independency feature should be understood? There are at least two ways of looking at it, (A) the classical Fregean one and (B) the Katzian proposal which partially motivates claim (IT):

(A) *Sense*'s independency lies just in the fact that the *sense*, expressed by any sentence after a that–clause does not determine its customary reference, *i.e.* its *extension*. In particular, if a proper name occurs after a that clause, it will contribute only its *sense* to the believed proposition, where its *sense* is taken to be "the mode of presentation of the bearer's name", or "a way of determining the bearer's name". In a nutshell *sense* must be independent from reference according to (2) but its characterization preserves *extensional* features, namely, *sense* is object involving.

(B) Criticisms to *sense*'s independency characterized in (A) partially motivates a new definition of *sense* that is not object–involving and leads to a sharp distinction between *sense* and reference. Katz's new definition makes *sense* always independent of reference and of any *extensiona*l feature. According to this claim, *sense* is defined terms of *sense* properties and relations like synonymity (sameness of *sense*), redundancy (repetition of *sense*), etc.[5] Because no *extensiona*l considerations enter in the definition of *sense*, it is claimed that *sense* never determines reference. Consequently mental content referred to by a sentence occurring after a that–clause is not a "mode of presentation of the object", a "way of presentation of the reference", and so forth.

Sense–intensional role characterized in (A) has received a great deal of attention in the literature and two salient problems have

---

[4]The standard Fregean and Descriptive view held that definite descriptions associated to a name contribute a *sense* to the proposition where the name occurs.

[5]I will discuss Jerrold Katz theory in section 5 of this paper.

been discussed. On the one hand, as Stephen Schiffer argues[6], it seems to be no satisfactory candidate to play the role of "mode of presentation of the object" when belief attribution is at issue. On the other hand, the Theory of Descriptions of the Frege–Russell style has been refuted by Kripke[7], thus there is no definite description that contributes a *sense* for a name ocurring after a that clause. These circumstances left even the two classical competing theories of names, namely the Millian and the Fregean. Millianism doesn't seem to be an impalatable semantics for proper names, as it did before, therefore I will discuss (ET) to my case. Notwithstanding the failure of descriptivism, Katz claims that his his new definition of *sense* is descriptional, intensional and provides an answer to Kripke's puzzle[8] thus, I will discuss (IT) to my case.

I will proceed as follows: in section **2** I will introduce my puzzle; in section **3** I will rehearse Kripke's conclusions where (ET) is the semantics for proper names; section **4** will be devoted to a brief characterization of the new *intension*alist theory supportting (IT); section **5** will test (IT) to my case, and finally I will end with a discussion.

2 Peter's Case

I'm endorsing two Kripkean claims: (i) that the puzzle about belief attribution is quite independent of any disquotational[9] and/or translational theory;[10] and (ii) that no Description Theory of Names[11]

---

[6]I'm assuming Stephen Schiffer's conclusion, according to which there is no good candidate to play the role of mode of presentation where belief attribution is at issue. Cfr. "Belief Ascription and a Paradox of Meaning", in *Philosopical Issues* 3, E. Villanueva (ed.) (California: Ridgeview, 1993).

[7]Saul Kripke, "Naming and Necessity" in *Meaning and Use*, D. Davidson and G. Harman (eds.) (Dordrecht: Reidel, 1972) pp. 284-308.

[8]Jerrold Katz, "Has the Description Theory of Names Been Refuted?", in George Boolos (ed.), *Meaning and Method.* Essays in Honor of Hilary Putnam. Cambridge University Press, 1990.

[9]I will use the weak version of the disquotational principle: If a normal English speaker, on reflection, sincerely assents to 'p' then he believes that p; rather than the strong one, i.e.: a normal English speaker who is not reticent will be disposed to sincere reflective assent to 'p' if and only if he believes that p.

[10]I will not discuss the principle of translation because my discussion does not hang on it. I will only assume, if necessary, that if a sentence of one language expresses a truth in that language, then any translation of it into another language also expresses a truth in that other language. J. Katz rejects this principle in footnote 49, "Why Intensionalists Ought not to Be Fregeans".

[11]Saul Kripke, *Ibid.* p. 11.

and no Rigid Designator Theory will help to solve the puzzle.[12] The argument for claim (i) is based on Kripke's objections to the Description Theory of Names. While the justifications for claim (ii) are found on the very nature of rigid designators, namely, that they are modally rigid and so irrestrictly satisfy the principle of substitution for modal contexts, but they may fail for belief contexts because "... sameness of properties used to fix the reference does not appear to guarantee in general that paradoxes will not rise".[13] In order to give rise to my puzzle about belief attribution we only need the following assumption:

(I) The assent on the part of sincere, reflective speakers, implies their belief in the proposition expressed by the sentence to which their assent is given.

Let Peter be a philosopher who struggles to restore, among many of his favorite claims, first–person authority, individualistic narrow content notion, *a priori* knowledge, the unnaturalizability of content, and so forth. He has read enough about Burge's, Davidson's, Donellan's and Putnam's thought–experiments, but he remains convinced that one cannot hold both the externalistic conception of language and the transparency of content. Peter has argued in a very suggestive way that in the so called "semantics of travel" there should be a point where thoughts do not switch notwithstanding the world to which the subjects of belief are tranferred to. As arguing against these theses is not that easy, Peter overcomes the stress that externalist's arguments impose on him by enjoying large sessions of Opera.

Peter loves opera singers. Among them, he specially enjoys Pavarotti's performances. He knows everything about Pavarotti's personal and profesional life. He recognizes how Pavarotti takes the top notes in a skilfully managed and probably quite powerfully edged, head voice. He believes that Pavarotti's voice is remarkable for its quality throughout a wide range because, he argues, in many opera singers when the high notes are free and strong the low ones will be relatively feeble and colourless, but in Pavarotti's case the tone remains firm and full–bodied throughout. Thus Peter sincerely reports: "Pavarotti is the King of the High C's". Accordingly, we are entitled to ascribe Peter his believing the proposition expressed by the sentence to which his assent is given:

---

[12] *Ibid.* *Cfr.* footnotes 7, 10 and 43.
[13] *Loc. cit.* footnote 43.

(1) Peter believes that Pavarotti is the King of the High C's.

Later on, learning that Pavarotti will perform *Il Trovatore* Peter rushes through to buy a ticket. However, unbeknown to him, while listening to Pavarotti's performance, Peter is transported to Twin–Earth, suffering no disruption in the continuity of his mental life. After listening to Twin–Pavarotti's feeble and colourless high C's, Peter sincerely says: "Pavarotti is not the King of the High C's". Once again, we are entitled to attribute to Peter his believing the proposition he assents to:

(2) Peter believes that Pavarotti is not the King of the High C's.

Since Peter was transported to a new enviroment without suffering any disruption in his mental life, without noticing the change surrounding him, and with no chance to discover he is in a new circumstance, Peter can easily arrive at the assumption that the name 'Pavarotti' refers to one and the same singer he is familiar with. In fact, the only salient difference that Peter can appreciate has to do with Pavarotti's tones while singing the High C's. Peter may acknowledge that the way Pavarotti sings has changed. And nothing seems to lead Peter to hold the beliefs reported in (1) and (2) which to him would be explicit contradictory.

Thus, the most natural suggestion is to say that Peter is assuming the co–referentiality of the tokens of the same type. On these basis, Peter may dismiss either of his two beliefs, (1) or (2):

(a) He could assent to (1) 'Pavarotti is the King of the high C's', because, Peter may think that an unsuccessful performance is not good enough evidence to stop believing that Pavarotti is the King of the high C's, after all induction is not reliable!

(b) Or maybe he could accept that Pavarotti is not the King of the high C's, say, because Peter thinks that he lost his voice! Thus he would change his belief reported in (1) in favor to that reported in (2). In fact, under the co–referential assumption we might regard Peter's assent to (2) as implying his disbelief in (1).

This is the description of Peter's case as well as of the two possible situations for Peter to rationally take. Our problem now is to accomodate these data in accordance with any of the two (ET), (IT) semantics for proper names when belief attribution is at issue.

3 (ET) Semantics and Peter

The core idea of Saul Kripke in his "Naming and Necessity" is that the Frege–Russell Theory of Descriptions is false. Not only Millians have a problem in attributing a belief reported after a that clause, but also Fregeans. Even worse, Kripke's criticism has left no *sense* at all for proper names given the fact that *sense–extensional* role fails because no definite description provides necessary and sufficient conditions to determine the reference of a proper name. *A fortiori*, *sense–intensional* role also fails because if a proper name occurs after a that–clause, it would not contribute any descriptive *sense* to the believed proposition. Thus let's deal in the first place with the thesis that names get their reference unmediated by any notion of *sense*.[14] The bare claim that I will discuss, namely, (ET) is that the meaning of a proper name is its bearer:

(ET) The meaning of a proper name is determined by its bearer.

Situations (a) and (b) just described assume that Peter takes the token–names to be co–referential, that's why he could be epistemically justified in any of these cases. Thus, our attribution of belief to Peter should honor his taking the token–names to be co–referential. Accordingly, the explanation of any of these situations requires a semantic theory for proper names that matches this feature.

As we have seen, Peter satisfies assumption (I), however if we assume (ET) to be the semantics for proper names, we will obtain different mental contents for each token of 'Pavarotti', because each token bears on different singers. Thus, theoretically, we would attribute to Peter two different beliefs, say P and $-Q$. Consequently, the explanation of the situation (a) cannot get started as it depends on the fact that Peter may think that one and the same Pavarotti had an unsuccessfull performance which is not good enough evidence to count against Pavarotti's being the King of the High C's.

Similarly, the explanation for Peter's change in belief appeals to the premise that his assent to (2) implies his disbelief in (1), but surely $-Q$ can hardly be the negation of P. (ET) Semantics determines two different mental contents to the believer, while the believer thinks that the content of one belief negates that of the other. We

---

[14] I will discuss in another paper Burge's and McDowell's theses. Let's recall that I'm assuming that no definite description of the Frege–Russell style serves as the *sense* of a proper name, neither Kneale's metalinguistic theory of names does. I also assume Schiffer's conclusion that there is no good candidate to play the role, of "mode of presentation" and the like.

can certainly think of Peter as having two contents widely individuated, but we can hardly be entitled to attribute to Peter those wide contents as he is in no position to see how every occurrence of the name 'Pavarotti' bears on different singers. It is just because he cannot see the difference, that he is internally and epistemically justified as described in situations (a) and/or (b).

Let us summarize. Peter satisfies assumption (I), but it is assumption (ET) which conflicts with Peter's mental contents. We have considered two possible situations according to the natural assumption, on Peter's part, that the names are co–referential. He may be rationally described as in situation (a) or (b). However to give any explanation of those situations we require co–referentiality on the believer's part while the semantical assumption (ET) precludes it. Hence, these explanations cannot take off the ground.

The bare diagnosis here would be that if our semantics for proper names can give no more than the bearer of the name to contribute to the believed proposition, it can easily happen that the bearer of a proper name conflicts our belief attributions. The obvious way out would be to append something over and above the reference and the name, that is, the old recipe of having one and the same *sense* for each token. However, if Kripke is right and the Classical Theory of Descriptions is wrong, no *senses* have been left to do that task. Proper names have no *sense* to contribute to the proposition. Consequently there is no way to honor Peter's rationality.

4 *Intension*alist Background for (IT)

Katz' *intension*alism promises to provide us of a notion of *sense* that is descriptional, utterly different from that of Frege's, which can play a role in the explanation of situations (a) and (b). Although Katz holds that the expressions of natural language have *sense* as well as reference, that *sense*s are abstract objects and that they are the proper objects of study in a theory of meaning, his semantics differs from Frege's by depriving the *sense* of its traditional *sense–extensional role*, but agrees with Frege in that the *sense–intensional role* should be preserved in order to explain the features[15] that motivated the celebrated *sense–reference* distinction.

His theory differs from Frege's and Russell's in that the *sense* only provides necessary conditions to fix the reference and in that it is

---

[15]They were discussed in section **1**.

definied in the Metatheory of *sense* by means only of intensional notions. The Metatheory is about the form of semantic descriptions and is not commited to any *intensions* for terms in natural languages, it is concerned with general features of *sense* descriptions for natural languages.[16] Katz unlike Kripke does not hold that the puzzle is troublesome for all theories of names. He thinks that it seems to be a problem for everybody because Millians have no notion of *sense* to restore *de dicto* belief,[17] and descriptional theories are false as Kripke did show. Thus, his diagnosis is that a notion of *sense* is missing and his strategy consists of providing an *intension*al semantics for proper names to avoid the puzzle. However, I will claim that if my case is not problematic for Katz's Meta–theory, then Kripke's Paderewski–case is a problem for his theory. The other way round holds too, I mean, if Paderewski–case can be explained by the Katzian theory, then my case would be a problem for it. I will present briefly, in the first place, Katz's theory and I will discuss my case. Later on I will go back to Kripke's Paderewski–case to conclude that Katz cannot have both puzzles solved.

Jerrold Katz has thoughtfully argued that a Theory of Meaning should be a theory of *sense* only. He defines the *sense* of an expression in terms of *intension*al notions, i.e. as that aspect of its structure which is reponsible for its *sense* properties and relations[18] like synonymy (sameness of *sense*), antinomy (opposition of *sense*), redundancy (repetition of *sense*), etc. and ends up with a twofold consequence. On the one hand, his definition makes room for a sharp distinction between *sense* and reference;[19] on the other, his notion of *sense* alone does not suffice to determine reference, hence, we're told, *sense* only mediates reference.[20] For *sense* to mediate reference is to give necessary conditions which along with referential principles and context fix the reference of the term. These referential principles should be stated in a theory of reference and should provide the necessary and sufficient conditions for an object to fall under the *sense* of the term.

---

[16] For instance, whether such descriptions are decompositional and compositional. *Cfr.* J.J. Katz, *Semantic Theory*, New York, Harper and Row, 1972.

[17] I agree on Katz's diagnosis and bet that he will ammend his proposal to answer the problems I will discuss in this paper. In fact, the way out to Kripke's puzzle is to restore *de dicto* belief by means of a new notion of *sense*.

[18] *Cfr.* "Prècis of *The Metaphysics of Meaning*", in *Philosophycal Issues* 4, 1993, E. Villanueva (ed.), Ridgeview Publishing Co. p. 131.

[19] *Cfr.* "Reply to Boghossian", in *Philosophycal Issues* 4, 1993, E. Villanueva (ed.), Ridgeview Publishing Co., p. 145.

[20] Jerrold Katz, *Ibid.* p. 37.

According to Katz, descriptivism is simply *intension*alism applied to names[21] and, he argues, his descriptional theory of names doesn't fall prey to Kripke's well known objections. He claims that a proper name 'n' has a metalinguistic *sense* (S), a *sense* that contains a concept uniquely representing the grammatical form of the proper name 'n' itself, and a concept representing the name relation. For the present discussion I will assume Katz's definition of *sense* (S) as the *intension*al semantics for proper names:

(S) The *sense* of a proper name 'n' = The contextually definite thing[22] which is a bearer of 'n':

In (S) the indefinite article points to a matter of the existence of multiple, contextually independent, correlations of a proper noun with bearers; while the definite article singles out definiteness as a matter of contextually dependent referential uses of a proper noun: "This definiteness can be understood as, roughly, the contextually fixed unique reference of a speaker's use".[23] Notice that the *sense* of a name is synonymous with the substantive phrase "The contextually definite thing which is a bearer of 'n'". Proper names are not predicative on Katz's theory neither their *sense*s are, and this features help to avoid Kripke's objections to Metatheories.[24] Given the fact that *sense* is taken to be a pure *intension*al notion, as (S) defines it, Katz holds that the *sense* of a name doesn't suffice to determine its reference, and claims that a proper name contributes only its *sense* to the believed proposition.[25]

5 (IT) Semantics and Peter

We have a *prima facie* candidate to restore *de dicto* belief and answer the puzzle under the new *intension*al construal. Our assumption for the semantics of proper names is (IT):

---

[21] Jerrold Katz, "Has The Description Theory of Names Been Refuted?", pp. 33-4.

[22] 'Thing' would be a categorized variable, to be replaced by a semantic representation of the concept of a sentient creature. See the discussion in J.J. Katz, *Semantic Theory*, pp. 259-60.

[23] J.J. Katz, "Why Intensionalists Ought not to Be Fregeans", p. 86.

[24] This thesis is used as a premise to forbid the substitution of 'Nixon' by 'a bearer of "Nixon" ' in 'Nixon might not have been Nixon' that would lead us from a false statement to a true one.

[25] In what follows I will assume, for the sake of argument, that Katz's semantics is immune to Kripke's objections.

(IT) A proper name contributes only its *sense* to the believed proposition.

Let's recall that an adequate *de dicto* belief attibution to Peter, should be sensitive to the false assumption, on Peter's part, that the two tokens of the name 'Pavarotti' co–refer, *i.e.* to Peter's understanding they have the same *sense*. Our problem was that the bearer of 'Pavarotti' is, as it were, one and the same to Peter's knowledge, while the names bear on different singers. The contextual information available to Peter does not suffice for separating the reference. The reason should be obvious. The externalist set–up for this case prevents this task. Peter will be endlessly told that "Pavarotti is identical with Pavarotti". Accordingly, the (S)–determined mental content could deliver (*) for any of the two occurrences of 'Pavarotti'

(*) The contextually definite thing which is a bearer of 'Pavarotti'.

It doesn't matter whether 'Pavarotti' refers to the Earthian or the Twin–Earthian singer. No considerations about what a name refers to are relevant for the grammatical structure responsible of the *sense* properties: "Facts about what this or that word is true of are not grammatical facts [*sense* facts], but referential facts".[26] If we are to satisfy this constrain we can only count on the grammatical structure that provides the *sense* of a proper name, namely, (*).

Thus, explanations of situations (a) and (b) can get started, either (i) Peter beholds his belief that Pavarotti is the King of the High C's or, (ii) we take Peter's assent to (2) as implying his disbelief in (1), *i.e.* Peter believes that Pavarotti is not the King of the High C's.

Let's summarize. We are told that we count only with *sense–grammatical* facts that do not determine reference. The definite and indefinite articles in (S) play mainly grammatical roles: (I) the definite article would single out the bearer of the name only upon contextual conditions of utterance and (ii) the definite article just points to the fact that a name may have more than one bearer, if any. Katz has claimed that there is no commitment in his theory to the reference, only to *sense*, in order to individuate Peter's mental content (beliefs).

However, this answer rests upon an ambiguity. Either (C) *sense* properties alone determine mental content (*de dicto* belief), or (D) *sense* properties and contexts determine it. The former answer

---

[26] Jerrold Katz, "Reply to Boghossian", p. 147.

could render Peter rational,[27] but it also could render Pierre irrational in Kripke's Padervski–case. The second answer could render Pierre rational, but my Peter couldn't be epistemically justified in any of the two (a) and (b) situations above described. Let's look at each case:

(C) *Sense properties alone determine mental content.* If only *sense* properties determine mental content, then every token of the same type would have to have the same *sense* of its type. My Peter is rational just because the (*) instantiation of (S) structure delivers the same *sense* for every token of 'Pavarotti'. However we cannot paralell this answer for Kripke's Paderevski–case. Let's look at the reasons. Kripke's Paderevski–case is one in which two tokens of the same type, namely 'Paderevski', refer to one an the same person and Pierre, the believer, mistakenly take them to bear on different people. Thus, in order to render Pierre rational we need to posit different *senses* for every token of 'Paderewski', as Pierre holds both that 'Paderewski had musical talent' and that 'Paderewski had no musical talent'. However, the instantiation (**) of (S) would deliver the same grammatical structure for each token–name, *i.e.*

(**) The contextually definite thing which is a bearer of 'Paderevski'.

Consequently we would attribute the same *sense* to each token while we require them to be different. In a nutshell, there seems to be no principled distinction to hold both that the (*) instantiation of (S) delivers sameness of *sense* for each token of 'Pavarotti' and the (**) instantiation of (S) delivers difference of *sense* for each token of 'Paderevski'. Thus, my case seems to be no problematic for the theory but Kripke's case is. Hence, let's look at the alternative answer (D) in order to avoid Kripke's Paderewski.

(D) *Sense properties and contexts determine mental content.* In order to render Pierre rational, who believes both that Paderewski had musical talent and that Paderevski had no musical

---

[27] I am assuming that it does, for the sake of argument. However there is a problem in situation (b) as it depends on the fact that (2) negates (1). For (2) to negate (1) we badly have to have truth–values. But in order to have truth–values we have to have the semantic value of a proper name, i.e., we have to determine the referent of the proper name notwithstanding the fact that (S) doesn't suffice to determine reference.

talent, we can get two different *senses* from (**) as Katz himself suggests. The referential indicator in (**), *i.e.* 'bearer of "*n*" ' is bare of identifying attributions, but it has to have some sort of place–holder in such a way that the attributions that are put there function as constrains on the object that can be the bearer of the name:

> ...Peter learns the name 'Paderevski' 'with an identification of the person named as a famous pianist'. Thus when we express what Peter believes in this connection, our expression of it will be: (16) Peter believes that Paderewski the famous pianist had musical talent.[28]

However we cannot pararell this answer to my case. As my case is described Peter uses the name 'Pavarotti' always relative to "musical contexts" so to speak and relative to a musician which Peter takes to be one and the same person. Once again, we have no principled distinction to hold both, that (**) instantiation of (*S*) gives rise to different *senses* for each token of 'Paderevski' while (*) instantiation of (*S*) gives rise to one and the same *sense* for each token of 'Pavarotti'.

(c) As the case is described, there are two kind of conversations where Peter hears the name, one has to do with politicians and the other with musicians. Assuming that each kind of conversations leads to a difference in contexts we can read off from the same (**) structure, two different *senses* for each token of 'Paderevski' and associate each one with the relevant conversation.

However reading (c) won't work for my case, as follows

(c′) As the case is described, by definition, Earthian and Twin–Earthian contexts differ, we have Earthian things and Twin–earthian things. In the absence of an explicit explanation of what "contextually" means in (*S*) and for by the same trend of reasoning used in (c), we could read off from the same (*) structure two different *senses* for each token of 'Pavarotti' and associate each one in accordance with the relevant Earth/Twin–Earth context of utterance. In so doing no explanation could be given of the natural assumption on Peter's part of taking the names to bear on the same singer. Thus, it seems that we are

---

[28] J.J. Katz, "Why Intensionalists Ought not to Be Fregeans", p. 88-89.

> back just where we started, namely, we have the same results as those we had when we used (ET). Is there another sensitve way of understanding "the *contextually* definite thing..."?[29]

If we assume (C) that only *sense* (grammatical, purely *intension*al) properties determine mental content my puzzle does not arise. However, two undesirable situations obtain: on the one hand we end up with the claim that every token of a type–name occurring after a that–clause contributes the same *sense* to the believed proposition. On the other hand, Kripke's Paderevski case cannot be solved.

If we assume (D) that *sense* properties and contexts of utterance determine mental content Kripke's case is solved. However, three undesirable situations obtain: first, we cannot give a theoretical explanation of the natural assumption on Peter's part that the names are co–referential. Second, (D) is no relevantly different of (ET) because it delivers the same results. Third, we violate the claim that no *extension*al consideration enters into the *sense* of a word.

---

[29]I am sure that Jerrold Katz will dispell this mystery in a forthcoming paper about *sense–reference* relation.

Contributors

Daniel Andler, Centre de Recherche en Epistemologie Apliquee
Louise Antony, Department of Philosophy, North Carolina State University
John Biro, Department of Philosophy, University of Florida at Gainesville
Ned Block, Department of Philosophy, Massachusetts Institute of Technology
Tyler Burge, Department of Philosophy, University of California at Los Angeles
Manuel Campos, Department of Philosophy, Stanford University
Jerry Fodor, Department of Philosophy, Rutgers University
Manuel García-Carpinteiro, Universitat de Barcelona
Roger Gibson, Department of Philosophy, Washington University
James Higginbotham, Centre for Linguistics and Philology, University of Oxford
Paul Horwich, Department of Philosophy, Massachusetts Institute of Technology
Pierre Jacob, Centre de Recherche en Epistemologie Appliquee
Jaegwon Kim, Department of Philosophy, Brown University
Joe Levine, Department of Philosophy, North Carolina State University
William Lycan, Department of Philosophy, University of North Carolina at Chapel Hill
Christopher Peacoke, Magdalen College, University of Oxford
David Sosa, Department of Philosophy, Dartmouth College
Ernesto Sosa, Department of Philosophy, Brown University
James Tomberlin, Department of Philosophy, California State University, Northridge
Josefa Toribio, Department of Philosophy, Washington University
Lourdes Valdivia, Instituto de Investigaciones Filosóficas, Universidad Nacional Autónoma de México
Crispin Wright, Department of Philosophy, University of St. Andrews

Enrique Villanueva is Research Fellow at the Instituto de Investigaciones Filosóficas in the Universidad Nacional Autónoma de México in México City where he has been doing research and teaching since he completed his Graduate Study at the University of Oxford in 1972. He has published a number of essays in Philosophy of Mind, Metaphysics of Persons and Philosophical History. Besides editorship of the present series he has edited the series *Simposio Internacional de Filosofía*, and the volume *El argumento del lenguaje privado*, at the Universidad Nacional Autónoma de México. He is author of *Lenguaje y Privacidad, Ensayos de Historia Filosófica* and *Las personas*.

Philosophy and Phenomenological Research

AN INTERNATIONAL QUARTERLY FOUNDED BY MARVIN FARBER

ERNEST SOSA, Editor RODERICK M. CHISHOLM, Associate Editor

PPR is fifty percent longer than in the past. This has made possible some new departures with no decrease in refereed material published. For example, regularly featured book symposia include the following:

1. Samuel Scheffler's *Human Morality*. Précis by the author. Discussants: Stephen Darwall, Amelie Rorty, and Susan Wolf. Response by the author.
2. John Searle's *The Rediscovery of the Mind*. Discussants: Brian Garrett, Pierre Jacob, Jaegwon Kim, and Robert Van Gulick. Response by the author.
3. Michael Slote's *From Morality to Virtue*. Précis by the author. Discussants: Stephen Darwall, Henry Richardson, and Michael Stocker. Response by the author.
4. Christopher Peacocke's *A Study of Concepts*. Précis by the author. Discussants: Jane Heal, David Papineau, and Georges Rey. Response by the author.
5. Crispin Wright's *Truth and Objectivity*. Précis by the author. Discussants: Terry Horgan, Paul Horwich, Phillip Pettit, Mark Sainsbury, James Van Cleve, Timothy Williamson. Response by the author.
6. Philip Kitcher's *The Advancement of Science*. Précis by the author. Discussants: Isaac Levi, Richard Miller, Peter Machamer, Dudley Shapere. Response by the author.

PPR can typeset directly from disks and from text received electronically. This enables us to publish substantially more material in each issue with no increase in the time from submission to publication, or in the time required for publication of reviews and book symposia.

Subscription Order Form

Enter my one-year subscription to *Philosophy and Phenomenological Research:*

❒ Individual $20.00 (foreign, $24.00 postpaid)

❒ Institutional $55.00 (foreign, $59.00 postpaid)

❒ Enclosed is my check (made payable to *Philosophy and Phenomenological Research*).

❒ Please charge my ❒ MasterCard ❒ VISA

Card # _____

Exp. Date _____

Signature _____

Name _____

Address _____

City, State, Zip _____

Send to: Philosophy and Phenomenological Research
Box 1947, Brown University
Providence, RI 02912 (USA)


9 780924 922220 >
EAN
* ISBN 0-924922-22-2