



Australian
National
University

THESES SIS/LIBRARY
R.G. MENZIES LIBRARY BUILDING NO:2
THE AUSTRALIAN NATIONAL UNIVERSITY
CANBERRA ACT 0200 AUSTRALIA

TELEPHONE: +61 2 6125 4631
FACSIMILE: +61 2 6125 4063
EMAIL: library.theses@anu.edu.au

USE OF THESES

This copy is supplied for purposes
of private study and research only.
Passages from the thesis may not be
copied or closely paraphrased without the
written consent of the author.

Explanation

A Causally Constrained Pragmatic Account

**A Thesis Submitted In Fulfilment of the Requirements for the Degree of
Doctor of Philosophy at the Australian National University**

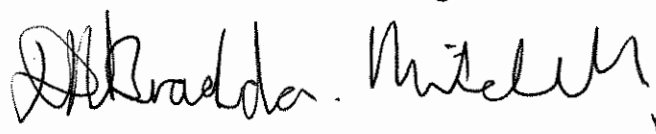
David Braddon-Mitchell

Department of Philosophy, RSSH

December 1988

For Deborah Lavers

This thesis is my own work, written while I was a Research Scholar in the Department of Philosophy, Research School of Social Sciences, at the Australian National University. As far as I am aware, all assistance and sources are acknowledged.

A handwritten signature in black ink, reading "David Braddon-Mitchell". The signature is written in a cursive, flowing style. The first part "David" is written with a large, stylized 'D' that loops around the 'B' in "Braddon". The last name "Mitchell" is written with a prominent 'M' and a trailing flourish.

(David Braddon-Mitchell)

Acknowledgements

First I must acknowledge the assistance of my supervisors, Frank Jackson and Philip Pettit. My intellectual debt to them will be obvious to any reader of these pages; what will not be so obvious is their generosity with helpful comment, and infectious enthusiasm for the discipline. For all this I thank them.

An earlier version of the Appendix was read to the Brisbane conference of the AAP in 1987, and to the Department of Philosophy in the RSS. Earlier versions of chapters 2 and 6 have been read to seminars in the RSSS, and the Department of Philosophy in the Faculties, respectively. I have in all cases benefited from the discussion. Much of the material in Ch. 7 is from a paper jointly written with John Fitzpatrick, a version of which was read to the Department of Philosophy in the RSSS. We both, I can say with confidence, benefited from the comments arising from the discussion.

Many people have helped in ways which have borne directly on the writing of this thesis. I have, where possible, mentioned particular points of indebtedness where they occur in the text. The following is a (probably incomplete) list of those who have helped by discussing material from the thesis, or in other direct ways: Barbara Davidson, Martin Davies, Kathinka Evers, John Fitzpatrick, Paul Griffiths, Virginia Hart, Loraine Hugh, Huw Price, Dominic Hyde, Deborah Lavers, Libby Prior, Kim Sterelny and Mark Walker.

The writing of a PhD thesis is not something which is causally isolated from those around its author. It is customary to acknowledge those who have in some way helped in its production. This seems to me a rather thesis-centred view of the world; one should in fairness add to it

those who have been affected *by* it. Amongst those who have both helped in, or put up with, the production of this thesis I must mention Marina Farnan, Virginia Hart, Margaret Jones, Geoff Kennett, Deborah Lavers, Mark Mitchell, Lisa Paul, Maureen Sheehan, Julian Thomas and Richard Webb.

Abstract

This thesis argues that explanation consists in giving pragmatically selected and often counterfactual information about causes. It begins by examining pragmatic theories of explanation, and argues that they are right insofar as pragmatics are properly a part of explanation, not just an adjunct to explanation. In the following chapters it is argued that while pragmatics may be *necessary* for an account of explanation, they are not sufficient. Additional constraints are imposed regarding the need for explanatory information to be causal information, and usually counterfactual explanation if higher-level explanation is to be accounted for. The account thus developed is then applied to problems in metaphysics, philosophy of psychology and philosophy of language. It is argued that causal reductionism can seem more plausible in the light of a good account of explanation, that propositional attitudes can be seen to have explanatory virtues (on a certain account), that Fodor's Language of Thought Hypothesis suffers in the light of the right account of explanation, and that taking into account certain interest dependencies in explanation makes the development of an account of the reference of kind terms easier.

A Note On Format

Each chapter has its own Table of Contents; these are quite detailed and include all sub-headings. The Abbreviated Table of Contents that follows includes only the highest level headings in each chapter.

Abbreviated Table of Contents

Chapter I General Introduction	2
--------------------------------------	---

Part One

Chapter II The Pragmatic Theory of Explanation	10
1 Historical Prologue: The Covering Law Model.....	12
3 Salmon's Probabilistic Account	19
4 Context and Contrast.....	27
5 Explanatory Relevance and the Problem of Arbitration	47
6 Van Fraassen's Logic of Why-Questions and Some Variations.....	53
7 Answers and Telling Answers.....	61
Chapter III The Causal Constraint.....	63
1 The Asymmetries Revisited and the Causal Constraint	64
2 The Tower and the Shadow.....	66
3 The Causal Constraint.....	76
4 The First Constraint.....	78
5 The Second Constraint.	85
6 The Third Constraint.....	87
7 Some Problems of Generality.....	88
8 Extension to the Probabilistic Case.....	92
9 The Limits of Causal Relevance.....	96
10 What kind of explanation is a causally constrained pragmatic account?.....	100
Chapter IV Levels and Explanation	104
1 Introduction.....	105
2 Levels of Description.....	107
3 Structural and Functional Levels of Description.....	114
4 Program and Process Explanations.....	132
5 Causally Efficacious Properties.....	141
6 The Constraints Reformulated.....	149
Summary of Part One, with some Last Remarks.....	153

Part Two

Chapter V Explanation and the Microstructural Cause Hypothesis.....	157
---	-----

1	Introduction.....	158
2	Why We May Want To Restrict Causation to the Bottom Level.....	159
3	The Competing Accounts.	163
4	The Case Against Causal Reductionism.....	164
5	Kim's Event-Supervenience	164
6	Menzies' Objections.....	166
7	Varieties of Supervenience and Menzies' Counterexamples.....	171
8	Explanation and Supervenient Causation.....	183
9	Against Supervenient Causation.....	184
10	Explanation and Event Identity.....	187
11	Summation.....	188
Ch. VI The Explanatory Rôle of Folk Psychology.....		190
1	Introduction.....	192
2	Positions on the Relationship.....	207
3	Explanation at Higher Levels and Broad Psychology.....	218
4	External Functionalism.	238
Chapter VII Explanation and The Language of Thought.....		259
1	Introduction.....	259
2	The Language of Thought.....	260
3	How Not To Get A LOT For Free	271
4	Explanation I: Synchronic and Diachronic Explanations.....	276
5	Explanation II: Implementation and Levels of Explanation.....	282
6	Conclusion	290
Appendix Explanation and the Reference of Kind Terms		292
1	Why-Questions and Kind Terms.....	292
2	The Programme.....	296
3	Causal Theories of Natural Kind Terms and the Sameness Relation	297
4	Natural Kinds and Natural Kind Terms.....	302
5	Reference Shift and the Bohr Atom.....	302
6	Some Twin-Earth Cases	306
7	Final Remarks.....	312
References		313

Part One

Chapter One

General Introduction

Table of Contents

Introduction.....	2
The Programme.....	3
The Court of Explanatory Theory	4
Some Minor Remarks (and a confession).....	6
A Last Preliminary About my Subject Matter.....	8

Introduction

Explanation is the pragmatic selection of usually counterfactual information about causes. This is the central thesis of this dissertation; and it is one which cuts across most of the received taxonomies of conceptions of explanation. The received taxonomy is best represented by Wesley Salmon's excellent book *Explanation and the Causal Structure of the World* [Salmon 1984], in which he divides conceptions of explanation into three kinds; *ontic* ones which are concerned with the real causal structure of the world, *modal* conceptions, and *epistemic* conceptions

I think explanation is all of these, if not necessarily exactly in Salmon's senses. It is ontic because it is concerned with real causal structures in the actual world; it is modal because it is also concerned with what would happen in worlds where the actual structures are different, and it is epistemic because not all information about causes is explanatory—only causal information which is an answer to a pragmatically disambiguated why-question is explanatory.

These are, roughly, the lines I shall be running, and in due course I shall outline how I plan to proceed. But my immediate concern is to suggest that they are lines that *matter*. The theory of explanation can sometimes seem like a dreary area; as though all that it involves is trying to find an account of explanation that will allow as many as possible of what seem intuitively to be explanations to indeed count as such—and of course will prevent us counting as explanations things which intuitively it would seem outrageous to count as such. But generally speaking the fitting of theory to intuition has not proceeded the other way. Theories of explanation have been tried in the court of intuitive judgement about what counts as an explanation, but rarely has anything been explicitly tried in the court of the

theory of explanation. I propose to apply the theory of explanation to some problems in various other areas, and find solutions to the problems in these areas in the light of what, independently, seems to be the best account of explanation.

It is this possibility which makes the theory of explanation exciting. In this thesis I will argue that the right account of explanation allows propositional attitudes—on a certain account—to have explanatory virtues, that it undermines some of the motivation for Fodor's language of thought hypothesis, and that it can protect the programme of causal reductionism from various objections. I even claim that we can be clearer on issues such as the reference of natural kind terms if we bear in mind certain features of a theory of explanation. In short, there's *lots* of work for a theory of explanation to do, so we had better have a good one.

The Programme

The thesis is in two parts. The basic account of explanation is outlined in Part One, and its chapters (chs. 2, 3 and 4) correspond roughly to the three components of the account. Ch. 2 deals with the pragmatic components, the elements close to some versions of what Salmon calls the epistemic conception. Here I argue, with van Fraassen [van Fraassen 1980], that taking pragmatics right into the heart of the account is necessary if many of the problems which have faced earlier accounts of explanation are to be solved. In the next chapter I argue that while the pragmatic theory is right in taking a pragmatic apparatus to be a *necessary* feature of an account of explanation (rather than an additional feature not even partly constitutive of explanation), it is not sufficient. I argue that a causal constraint has to be placed on explanation if we are to find principled reasons for preventing the mushrooming of supposed 'explanations' that would be licensed by a purely

pragmatic account. At this point two features of the account will have emerged: that explanation is the giving of information about causes, but not the giving of *any* information about causes. Only information licensed by the pragmatics is explanatory. So explanation as such remains a pragmatic notion, sensitive to explanatory purpose and context, but not unconstrainedly so. It is constrained by the requirement to provide information about causes, and what counts as information about causes may not be a pragmatic matter.

In ch. 4, I add the final ingredient to the mix. I argue that explanation can involve the giving of *counterfactual* information about explanation. We are not just concerned with information about the *actual* causes of something to be explained, we are also concerned with information about what *relevantly similar* causes might bring about states of affairs relevantly similar to those which are to be explained. This is one way of capturing explanatory generalizations. I approach the problem by way of an account of *higher-level* events, objects and so on, which may supervene on levels of nature where causal processes work. It turns out that it is by giving counterfactual information about causal processes that one supervenient event is able to explain another, on the assumption that the explanation should hold good for all worlds in which the token events are realized, even if realized differently at the causally efficacious level of nature.

The Court of Explanatory Theory

I mentioned above that what makes explanation exciting as a topic is that views about explanation have ramifications elsewhere. Part Two is concerned with these ramifications. It begins with ch. 5, which deals with problems emerging from the discussion of explanation and causation in the first four chapters. In this respect it is a kind of transitional chapter; while it

is primarily concerned with using the theory in Part One to provide a defence of causal reductionism against certain objections, it also develops and amplifies the account of the relationship between explanation and causation offered in ch. 4. The objection to causal reductionism with which ch. 5 is primarily concerned is that which has been recently raised by Peter Menzies [Menzies 1988]. He claims, roughly, that there are features of our ordinary discourse about causes which makes it impossible to claim that much ordinary talk about causation in the macroscopic world is reducible to *ur*-causal relations in the microstructure.

I provide a diagnosis of this problem as follows: in ordinary discourse about causal relations we are making claims about *counterfactual* relations. So of course this talk will not be reducible to talk only about actual causal relations in the microstructure. It will, however be reducible to counterfactual claims about causal relations in the microstructure. Since, in addition, pragmatics constrain what are often considered to be causal relations between macroscopic entities, a familiar pattern emerges: macroscopic 'causation' is the giving of causally constrained counterfactual information about microscopic causes governed by a pragmatics. This is, of course, the account of explanation I am defending. So perhaps the right way to characterize macroscopic 'causation' is as *explanation*: and causal reductionism might be true if the only kinds of causal facts that explanations depended on were ultimately microstructural ones.

With these thoughts in mind, I turn my attention in the next two chapters to the philosophy of psychology. In Chapter 6 I argue that a certain view of folk psychological entities, which I call external functionalism, can be construed as giving explanations insofar as it provides counterfactual information about causes in a way that would be redundant if one's account of explanation required that explanations always cite only actual causes.

Chapter 7 has a more negative claim; that Fodor's Language of Thought hypothesis is motivated by a view of explanation which requires that high-level causes be cited in the explanation of behaviour. I propose instead the replacement of that view by one which requires that there be *some* causes of behaviour, but that they need not be at a high level, and that just so long as there is a motivating story about why we should expect there to be such causes (i.e. we have provided counterfactual information about the causes) we have met the explanatory requirement. On this view, much of the motivation for the Language of Thought hypothesis evaporates.

Finally, I provide an appendix in which I give an account of the reference of natural kind terms which takes into account the explanatory motivations that underpin our use of such terms, and the pragmatic component of explanation. I include it as an appendix rather than a chapter because while certain pragmatic, interest dependent features turn out to determine the reference of kind terms, and in addition explanatory motivations play a rôle, you do not have to buy the detail of the account in Part One to accept the position outlined in the appendix.

There is no conclusion, though there is a summary of Part One at the end of Ch. 4. Conclusions to theses tend to read like versions of their introductions, with all the future tense verbs moved into the past. This introduction outlines what I am attempting; it is the reader's job to judge how far I have succeeded.

Some Minor Remarks (and a confession).

It is difficult to discuss explanation without using some terms that mean, roughly, 'thing to be explained' and 'thing which does the explaining'. In order to avoid unnatural Latinisms I first tried using these very phrases; but the result was tedious, inelegant and confusing. Variations on them turned

out to be even worse; the search for variety was successful only with considerable loss of clarity. So I have fallen back on the artificial Latin constructions popularized, amongst philosophers, by Hempel (he even gets a mention in the Supplement to the *OED*). So *explanandum* and *explanans* it is; 'explanandum' for that which is to be explained; 'explanans' for that which does the explaining. I have not distinguished terminologically between cases where people think that it is sentences or propositions which explain or are explained, where information explains, or where events or things explain or are explained. In particular I do not distinguish the use of 'explanans' to denote the information that an event occurs, and its use to denote the event itself. I hope that it will be clear from the contexts in those cases where the difference matters.

The plurals of these words are regrettably very useful. I have been torn between pedantry and a desire to Anglicize the words. Partly because of the thought that 'explananses' is very ugly, and looks like a mistaken Latin plural, I have decided to be Latinate the whole way. I use *explananda* as the plural form of *explanandum*, and *explanantia* (the plural form of the present participle) as the plural form of *explanans*.

On a more serious note, a confession. In various chapters I help myself to a possible worlds analysis of modality, since I think such analyses are the clearest way of formulating modal thoughts. But I am not prepared to pay the price. I am not a modal realist. This would be fine if I thought that there was a convincing account of modality available in terms of *ersatz* worlds, but I know of no convincing reply to David Lewis's [Lewis 1986e p. 136ff.] objections to such accounts as presently on offer. Those who are modal realists, or who do not share my conviction that Lewis has shown that there are serious problems for extant versions of *ersatzism* will not think that my use of the possible worlds idiom is problematic. The rest of us just have to

live in hope that a better account of the utility of possible worlds talk will be forthcoming. All I can say is that I do not know of a better way of dealing with modal notions.

A Last Preliminary About my Subject Matter

Sometimes the word 'explanation' is used to cover things to which my account could not possibly apply. Sometimes we say that the fact that some expression is a theorem is explained by certain lemmas being provable, or by the fact that proofs of a certain kind tend to go through. Presumably no information is given about causes here, counterfactual or otherwise, constrained or unconstrained by pragmatics. Words in ordinary language are always used in a variety of ways, and the philosopher attempting to give an account of things, always has to make a choice about how much of the actual usage she aims to cover. Often this means difficult choices have to be made. Try to include too much, and the theory will look weak and unmotivated. This is what I think happens to those philosophers who try to make a theory of explanation cover *every* use of the word; all that might be found in common between all uses, for example, is that explanations somehow 'increase understanding'—which leaves understanding as a difficult notion in need of explication [see Achinstein 1983]. Cover too few of the uses, and one is open to the charge that you've in some culpable sense done too little of the work.

I hope that I have done just enough of the work. I could be content, like Lewis [Lewis 1986d], just to stipulate that I am talking about *causal* explanation—and thus rule out the mathematical sense. I want to claim a little more, however. I think that there are lots of contexts in which no information other than causal explanation can with any plausible explanation be said to be explanatory. So I think that the causal sense of

explanation is a primary one; only information relevant to causes properly answers why-questions in most contexts, as I argue in Ch. 3. This is not because the contexts suggest that, out of the many kinds of explanation that could be offered, only causal ones are appropriate. It is because it is only causal information that could be explanatory in these contexts; nothing else would answer to our intuition that it is only information however distantly about causes that explains *why* something is so in these contexts. So I do not view causal explanation as a *kind* of explanation, and mathematical explanation as another kind. Instead I think that the word 'explanation' is being used in these mathematical cases to denote a *different*, if related, notion. In some sense such stipulations are arbitrary, but the bottom line is this. The theory of what mathematical explanation and causal explanation have in common is not very interesting or broadly applicable. I hope to show that the theory of explanations in causal contexts is both; so if we think that explanation is an important notion which can constrain our theorizing in various ways, we had better reserve it for the latter enterprise.

Chapter Two

The Pragmatic Theory of Explanation

Table of Contents

1	Introduction	12
2	Historical Prologue: The Covering Law Model	13
	2.0.1 The Symmetry Thesis.....	14
	2.0.2 Problems of Selection and Completeness	15
	2.0.3 The IS Case.....	17
	2.1 The Problems to be Solved	18
3	Salmon's Probabilistic Account	20
	3.1 The Probabilistic Account and the Problems for Explanation.....	23
	3.1.1 The Problem of Rejection.....	23
	3.1.2 The Asymmetries.....	24
	3.1.3 Illegitimate Inclusions and Exclusions	25
	3.1.4 Arbitration	26
	3.2 Conclusion	28
4	Context and Contrast.....	28
	4.1 Contrastive Focus.....	30
	4.2 Contrastive Focus and Contrast Class	31
	4.3 Contrast Class and the Problems for Explanation	34
	4.3.1 The Problem of Rejection.....	34
	4.3.1.1 Contrast Class and Paradigm Shift	35
	4.3.2 Asymmetries.....	37
	4.3.3 Exclusions and Inclusions	38
	4.3.3.1 Choice Class and Explanatory Completeness.....	39
	4.3.4 Arbitration	40
	4.3.5 Contrast, Choice and Explanatory Irrelevance	41
5	Explanatory Relevance and the Problem of Arbitration	47
	5.0.1 An Example	49
	5.2 Interest and Purpose Relativity.....	51
	5.3 Relevance Relations.....	52

6	Van Fraassen's Logic of Why-Questions and Some Variations.....	53
6.1	The Topic and the Contrast Class.....	55
6.2	Choice Class of Explanans.....	56
6.3	The View Reformulated.....	58
6.4	Choice Class of Explanandum and Explanans.....	60
7	Answers and Telling Answers	61

1 Introduction

In this chapter I argue that pragmatics are a necessary part of a theory of explanation, rather than an additional feature which determines which explanations you want from a range of perfectly acceptable, fully explanatory alternative explanations with the same explanandum. I do this through a sympathetic examination of the so-called pragmatic theory of explanation—a theory which arises from the work of—*inter alia*—Fred Dretske [Dretske 1978], Alan Garfinkel [Garfinkel 1981] and Bas van Fraassen [Van Fraassen 1980]. By the end of the chapter I hope that the necessity claim made by this kind of theory will seem compelling. I set the (more or less!) sufficiency claim—pressed at least by van Fraassen—aside, and will in fact argue *against* it in the next chapter.

I start by way of a prologue with a piece of modern history; a brief account of Hempel's covering-law model, and some of the reasons why it is history. I then motivate the examination of the pragmatic component of explanation through the deficiencies in a representative post-Hempelien but pre-Pragmatic theory—the earlier one of Wesley Salmon [Salmon 1971]. I shall then examine two leading ideas of pragmatic theories—contrastive difference and explanatory relevance—to see how they fare against these difficulties. At the same time I hope to show how apparently differing formulations of the theory in the literature can be reconciled under a more general theory of contrasts. I end with a formulation of the pragmatic elements of explanation which is suited to act as the framework onto which the other elements will be added in later chapters.

2 Historical Prologue: The Covering Law Model

Modern work on explanation really begins with Hempel's accounts of explanation in [Hempel and Oppenheim 1948]. The basic idea is that explanations are logical arguments, with the explanantia featuring as premises in the arguments, and the explananda featuring as conclusions. This idea became common to a range of views which differ in detail, such as those of Braithwaite [Braithwaite 1953] and Ernest Nagel [Nagel 1961] and Karl Popper [Popper 1959].

Hempel actually had *two* accounts of explanation; the original *deductive-nomological* (DN) theory of explanation, and the later *inductive-statistical* (IS) account (revised and improved in [Hempel 1965]) to cover cases of statistical explanation. What they have in common is that on both account explanations are arguments, and in both cases the premisses must include law-like generalizations. Hence the general description 'covering-law' account. Where they differ is that in the DN case the explanation is a *deductive* argument featuring at least one universal generalization amongst the premises and which yields the explanandum as a logical consequence. In the IS case the explanation is an *inductive* argument featuring at least one statistical generalization amongst its premises, yielding the conclusion that the explanandum has a high probability.

What made this the beginning of modern thought on explanation was that it conferred some kind of respectability on explanations. Previously, in an environment dominated by a positivist hostility to theoretical terms, it was very unclear how a scientific *theory* could explain an event. With the advent of the DN and IS models this became possible; since explanations could be concerned with logical relations between terms, it didn't matter if you weren't a realist about those terms. This opened up the possibility of a

certain kind of *instrumentalism* about explanation. Theoretical terms were already regarded as instrumentally useful for *prediction*, and construing explanation as a logical relation between (*inter alia*) theoretical terms, with much the same sort of logical structure as figures in arguments used to make theoretical predictions, meant that some of the same instrumental benefits might be enjoyed in the explanation case as had been enjoyed in the prediction case. You could happily be an instrumentalist of sorts about explanation, then, because of the extra thesis that there is supposed to be a *symmetry* between explanation and prediction. This was motivated by the consideration that all that really matters is that the premisses of the argument are known; from these premisses the conclusion should follow, regardless of whether the conclusion is that the explanandum will occur, or that its previous occurrence should have occurred (i.e. is explained.). So if theoretical terms are instrumentally useful for prediction, which the most hard nosed positivist granted, there should be no reason why they should not be instrumentally useful for explanation.

Almost no-one now believes the covering law model; at least not in anything closely resembling its original forms. A range of objections have been raised, from which I shall mention enough to give a feeling for why the model is history.

2.0.1 The Symmetry Thesis

The symmetry thesis has been attacked from both directions: by showing that there are predictions that are not explanatory, and explanations which are not predictive. The most famous of the former cases is Sylvian Bromberger's example of the flagpole and the shadow [Bromberger 1966], which, in various forms, has haunted the theory of explanation ever since. The problem is that, given various lawlike correlations of the angle of the

sun, the height of a flagpole and the length of its shadow, we can certainly *predict* what the height of the flagpole will be from the length of the shadow and the position of the sun, but surely this is no *explanation* of the flagpole's height.

Prominent among the latter cases is the evolutionary example of Stephen Toulmin [Toulmin 1961]—in which it is suggested that while the theory of evolution seems to give explanations, it does not seem to have made any successful predictions.¹ Michael Scriven [Scriven 1962] also has a range of counterexamples as well, most of which rely on the intuition that there seem to be perfectly good explanations to be had in cases where the information is not nearly complete enough to *predict* the explanandum. A particular event may be explanatory *ceteris paribus*, but actually spelling out the *ceteris paribus* clause may require more information than is at hand. It is the Hempelian response to these latter problems which leads to the problem of completeness of explanatory information.

2.0.2 Problems of Selection and Completeness

In the evolutionary case Hempel's response [Hempel 1965] was to introduce the notion of *retrodiction* to account for the evolutionary case. The thesis of symmetry between explanation and prediction turned into the thesis of symmetry between explanation and prediction or retrodiction. This was an *ad hoc* manoeuvre if ever there was one. It has no independent motivation, and it is hard to see what the difference between retrodiction and

¹This is not strictly true: the prediction that the proportion of moths in London that would lighten in colour after a clean-up campaign has, for example, been borne out. But this is a just one isolated evolutionary prediction, with none of the power and range of the instances of evolutionary *explanation*.

explanation is; and if there is none the thesis degenerates into the (of course perfectly true!) thesis of symmetry between explanation and (prediction or explanation).

Hempel's response to the other kind of cases was to argue that when there was insufficient information for proper prediction, then there was indeed no explanation, though there might be an 'explanation sketch'. Such explanations are enthymematic, even when the explanations they are enthymemes for are not known. This is more damaging than it seems; in the case of particular events, enough initial conditions are almost *never* known to allow the explanandum to be predicted as a *logical consequence* of the initial conditions and general laws. If we are to continue to regard some of our best scientific explanations as successful explanations, we had better not apply the strict DN requirement. An excellent discussion of this can be found in [Lewis 1986d].²

The other side of the completeness coin is the problem of selection; even if we know *some* explanations which satisfy the strict DN requirement, there is no guarantee that all explanations which do satisfy the requirement will be satisfactory. From the standpoint of a particular event, if one traces its causal history back a long way, you never reach a point where some exhaustive list of initial conditions and laws in the distant past would not, if it were available, provide a DN argument. Of course the further back one

²Lewis concludes from this problem that any constraint on how much (on his account) causal information you have to have to count as an explanation is inappropriate: explanation is just something you can have more or less of. As will become clear below I think this is too strong a conclusion to draw from an instance of too strong a constraint; there are, in particular contexts, often very particular kinds of causal information which will count as explanatory, and particular kinds which will not.

goes in an event's causal ancestry, the more scattered and irrelevant-seeming these laws and initial conditions will seem. And most of these DN arguments are not what we are looking for when we seek explanations; in response to a request for an explanation of the Gulf war, a comprehensive list of the initial conditions of all the particles moments after the Big Bang together with all the relevant laws would not, even if it were available, constitute the kind of answer that (given most contexts) would have been sought.

2.0.3 The IS Case

There are at least two significant problems which have beset the IS version of the covering law model. The first is to do with the nature of probabilistic inference. If $P(Q/R)$ is high, this is no guarantee that $P(Q/R\&S)$ is high. If we know that the probability of death from disease given that one is in a refugee camp is high, this may make it look like a good candidate explanation for a particular death from an infectious disease. But although the probability of death given that one is in a refugee camp is high, the probability of death given that one is in a refugee camp *and* one has been immunized against all known diseases is not high. So the putative explanation breaks down in the presence of extra information: and there is surely something wrong if more true and relevant information makes an explanation worse. This is a problem from which the DN account is free, of course, since if certain laws and initial conditions *entail* the explanandum, then, at least in Classical logic, no extra premisses could upset the entailment.

Hempel's solution to this (the best attempts are in [Hempel 1965] and [Hempel 1968]) is a *requirement of maximal specificity*. Crudely, this is simply to require that all pertinent information be taken into account when making

the probability judgements.³ Salmon has an argument in [Salmon 1984 ch. 3] to the effect that the most refined version of the requirement has the consequence that there is no real IS form of explanation, but that IS arguments in the Hempelian scheme are just more enthymemes for DN arguments.

More worrying still is the fact that on the IS model, events with a very low probability simply can't be explained, since the IS scheme requires that an argument be produced to the effect that the explanandum has high probability. If the explanandum in fact has low probability, then there can be no valid argument from true premisses which concludes that its probability is high!

It was as a remedy to this last problem that the statistical-relevance model once adopted by Salmon was first conceived. This model will be briefly outlined in the next section prior to a discussion of the problems it is still beset by, and what to do about them.

2.1 The Problems to be Solved

Some of the problems which faced the covering-law model, and others, continue to bedevil accounts of explanation which were designed to cope

³It is expressed as a requirement that we not know of any way to partition the class, membership of which gives the explanans its statistical significance,—in this case the class of people in refugee camps of a certain kind—into partitions one of which does not have the required property. So in this case the problem would be avoided by ruling out the explanation if we know of a way to partition the class into the class of those people in the camp who are immunized, and those people in the camp who are not. If the person were not immunized, of course, the explanation (*ceteris paribus*) could go through with the class of immunized people in camps as the statistically significant class.

with difficulties in the Hempelian account. I organize them into five groups for expository convenience.⁴

- 1) The Problem of Rejection: none of these accounts can explain why some apparently reasonable requests for explanation are rejected.
- 2) Asymmetries: nor can they explain why certain probabilistically identical propositions do not explain each other, or give an argument to show that they should.
- 3) Exclusions: apparently good explanations are excluded by the theory.
- 4) Inclusions: other apparently hopeless explanations are included by the theory.
- 5) Arbitration: different explanantia for some explanandum, with no way of arbitrating between them, or even of accounting for the differences or describing the relations, that hold between them.

⁴There is some overlap here; the asymmetries are a kind of inclusion, but for expository purposes it has been more convenient to keep them organized in this way.

3 Salmon's Probabilistic Account

Wesley Salmon [Salmon 1971] once maintained⁵ that explanation consists in citing factors which are statistically relevant to the explanandum. A factor F is statistically relevant to a phenomenon A exactly if $P(A) \neq P(A/F)$. Thus the only arbiter of statistical relevance is that the probability of the phenomenon given the factor is different from the probability of the phenomenon *simpliciter*.

This is supposed to overcome some problems with Hempel's insistence that an explanation at least give good grounds for believing that the explanandum occurs. The requirement is removed, and all the explanans has to do is to be statistically relevant to the explanandum. The well known example of paresis is the paradigm case which Salmon's model was supposed to be able to account for,⁶ but Hempel's models cannot.

Paresis is the unpleasant mental disorder that can result from latent untreated syphilis. (Robert Schumann is perhaps the most famous of many who have been afflicted by it). It is contracted only by people who have latent untreated syphilis, but not all—in fact very few—people who have latent untreated syphilis contract paresis. Believing that someone has latent untreated syphilis does not give good grounds for believing that someone has paresis. Yet surely the answer 'because they have latent untreated syphilis' is in some sense an explanation of why they have paresis. On Salmon's account, the explanatory relevance of having syphilis to the

⁵He no longer holds this view, but a discussion of it is useful to establish the direction of the argument. There is some discussion of his current views, best collected in [Salmon 1984] in ch. 3.

⁶See below for a trickier version which Salmon may not be able to account for.

contraction of paresis suffices to give it status as an explanation. One response to this claim would be to say that most so-called explanations are not explanations at all, but rather gestures towards a complete explanation—in this case one which provides sufficiently many added factors which increase the probability of paresis to the point where it satisfies Hempel's criterion. There are at least three objections to this response more or less explicit in the literature.

1) There would then be a significant loss of generality in our explanation, for surely what makes medical science possible is that it invokes the same explanation for various cases of paresis. If invoking latent untreated syphilis were merely an enthymeme for a fuller causal account, then it may presumably only be an enthymeme for a small class of such accounts, since the other factors which predispose latent untreated syphilitics to paresis may be very varied.

2) A reference to the full account that will make paresis sufficiently probable had better not be a reference to the complete causal story, since that is a story which goes back to the beginning of the universe, and has as its *dramatis personae* everything in the light-cone of the particular case of paresis that we are trying to explain. Of course by extracting part of the complete story, and relegating the rest of the plot to the status of background conditions, we can produce a DN explanation which will give the probability of paresis in this case as 1. But this will be a one-off explanation, and will succumb to the first objection. In any case, the theory does not give us any motivation for particular taxonomies of the story into background conditions and explanantia. Surely the explanation should not be that, other things being equal, Schumann got paresis because a shower of meteorites hastened the extinction of the dinosaurs? I propose an intuition: the right taxonomy must allow us to have syphilis right in there with the

other regularly occurring medical phenomena which are generally present in cases of paresis and which constitute the explanantia. I do not propose a justification for this yet, but the Hempelian model neither provides a motivation for preserving this intuition, nor explains why we should not have it. Since the regularities, other than syphilis, in cases of paresis are in fact few, the model does not even allow us to preserve it.

3) Perhaps at this stage it is objected that no matter what intuitions are sacrificed, it must be worth the sacrifice to prevent the absurdity of some event with a low probability acting as an explanation of it. If it can be shown that there are cases where this conclusion is irresistible (or at least much less resistible), then such explanations may as well be applied to cases such as paresis, with much benefit and no increase in the grotesqueness of the theory.

Consider the case of non-deterministic events. In particular, consider the case of the changes in energy-levels of electrons. From time to time, electrons move from high energy shells to low energy shells. The explanation of this is presumably that they do so in virtue of the structural features of the atom which are described by the true theory of atomic structure. Now the best candidate for this theory says that, in virtue of said structural features, there is a small but finite chance⁷ that such a change will occur in any given small interval of time. And this is all: the phenomenon is supposed to be irreducibly probabilistic. These chances are not disguises for our ignorance of the full, determinate, story. They are not subjective probabilities, rather they are objective ones. So if the explanation of these

⁷Taken as singular chances; I do not think the point suffers on other construals of the probabilities.

energy level changes is indeed in terms of the true structural theory, then that is a theory which gives as the explanans a low probability. *Prima facie* it is hard to see what is wrong with this, after all (runs the intuitive response) if it weren't for that low probability, the change wouldn't have occurred.

These cases—of paresis and of irreducibly probabilistic events—seem to be dealt with by Salmon's account. At least they are dealt with better than in Hempel's account. Since it is only statistical relevance that is at issue, both paresis and a quantum-theoretic account of electron energy-level change will qualify as respectable candidate explanantia. It will also at least point in the direction of a solution to the problem of alleged explanations which have as explanantia phenomena which are entirely irrelevant to the explanandum, but when added to the background theory guarantee the explanandum's likelihood just because the background theory does.

3.1 The Probabilistic Account and the Problems for Explanation

I shall now examine how successful Salmon's 1971 account is in dealing with these problems. The difficulties which still abound motivate the search for solutions through pragmatics.

3.1.1 The Problem of Rejection

Examples which illustrate how this problem can still arise for Salmon are in a way the reverse side of the coin which we tendered to buy Salmon's solution to the irreducibly probabilistic cases. We may agree that the reason that an electron changes energy level is that there is a small probability given by the theory that it will change energy level in a given period. So the explanation of the token change is in terms of this probability. But is this an explanation of why it changed energy level at that time? Consider a similar example of van Fraassen's. Suppose some U-238 atom decays during the

interval before its half-life. We ask why it decays during this interval. On Salmon's account the answer is that it is because it is a U-238 atom which has a certain probability of decaying during that interval—by definition $1/2$. Now while we may be happy saying that the atom decays because of some probabilistic features associated with its structure, surely to ask why it decays during that interval is precisely the kind of explanation request that is rejected by modern physics. Part of what it is to be an irreducibly probabilistic event is for the probabilistic theory to explain why it is that something happens, why it happens with a certain probability distribution, but not why a token event takes place at a certain time. Modern physics tells us that there *is* no explanation of that.

Yet on Salmon's account, the structure of the atom and its associated probability of decay must count as an explanation, because the probability of decay during the interval given the structural theory is different from the probability of decay during the interval simpliciter, a determinate probability being different from an indeterminate one. If we narrow down the interval we get decreasing probabilities, the limiting case being the one in which Salmon will have to claim that the explanation for decay having commenced at a certain time t is that the probability of decay commencing at t is infinitesimal. So if we are attracted by the notion that the occurrence of irreducibly probabilistic events at certain times is inexplicable, then we cannot use Salmon's account.

3.1.2 The Asymmetries

Salmon's theory does nothing to solve the asymmetries of explanation. In the notorious flagpole case, the shadow of the flagpole can still explain its height since the probability of the flagpole's being a certain height given the

length of its shadow and background features of topology, geography and time is different from the probability of its being that height simpliciter.

3.1.3 Illegitimate Inclusions and Exclusions

Bas van Fraassen [van Fraassen 1980] gives an example of what I have called illegitimate inclusions and exclusions. Let us consider the paresis case again. Suppose, contrary to fact, that paresis is caused by either latent untreated syphilis or by epilepsy, and that the probability of contracting paresis is the same in either case. Suppose that Mary comes from a family all of whom have either syphilis or epilepsy. If Mary has paresis, surely it is either because she has syphilis, or because she has epilepsy. However the probability of her contracting paresis given she is syphilitic, and the probability of her contracting paresis given she is epileptic are both the same as the probability of her contracting paresis given the disjunction of syphilis and epilepsy. Suppose she is in fact syphilitic but not epileptic: then surely this is the explanation; though on Salmon's account there is no way to distinguish this from the disjunctive case.

Van Fraassen thinks that this is a case of a good explanation being ruled out, and a patently bad explanation being accepted. He thinks that the disjunctive explanation simply isn't "the" explanation, and for the correct explanation to be said to be included it should be uniquely singled out. This seems to me to be too strong: the disjunctive explanation surely constitutes some kind of explanation. Such explanations abound, in medicine and elsewhere. For example malaria is carried by various kinds of mosquito. If someone contracts malaria in an area where two kinds of malaria carrying mosquito abound, then *prima facie* the fact that these kinds of mosquitos abound seems to explain the fact. Even if we know that the case of malaria was caused by a member of species A, this does not reduce the temptation

to explain it by reference to the disjunction in actual epidemiological practice—perhaps because the utility of such explanations is in devising programs for malaria elimination, and eliminating one species (or indeed the token mosquito that bit our victim) will hardly do. Admittedly such explanations are not strictly ætiological—they do not cite only actual causes—but van Fraassen himself wants to allow explanations that are not strictly ætiological. I am not persuaded by van Fraassen's claim that these kinds of cases constitute unproblematically illegitimate inclusions or exclusions. However, to the extent that that Salmon's account fails in any way to distinguish between the disjunctive case and van Fraassen's "real" explanation, the example remains a problem for Salmon. At very least it must be an instance of the problem of arbitration. There are two explanations here, and the theory does not give us any mechanism for choosing between them or even of accounting for the differences.

An example due to Nancy Cartwright [Cartwright 1979] shows that Salmon's theory suffers from illegitimate inclusions, and thus does not provide sufficient grounds for allowing that something is an explanation.

Suppose that some plants have been sprayed with Agent Orange, and ninety percent of them have been killed. Salmon's criterion will not only allow their having been sprayed to account for why some plant which is dead is dead, but will also allow the spraying to figure in an account of why some *living* plant has managed to survive: after all, the probability of a plant's living after it has been sprayed by a deadly defoliant is hardly the same as the probability of its surviving .

3.1.4 Arbitration

The problem of arbitration is well illustrated by an example which has had currency in the literature since Collingwood [Collingwood 1940]. Suppose

we are trying to explain a car crash. The competing explanations are negligent driving, a wet road surface, faulty steering and so forth. Call each of these candidates $E_1, \dots, E_n$. It is likely that, if there is anything remotely plausible about these candidates, that where C is the crash, $(P(C) \neq P(C/E_1)), \dots, (P(C) \neq P(C/E_n))$. The remote plausibility requirement guarantees this if it only means that the purported explanation makes some difference to the probability of the crash, however small. Indeed, by careful conceptual surgery of the various components of the crash's causal history it should be possible to ensure that pairs of explanations and background conditions are generated so that $P(C) \neq P(C/E_1)$ and $P(C/E_1) = P(C/E_2) = \dots = P(C/E_n)$. In this case the Salmon theory does not provide any arbitration between the rival explanations whatsoever.

Perhaps this failure to arbitrate is not alarming: the statistical relevance theorist may object that explanation just consists in listing probabilistically relevant factors. These "rival" explanations just are the explanations, and this talk of choosing between them is the rhetoric of insurance companies. The problem is, though, that Salmon's theory will not just have us listing plausible enough candidates like negligence and poor road surfaces. By the right sort of carving up of the complete history in which we choose selectively what to take as given, we can generate as many factors as we like which, together with the things we have chosen to take as fixed, will give us the desired probability of the crash. The breaking strain of some part of the gears (regardless of whether it is average for such gears), some factor in the decision-making process that led to the journey being undertaken in the first instance, or any one of many events in the past life of the town planner who designed the road system (long before she was engaged to design the model highway system in question), can all be gerrymandered into being explanations of the crash.

3.2 Conclusion

The purpose of this discussion of Salmon's earlier model has been to identify the problems which post-Hempel theories suffer from, and to use these as a motivation for an examination of the pragmatic solutions to some of these problems. What the various elements of a pragmatic account of explanation have in common is that the semantics of explananda, explanantia and explanation requests are supplemented by a pragmatics; they are relativized on the basis of interests and information to the particular contexts in which they occur.

4 Context and Contrast

Contextual contrast has been noticed as a puzzling, and not fully accounted for, aspect of some expressions in natural language since the linguistic work of Cook Watson (mentioned in Chomsky's *Aspects of the Theory of Syntax* [Chomsky 1965]). It is first developed in a philosophical way by Fred Dretske in [Dretske 1972]. The point of this paper was that the meaning of two apparently identical sentences can vary depending on a difference of contrastive focus in some expression embedded in both of them.

Take for example the sentence:

(1) Clyde gave me the tickets.

Dretske distinguishes between two possible ways of stressing this:

(2) Clyde gave me the tickets.

and

(3) Clyde gave me the tickets.

The effect of the stress is the same as it would be in natural language: the stress shows which part of the expression is to be focussed on, and what the sentence is being asserted in contrast to. So in (2), (Clyde gave me the tickets) the important feature is that it was me (rather than someone else) that was given the tickets. In (3) the important feature is that it is the tickets (rather than something else) that have been given to me. Dretske does not maintain that there is any strictly *semantic* difference between (2) and (3), on the ground that there are no conceivable states of affairs that would make one true and not the other, and they therefore do not vary in truth-value. This is not a general feature of such contrastive statements, however. Consider the sentences which are formed when (2) and (3) are embedded in an identical framework to form:

(4) Clyde gave me the tickets by mistake.

and

(5) Clyde gave me the tickets by mistake.

Suppose Clyde is instructed to give the tickets to Harry, but gives them to me by mistake—then surely (4) is true but not (5). On Dretske's account this would be a semantic difference in the expressions as a whole due to a difference of contrastive focus in the embedded (semantically identical) fragments. The exact boundary between pragmatics and semantics is not, I think, absolutely relevant to the importance of (prima facie) pragmatics to explanation. Perhaps one could construct a very rich, possibly non truth-functional, semantics which include disambiguation procedures along these lines. If Dretske is wrong about the non-semantic nature of much of the contextual and interest-related disambiguation of expressions, then it is no problem for the so-called pragmatic accounts of explanation if what is important about them is not that they are accounted for by a formal theory

of pragmatics in contradistinction to semantics, but rather that they are context and interest dependent.

4.1 Contrastive Focus

The relevance of Dretske's notion of contrastive focus can be seen by a consideration of another of his examples. Consider

(6) Clyde lent Alex \$300.

Depending on whether, for example, this is given its contrastive focus as

(7) Clyde lent Alex \$300.

or

(8) Clyde lent Alex \$300.

different explanations seem appropriate: perhaps, for (7), that Clyde thought giving Alex the money would embarrass him, or, for (8), that Alex needed the money more than George.

Garfinkel and van Fraassen take over this point, but develop it rather differently. Rather than using stress as a way of giving contrastive focus to one part of an expression and fixing the rest of it, they talk of a class of contemplated alternatives to the explanandum over which an explanation is supposed to in some way favour the explanandum. Garfinkel talks of a 'contrast space', van Fraassen of a 'contrast class'. I shall use the latter term for now, but will shortly introduce the notion of *choice* class, for reasons that will become apparent.

4.2 Contrastive Focus and Contrast Class

Is there a real difference between Dretske's contrastive focus and Garfinkel and van Fraassen's contrast class? The latter seems more concerned with which part of an expression is perceived as variable, and which fixed; while the latter is concerned with what the contemplated alternatives to an expression actually are. Thus Dretske is concerned with the fact that Clyde lent Alex \$300, emphasizing that it is 'Alex' rather than any other part of the sentence which is variable in this pragmatically elucidated version of the sentence. Garfinkel and van Fraassen would be more interested in the actual class of alternatives to Alex, for example

(9) Clyde lent Alex [rather than: George, Mary]
\$300.

which could well be explained differently from, say,

(10) Clyde lent Alex [rather than: Bill] \$300.

Clyde may well think that George and Mary are spendthrifts and need to be taught a lesson even though they need the money, and therefore favour Alex over them. On the other hand, he might loath Bill, and this would be the reason for his favouring Alex over Bill.

Contrast class can, I think, be seen as a generalization of contrastive focus if we allow each term in an expression to have a contrast class. We could well be interested in why

(11) Smith [rather than: Jones] is head of [rather
than: reader in, research assistant in] Anthro-
pology [rather than: Pre-history].

Two measures are useful at this point. The first is to allow a term to be included in its own contrast class, and the second is to allow a special, trivial case, classes which contain *only* the original term. This would allow the formation of constructs such as:

(12) Smith [rather than: Smith] is head of Anthropology [rather than Prehistory].

This creates a kind of null contrast class. When the contrast class contains only the same term as it is in contrast with, then there can be no facts which favour the original term over the class. This is quite different from a contrast class which contains *nothing*, i.e. the null class strictly, since presumably facts often, statistically or otherwise, prefer the original term to nothing. This allows us to define contrastive focus in terms of contrast class: the contrastive focus of an expression is on those terms in the expression that have contrast classes containing elements other than themselves. The full notion of contrast class would go on to list what they actually are.

Including a term in its own contrast class has the further advantage of allowing us to define the contrast class of a whole expression as the class of expressions formed by the cross-multiplication of the contrast classes of the terms in the expression. Thus an explanation of why

(13) Clyde [rather than: Clyde] lent [rather than: lent, gave] Bill [rather than: Bill, Jack, Harry] \$300 [rather than:\$300].

could be seen as favouring

(14) Clyde lent Bill \$30 (15) Clyde lent Jack \$300.

(16) Clyde lent Harry \$300.

(17) Clyde gave Bill \$300.

(18) Clyde gave Jack \$300.

(19) Clyde gave Harry \$300.

Thus defined, though, the contrast class would include (14). This is not a problem if we avoid the slightly misleading "rather than" expression and take the contrast class to be the full range of choices, and an explanation as being the function which favours one or some of them—in this case, (14).

In fact this more general contrastive apparatus is not so much a *contrast* class as a *choice* class, of which the contrast class in van Fraassen's sense is a proper subclass⁸. The explanans becomes a kind of choice function which selects, on various grounds I shall discuss later, one member of the class. When the one member selected happens to be true, and the others false, then the whole apparatus is a potential explanation of the selected member in contrast to the rest of the class. I shall continue to use the term contrast class fairly often, however, since it is closer to other formulations in the literature, while occasionally using my notion of choice class when it is especially important to bear in mind that a contrast class for a term may include the term itself.

This analysis is in the general spirit of van Fraassen's use of contrast class in [van Fraassen 1980]. Although when he first discusses the notion he talks as though they are classes of strict alternatives to a particular term in an expression, when he introduces his version of the notion in his formal

⁸Or perhaps van Fraassen's contrast class is already a choice class; as I mention in section 6 it seems at times as though he takes the explanandum to be included in the contrast class.

theory of Why-questions, he introduces the term 'topic' to mean the proposition to be explained, and stipulates that the contrast-class *includes* the topic (but that only the topic is true). Garfinkel's [Garfinkel 1981] idea of contrast space, however, seems to include only the counterfactual alternatives, although his notion is not formally worked out.

4.3 Contrast Class and the Problems for Explanation

Now we shall see how the notion of contrast class helps with the problems outlined in section 3.1. First, the problem of rejection.

4.3.1 The Problem of Rejection

We saw that Salmon's theory had difficulties in that it provided explanata which seemed to be irrelevant to the explananda in some cases, and where in any case the best theories of science would have us believe the explananda are inexplicable.

Consider the example of the decaying U-238 atom again. The problem was that while we might concede that the explanation of why the atom decayed might be that it had a certain probability of decaying, a problem arises when we ask why it decayed during the interval *I* before its half-life. There is a certain tension in our intuitions here. On the one hand, something tells us that the reason it decayed during that time had better be simply the reason it decayed; but also there is a sense in which, precisely because this is an irreducibly statistical phenomenon there should be no explanation of why it decayed *during that interval*.

If we think in terms of contrast-class, some of this ambiguity is removed. Consider the two propositions:

(1) The atom decayed during the interval I [rather than: never decaying].

and

(2) The atom decayed during the interval I [rather than: after its half-life].

A modified version of Salmon's theory, or any similar statistical or causal one which includes choice class would not give us the undesired explanation for (2), since the probability of decay during the interval I given the structural theory of the atom, is not different from the probability of decay after the half-life, given the structural theory. On the other hand, consider another case. Suppose that the half-life of some atom is N years. Suppose that the atom indeed decays after N years. It may seem plausible that some fact about the atom expressed by its half-life explains why it decayed in the interval 0 to N years, rather than the during some interval between fractionally before N years, and N years.

So there are good grounds for supposing that the notion of contrast or choice class is going to be an indispensable one in any account of explanation, to deal with cases where whether or not we think something is explicable depends on the contrast.

4.3.1.1 Contrast Class and Paradigm Shift

A related point is made by Garfinkel, who thinks that contrast class can be used to express differences in the kinds of questions permitted under different Kuhnian paradigms (or after epistemological breaks, to use Bachelard's term, which is favoured by Garfinkel).

His example is of the difference between mediaeval and Newtonian physics. Mediaeval physics is supposed to be addressing the question of why

(3) The thing is moving [rather than: not moving].

whereas Newtonian physics is taken to be concerned with why

(4) The thing has a given acceleration [rather than:
some other acceleration] at t .

To answer (3) Garfinkel maintains that some kind of force has to be postulated working at t , whereas to answer (4) one cites parts of the thing's causal history.

I am not convinced by this on two counts. First, I do not think it offers a complete diagnosis of the conceptual differences between the two physical theories. The difference between motion and acceleration is not captured by any difference of instantaneous states in possible contrast classes which is available without having the notion of acceleration available in the first place. Second, Garfinkel believes that this formulation shows that mediaeval and modern physics were asking different questions, and each provide correct answers to their own peculiar questions. This doesn't seem right. The story that requires a force to be acting on an object of constant velocity is a fiction, even when understood as an answer to the mediaeval question. If the contrast class is objects which are not moving, then if that term is understood extensionally (well within the Greek conceptual apparatus) then even without an understanding of the essentially relative nature of motion, the contrast can be supplied by those objects regarded as paradigmatically at rest: mountains, boulders and so on. And the right answer to the question about the difference in motion between some arrow whistling through the

air and the mountain (and indeed the air itself) is the difference in the causal histories of the arrows and the mountain, and not any story about a continuous force acting on the arrow. This failure of the contrastive apparatus to rescue mediaeval physics I take to be a desirable consequence, though. A contrastive account of explanation should not lead to pure relativism. Sometimes apparently different explanations may be equally good answers to different questions, but not always: sometimes there can be different answers to the same question, or different answers to different questions, when one or both of them are bad answers. The disambiguation does, however, while not showing that the mediaeval answer was the right answer to a different question, provide a useful *diagnosis* of why the answer was wrong. Drawing attention to the inclusion in the choice class of 'not moving' highlights the basic work that contrast played in their account, and explains why a physics that uses it is likely to go wrong.

4.3.2 Asymmetries

The notion of choice or contrast class does not solve the asymmetries of explanation. Statistical relevance plus contrast-class will still give 'because the shadow is *h* metres long at *t*' as an explanation of why the flagpole is *k* [rather than: *l*] metres tall.

There is some improvement in this sort of example, however, even if not strictly an improvement in terms of asymmetry. Some of the plenitude of explanations is excluded thanks to contrast class. If we make it clear that context determines that we do not mean, for example, that we are asking why

(5) The flagpole [rather than: the water tower] is *k* metres tall.

we at least exclude an answer in terms of its being too expensive to build a water tower of the right type to act as an appropriate landmark.

4.3.3 Exclusions and Inclusions

Contrast class does not help greatly with exclusions, but there are inclusions which it despatches nicely. Let us return to the paresis case. We saw how Salmon's theory allowed us to regard 'because he had latent untreated syphilis' as an explanation of why someone—say Phil—has paresis. I am sure that the remaining unease that we have at this result is not just the vestiges of deductive-nomological purity, even though this outcome is surely better than its counting as no kind of explanation at all. Equally, though, it is not a complete explanation in some sense, since, after all, disease is not (we hope!) an irreducibly statistical phenomenon, and not everyone gets paresis who has latent untreated syphilis. A notion of genuine completeness, however, would reduce every explanation of complex phenomena to a state of incompleteness unless it were so unwieldy that it would be an indigestible lump of causal history of no use until pruned (and therefore incomplete). So the mixed feelings which the paresis case provoke cannot be accounted for simply by saying that the explanation is incomplete. *Every* explanation accepted in real life is incomplete, and an attempt to complete the paresis explanation would in any case lose the explanatory generalization which is the benefit of the explanation in terms of syphilis, by including the idiosyncratic causal factors in each case. What contrast class can do, is to offer a satisfying account of why there is some residual unease about both accepting and rejecting 'because he has latent untreated syphilis' as an explanation of Phil's paresis.

Suppose that Jill, Bill, and Will all have latent untreated syphilis, and each of Jane, Shane and Wayne does not. Now,

(6) Why does Phil [rather than: Jill, Bill, Will] have paresis?

has no answer in terms just of having latent untreated syphilis on a statistical account plus contrast class, whereas

(7) Why does Clyde [rather than: Jane, Shane, Wayne] have paresis?

does have such an answer.

4.3.3.1 Choice Class and Explanatory Completeness

The size of choice class can be seen as related to the notion of explanatory completeness. As the choice class grows larger, so the explanations become more cumbersome and involved, so as to favour the explanandum over progressively more other members of the choice class. But unlike the pre-theoretical notion of completeness which is supposed to be the *aim* of explanation and hence desirable, it is obvious that the details which are relevant to preferring Phil as a candidate for paresis to the rest of a choice class which contains enormously large numbers of people or things becomes more and more ungeneralizable and useless from a medical point of view. Why Phil got paresis rather than an ant (ants don't get it !) is hardly an interesting question, and why Phil got paresis rather than a class of latent untreated syphilitics each of whom had a *slightly* different causal history both from Phil and from each other is going to produce an answer in terms of a wildly disjunctive property the disjuncts of which will be factors with minuscule probabilistic influence on the likelihood of paresis. Such an explanation will be forthcoming, but the reason that there is a tendency to say that Phil is just unlucky is that such an explanation is not much *use*, a useful explanation is one which selects from a *salient* choice class. Choice

classes are chosen for a purpose; the only thing which can fix the desired contrast is the particular pragmatics of answering the question.⁹

4.3.4 Arbitration

Further light is shed on the question of salience if we look at the problem of arbitration.

Let us consider again the problem of the car crash. Contrast or choice classes will enable us to clear up at least part of the problem. For example, most of the explanations that invoke factors causally or statistically relevant to the journey's being made *at all* are irrelevant when the pragmatics are made explicit in most cases. In these cases the post-theoretical reframing of the question which best matches the pre-theoretical question and its pragmatics might be something like:

(1) Why did the car crash [rather than: continue safely]?

and not

(2) Why did the car [rather than the bicycle in the garage] crash?

or

(3) Why did the car crash [rather than: skid]?

⁹A philosopher looking for ways of outlawing as not well formed the big and possibly meaningless metaphysical questions—like why is there something rather than nothing—would well be advised to investigate the peculiar properties of genuinely null or very large choice classes.

The contrast class carves up reality; only part of the causal web is relevant to preferring this event or that event over specified alternative(s). Explanations are in some way about why the alternatives do not happen. So the contrast class works by eliminating much of the material that would otherwise bog down the explanation in irrelevance.

This elimination is especially effective in cases where we are tempted to cite the counterfactuality of counterfactuals as explanations. If we were to say that an explanation of why the car got from home to work was that no meteorite struck it, we would be laughed at, and, in many contexts, rightly so. But a question like 'Why did the convoy get to the supply base?', answered by 'There were no enemy ships in the area' seems less absurd. We can account for this by considering the specified alternatives built into the question. Consider this formulation:

(4) Why did the ship get to the supply base [rather than be blown up by the enemy]?

and the answer does not seem so empty. It provides a distinct answer which supervenes on quite different material states of affairs than, say, the enemy's having run out of ordinance, and no answer at all to

(5) Why did the enemy get to the supply base [rather than: strike an iceberg]?

4.3.5 Contrast, Choice and Explanatory Irrelevance

Contrast class can also be used to account for what Alan Garfinkel calls the 'irrelevance geometry' of explanations. Let us consider, yet again, the car crash example. Suppose in trying to evade a claim that I was driving negligently, I claimed that the explanation of the crash was that I was

wearing a certain pair of shoes. After all, I claim, *that very* car crash would not have happened if I had not been wearing those shoes.

This is not, of course, very convincing since presumably not *everything* in an event's constitution is essential to it. When we try to explain events, often we are concerned with a degree of generalization which is greater than the particular details of the explanandum *as it occurred*. Rather, we are concerned with an explanation that would have been good even in different, but relevantly similar circumstances—such as where different shoes were worn.¹⁰ To overcome this, Garfinkel introduces an equivalence relation 'differs inessentially from' and casts the explanandum as the class of propositions that bear this relation to the explanandum. So for the car crash example we end up with something like

(1) Why did the car crash [or any of the equivalence class] rather than [contrast class]?

This kind of question is particularly interesting in cases where there are structural preconditions at work. When we ask 'Why is Sally unemployed?', the equivalence class could be very large indeed. In this case it could include any individuals relevantly similar to her, in which much of her particular causal history will be irrelevant. This is the version of the question in which answers like 'Because of the stock market crash' can be appropriate. The equivalence relation is so wide that most unemployed people can fit into it, and so facts which structurally account for why that class of people is unemployed will count as explanations. If we take a narrow equivalence

¹⁰This introduces a topic that is dealt with more fully in Ch. 4. Here I am mainly concerned to note that much of Garfinkel's intuitive notion of inessential difference can be largely captured by contrastive differences.

class then more and more specific details of the individual's causal history become important, and we are looking for particular facts which make it the case that *that* person is unemployed. We are, in fact, asking why it is that that person *rather than someone else* is unemployed.

The language of this last sentence may sound familiar: I think this is because the work of Garfinkel's equivalence classes is already being done for us by contrast or choice class.

Our faithful car crash can be put to work here. Let us select contrast classes so as to generate the following two versions of the problem:

(2) Why did the car crash [rather than: behave in all respects similarly except that it did not crash at that time]?

and

(3) Why did the car crash [rather than: be involved in a different crash in all respects similar except that different shoes were being worn by driver]?

I think choice or contrast class analysis along the lines of (2) may save us from the need for an explicit class of relevantly similar alternatives in addition, as demanded by Garfinkel. The point of his demand is that we are typically not trying to explain why the event (in this case the crash) happened in all its inessential detail; rather a good explanation ought to be one which will do for all inessentially differing events, where what counts as inessentially different is something for pragmatic determination. The danger is supposed to be that each inessential detail may turn out to be explanatory unless it is explicitly postulated that it is inessential and is therefore not part

of the identity conditions for the explanandum. The schema would be something like this:

(4) Why did the car crash [or any relevantly similar event] [rather than: not crash [or any relevantly similar failure to crash]]?

I hope my formulation of the proposal as (4) helps make it clear why ordinary contrast class alone should suffice: what gives the Garfinkel problem its bite is the presence in the contrast class of unnecessary relevantly similar failures to crash. If the contrast class contains, say, failures to crash in which the driver has different shoes or (especially) crashes in which the driver has different shoes, then certainly the wearing of the shoes that the driver did indeed wear will turn out to have at least statistical relevance which will favour the explanandum over the contrast class. But there is no reason to include such events in the choice or contrast classes. The content of such classes is *stipulated* according to interest; and the most common interests are in line with (2) above. In this case, the wearing of the particular shoes does not increase the probability (or otherwise select) the explanandum over the contrast class *just because the contrast class is the same in respect of the shoes*. The problem only arises when events different in this respect are included in the contrast class, in which case a formulation like (4) and a rule which requires the explanans to favour at least one member of the equivalence class to each of the contrast class might prove necessary.

Of course the story without the equivalence classes *allows* one to make whatever you want essential to an event—there is no bar against asking questions of type (3); in which case a rather narrower part of the causal history of the event is going to be able to select the explanandum from its alternatives along the axis of variation of footwear. But there is nothing

wrong with its being allowed, of course. For some purposes this might be the salient question, and in any case the equivalence class story allows it as well.

The motivation for these equivalence classes, or perhaps for the construction of contrast classes which require them, is of course not only to prevent the explosion of details which are irrelevant given certain pragmatics. It is also to allow the construal of explanation requests in such a way as to admit variations of generality, as in the unemployment example above. We are supposed to be able to explain things where quite different classes of things will count as relevantly similar. So, to return to our unemployed friend Sally, we might be interested in why she is unemployed when *anyone* in the same society counts as relevantly similar, we might be interested in why she is unemployed where the class of unemployed people count as relevantly similar, or we might be interested in why she is unemployed where few if any people count as relevantly similar.

These distinctions, though, are capturable with just contrast or choice classes. Consider

(5) Why is Sally [rather than: no-one] unemployed?

and contrast it with

(6) Why is Sally [or anyone else] unemployed?

I take it that these things boil down to the same question. Certainly it will be the same explanantia that do the work here. It will be answers in terms of the economic conditions of the society which govern the existence of unemployment which are needed for both of these questions. In the same spirit:

(7) Why is Sally [rather than: (some of?) those people with jobs] unemployed?

and

(8) Why is Sally [or any of her co-sufferers in unemployment] unemployed?

In these cases it is the complement class of the similarity class which appears in the choice class (or the complement class of the conjunction of the particular individual and the similarity class which appears in the contrast class). It is not always possible to provide *unambiguously* a contrast class analysis of a question couched in terms of a similarity relation, but this shows only that contrast class is a more general notion. Consider

(9) Why is Sally [or Jenny] unemployed?

In the absence of further pragmatic information there is nothing to choose between, say,

(10) Why is Sally [rather than: all other unemployed people other than Jenny] unemployed?

and

(11) Why is Sally [rather than: all other unemployed people under 18 except Jenny] unemployed?

but this is part of the contrastive ambiguity of explanation requests in general, and serves only to reinforce the need for a contrastive analysis. In some cases the contrastive analysis of sentences couched in equivalence class terms becomes either trivial or Byzantine: but this is when there is genuinely

little in common between the things in the equivalence class, in which case the messiness of the analysis reflects the artificial (or, more neutrally, complex) nature of the equivalence claim. In any case, in such cases the claim that such an equivalence is tacitly understood becomes weak, and if it is to be articulated it should be done simply as a request for explanation of a disjunction. And there is surely no general requirement that every request for explanation of a disjunction be expressible as a request for explanation of one of the disjuncts and a contrast class in order to claim that equivalence class is not required for complete expressibility of explanation requests. So while I sympathize with part of the motivation of Garfinkel's equivalence class and irrelevance geometry, I think its work can be done with the tools already at hand.

5 Explanatory Relevance and the Problem of Arbitration

So far we have managed to prune back the bewildering variety of factors which might count as explanations by appealing to context. Nevertheless we still have part of the force of the problem of arbitration to deal with. Suppose we have cashed out the car crash question as in (4.8) above, and ask:

- (1) Why did the car crash [rather than: continue safely] at t ?

We are still left with a multiplicity of answers such as

- (2) There was tall shrubbery at the road at that bend.

or

- (3) The driver was negligent.

or

- (4) The brakes were constructed in exactly such
and such a way.

which could be equally probabilistically significant (or count as necessary but not sufficient parts of sufficient but not necessary conditions *a la* Mackie [Mackie 1965] [Mackie 1974] or whatever your preferred theory of causation is). Even if we accept as unproblematic a theory of real causation, and see events located in causal networks as in Salmon's later theory [Salmon 1984] we will be left with too many to count usefully as an explanation. Indeed we are still left with the plenitude of them that we had before: although we have chopped limbs off the causal tree, we still have some infinitely long (and probably dense at that) ones to deal with. Some more pragmatic surgery is called for if we are to capture explanation.

The first temptation which one might have is, again, to accept all of (2), (3) and (4), and say that once we have analyzed the problem in terms of contrast class, explanation is just listing all the factors relevant to favouring the explanandum over the rest of the choice class. This temptation fades a little, though, upon reflection. Explanantia (2), (3) and (4) were, of course, chosen to have the flavour of plausibility as they are the kind of explanation that is typically being looked for given certain interests.

A town planner might be looking for some feature of roads which is easily changed and that, if changed, would reduce the likelihood of accidents. A lawyer might be looking for some behaviour on the part of the driver statistically or causally relevant to the crash, and which can be regarded as being culpable. An engineer might be looking for some feature of the car which varied from the maker's specification and which was

involved: an engineer from a rival firm may look for a feature present only in cars that meet the first manufacturer's specification!

There are, though, countless factors which are not interesting in these ways and which will still have the required causal or statistical role. The entire mechanism of the brakes, say, might be able to be implicated by the same argument, which implicates that aspect of them which is non-standard: had they been other, the crash mightn't have happened. Every fact about the driver and the causal history of his or her psychology will give us causal clay from which to mould chunks to get the desired statistical property called for by Salmon's theory modified by contrast class. If we accept all such comers, then we will still have a uselessly vast list.

5.0.1 An Example

A useful example here is looking for a bug in a computer program. We have seen some strange behaviour in the program. A peculiar symbol flashes up in front of us in the middle of the screen and won't go away.

Why? A short answer is because it is a functional output of the structure of the program. This doesn't help much, the computer centre manager wants to get to the bottom of the problem (and preferably *fix* it). So a laborious search is made through thousands of lines of source code, the flow of control flickering confusingly through the program in tortured tangles (the program is written in FORTRAN by a programmer with a lofty contempt for the pretensions of structured programming).

What is being searched for? Answer: the explanation of the glitch. For the deductive-nomological model, or Salmon's model, we are looking for some factor which, together with the laws in the field and the background

theory will be related in the proper deductive, statistical or irreducibly causal way.

But the theory simply is the principles of the programming language and an overall understanding of the program and the related hardware. Now most parts of the program, together with the theory, will produce the desired result. If a line of code (let that be the size of unit we are dealing with) is at all important to the functioning of the program, then there is a good chance that the program just won't run in its absence, so that the possibility of the glitch given the line and the theory is certainly going to differ from the probability of the glitch given the background *simpliciter*.

Nevertheless there will often be agreement that a certain module or line of code contains the code. This will in part be because of agreement that we are trying to explain why the glitch happens rather than the program functioning normally without the glitch. But only in part; there are many artificial carvings up of the program which could locate the problem in some (wildly distributed) module which obeys that constraint. One account of the agreement which would be in line with White Beck's views is that the line is that thing we have control over, such that if we change it we could iron out the glitch. This won't do: there are vastly many changes that could fix the glitch, though most of them might be terrible programming practice. So still more contextual factors will have to be found.

Perhaps the part of the program which can be invoked to explain the glitch is that part which can be modified with the least trouble. An explanation would show why this section rather than other candidates is most easily (or elegantly) modifiable.

Perhaps the problem proves so intractable, and the program so messy, that the only real solution is to scrap the program and rewrite it from

scratch. In this case a system consultant (called in by the despairing manager) would, in giving an explanation of the glitch, not invoke lines of code or modules, but rather say “the whole thing is useless, and it produces this wretched glitch”: exactly, and it is the whole thing that needs to be replaced.

5.2 Interest and Purpose Relativity

This example is not meant to be generalizable into a theory of explanation in terms of repair; rather the point is that there are purposes and preconceptions dependent on context which makes certain factors cited as explanations in a context. Van Fraassen points out that most post-Hempel and Hempel writing on explanation acknowledges that explanation points out ‘salient’ features of a causal network. It is just that they have taken salience to be unproblematic, and have nevertheless come up with a bewildering variety of criteria for salience. Gasking suggests control over the factor, Nagel [Nagel 1961] suggests lack of control, Braithwaite [Braithwaite 1953] ignorance of the factors and Bohm [Bohm 1957] variability¹¹. And each of these accounts is meant to be *constitutive* of salience, not just an example of it. The bewilderment fades a little if we realize that, depending on what one is trying to do with an explanation, what the purpose of it is in its context, different ones of these might come into play. Variability helps if you are trying to repair something (and we want to know why it won’t work) as does control. Lack of control helps if we are trying to explain why it can’t be fixed, ignorance if we are trying to reduce our puzzlement, and so on.

¹¹For the last three references I am indebted to the survey in [van Fraassen 1980], which is in turn due to a survey in [Zwart 1967].

5.3 Relevance Relations

This leads to the extra ingredient required to disambiguate the notion of explanation enough to make it look like the folk notion: some stipulation as to the *kind of factor* which is being looked for in an explanation request. Van Fraassen calls the contextual determination of which features are interesting or important the *relevance relation*. He does not deny that science allows us to describe the processes which lead up to and go past events in a causal network—to that extent he (and I) agree with the later Salmon¹²—or that it can describe the standing conditions that must and must not obtain. On the other hand he does not seem to think that citing information in some way about causes is the only kind of explanatory information. In any case the doctrine is that all there really is to say about explanation is an account of how to choose from a mere description, the contextually relevant extra features. He has a motto: “no factor is explanatorily relevant unless it is scientifically relevant”, and among scientifically relevant ones, context determines explanatorily relevant ones. Just how the motto is justified within the framework of his own account, though, is puzzling—for the stipulation that nothing but a scientifically relevant feature could be explanatorily relevant does not feature in the formal apparatus; so presumably it, too, must be a merely pragmatic constraint. Yet this does not seem consonant with the boldness of the motto; nor with the importance of the motto. Nor, for that matter, is any account of what scientific relevance might be provided. That too seems to be a *purely* pragmatic notion on his account.

¹²[Salmon 1984]: this is a causal view of explanation. Causal aspects of explanation are discussed in the next chapter.

There is no precise account of the various different ways contextual determination of explanation proceeds in Garfinkel, aside from the stipulation that there needs to be a contrast space and an irrelevance geometry (through relevant similarity relations) as described above. Van Fraassen, however, does give a precise account. He makes use of Belnap's [Belnap and Steel 1976] formal analysis of why-questions in a formal account of explanation-requests and explanations.

6 Van Fraassen's Logic of Why-Questions and Some Variations

Van Fraassen characterizes the why-questions that explanation requests are supposed to be as composed of three elements, in addition to which there is a background theory to which any answer is relativized.

(1) The *topic* (i.e., the explanandum) P_k .

(2) The *contrast class* $X = [P_1, \dots, P_k]$.

(3) The *relevance relation*, R .

As well as these factors, the *background theory*, K , is required to formulate answers. The purpose of the background theory is to provide the source of adjudication as to whether some proposition is indeed scientifically relevant in the required sense. Thus the background theory might be Newtonian physics, relativistic physics, evolutionary epistemology, sociobiology, long-wave theories of Mandelian political economy, macroeconomic theory, or whatever else takes the enquirer's fancy. As this provides the criteria of scientific relevance, it is clear that this theory of explanation is not meant to be a theory of *correct* explanation, since the notion of scientific relevance which is a necessary condition on explanation is not a condition of scientific relevance according to true scientific theories. I think there is some confusion

here, however. The van Fraassen motto seems to be designed to give a basis in the way the world works (for him, construed anti-realistically) in explanation. But if there is no constraint on truth to provide this basis, then it is unclear what work the notion of scientific relevance is doing. For the motivation for an *unconstrained* account would presumably (none of this is explicit) be to give a story about what it is to be an explanation in terms of some given theory, rather than a *correct* explanation. But if this is the case then astrological and theological explanation ought, presumably, be captured by a theory of explanation since they are, historically, almost paradigm cases of (bad) explanation. If they are to be excluded, it seems odd to do so on the basis that they are not sciences, rather than that they are false theories. If they are to be included on the grounds that they *are* sciences, then the notion of science has been extended to the point where it is hard to see what it would *exclude*.

These three factors, together with the background theory, are all that van Fraassen uses to characterize the notion of a why-question. The why-question in general is characterized as an ordered triple $Q = \langle P_k, X, R \rangle$ with a background theory K . A proposition is called relevant to Q iff it bears the relation R to $\langle P_k, X \rangle$.

He then defines a *direct answer* to Q , given the background theory K , in the following way:

B is a *direct answer* (given K) to question $Q = \langle P_k, X, R \rangle$ exactly if there is some proposition A such that A bears relation R to $\langle P_k, X \rangle$ and B is the proposition which is true exactly if $(P_k$; and for all $i \neq k$, not P_i ; and $A)$ is true [van Fraassen 1980 p144].

In other words, B is a direct answer to Q if there is some candidate explanans A which is relevant to $\langle P_k, X \rangle$ and this answer B is true if and only if

(1) The explanandum ('topic' in van Fraassen's terminology) P_k is true.

and

(2) No (other) member of the contrast class is true.

and

(3) The candidate explanans, A , is true.

This is supposed to capture the idea that a direct answer to a question (a stronger notion than would normally be implicit in this phrase but weaker than the idea of a fully telling or successful explanation) is some true statement, relevant to some question seeking explanation of why some true statement is true rather than some alternatives which are not. Thus no question seeking 'explanation' of something that is not the case could have a direct answer.

This is captured formally by the notion of a *central presupposition*. The central presupposition of a why question, Q , is (1) and (2) above. The question Q can be said to arise iff the background theory K entails (1) and (2) above, and does not entail that there is *no* proposition, A , which bears the relevance relation R to $\langle P_k, X \rangle$. So a question can arise even though there is the (epistemic) *possibility* that there is no direct answer to Q in virtue of there being no A properly related.

6.1 The Topic and the Contrast Class

One feature that should be observed about van Fraassen's formulation is that the topic appears as a member of its contrast class. This is not motivated in the text of van Fraassen, and I suppose it is this way for the sake of formal elegance of a sort. He nevertheless continues to talk about *contrast* class. As I have argued above, however, I think there are good independent

motivations to do with contrastive focus for including the explanandum in the class. This is why I prefer the term 'choice class'.

6.2 Choice Class of Explanans

Much of the work in this scheme of van Fraassen's is being done by the relevance relation, but the exact nature of this relation is not spelled out, other than in terms of the general claim that it further trims the causal or other tree to provide an explication of what it is to ask for an explanation in a certain respect. Thus, asking for some kind of local political decision which explains some event rather than asking for some structural or international political effect which would also in some way favour the explanandum over the rest of the choice class would constitute a difference of relevance relation.

The kind of *contrast* which we have been so far considering has been contrast of explanandum, in which the explanandum is contrasted with various counterfactual alternatives. A suggestion by Philip Pettit that contrast of *explanans* might also be involved in an account of explanation leads to what I think may be an alternative to the relevance relation of van Fraassen.

Let us consider the sort of work that the relevance relation has to do. It does not specify the (enormously large) class of candidate facts which do in some way select (by being linked causally or probabilifying or whatever) the explanandum from the choice class; that work is done by the background theory. Rather it selects the kind of factor which we are interested in. One way of seeing this is as a kind of class of admissible answers from which investigation will choose at least one which is true. Such a list would have the features of a contrast class or a choice class: either it would be a certain answer rather than specified others that would be the properly understood

answer to a question, or else there could be a kind of choice class of answers included in the explanation request, from which we must select.

This mirrors the actual procedures in our examples so far. In the case of the car crash, the mechanic who is interested in the deviant mechanical feature of the car discovers that it is one particular such feature of the car rather than the class of other such features which is actual and which therefore gets to be the explanans. And the brakes failing rather than any other mechanical factor is a very different sort of contrast from the brakes failing rather than, say, negligence since in the first claim negligence isn't in the class and so cannot be selected.

If we set up the problem and the probabilities in the way they were when this example was first discussed above (i.e. where C is the crash and the E s are competing explanata $P(C) \neq P(C/E_1)$ and $P(C/E_1) = P(C/E_2) = \dots = P(C/E_n)$) then the first claim would be true and the second false. But it will be false in an interesting way: it is a hybrid choice-class of explanans, and I think that these underlie much explanatory perplexity. There is no scientific factor which, in this example, favours the brake story over the negligence story. But for most concrete purposes when explanations are actually used implicit choice classes of explanans are set up so as to contain factors of a single kind not many of which will be actual, since in such conditions enquiry is most likely to be successful: not many investigations are set up to consider vast numbers of different types of factors, or indeed have interest in so many types. But when reflection runs riot, or when compensation cases depend on it, large choice classes of explanans get created in which a very large number of types of factor occur, and in which many of the members may be actual. The car case is an instance of the latter, the former includes questions like 'Is the tower there because of the labour of its builders or because of the need perceived by the King?'

A consequence of my view (and if I am right about the relationship between relevance relation and contrast of explanans, then also of van Fraassen's) is that before a why-question can be properly framed, we need to have some idea of the class of things which would count as answers were they true. If we require not only that they be specified by type but as tokens at some fairly high level of generality, then we would require at least a partially worked out list of hypotheses before we have fully formulated the question. This is not as strange as it sounds; much investigation does proceed in just this way, by working through hypotheses and eliminating them leaving the true ones in the class. If it is objected that it is at least *possible* to ask why something happened without the remotest hint of a list of possible answers or answer-types, then I am inclined to say that, at least initially, what is being requested is a fairly unconstrained description of the causal (or other) history of the event. When this begins to emerge, pragmatics will begin to determine some list at least of the *interesting* factors and explanation proper can begin. I offer this as a diagnosis of why, despite great efforts to constrain them, Royal Commissions set up to explain things get so bogged down with competing claims. It is because in the absence of the powerful pragmatics which constrains scientific research, more causal descriptive work is being done than narrowly explanatory work.

6.3 The View Reformulated

In the light of these observations about choice classes, and to facilitate the further development in later chapters, I propose to reformulate the notion of a direct answer to a why question. A why-question (or explanation request—I use this term occasionally because of its applicability outside formal contexts) comprises a choice class of explanans $A = \{a_1, \dots, a_n\}$, a choice class of explanandum $B = \{b_1, \dots, b_k, \dots, b_n\}$, and an explanandum b_k . Direct answers to why-questions say that some proposition, a_j , explains the

explanandum. For the proposition to explain the explanandum, it must *select* the explanandum from its choice class. This will be more fully spelt out in the next chapter, but the idea is that there must be some member of the choice class of explanans which selects the true member of the choice class of explanandum from the competing members in virtue of various explanatory features. Being a direct answer is a necessary condition for such selection, which is in turn a necessary condition for being a bona-fide explanation.¹³

Some proposition P is a direct answer to the why question $\langle A, B, b_k \rangle$

iff

P is true *only if*

- (a) b_k is true
- (b) no other member of B is true
- (c) a_j is in A
- (d) a_j is true.

So, on this analysis, it is a necessary *and* sufficient condition for being a direct answer to a question, that a certain *necessary* condition for selection and hence being a bona-fide explanation is met. Thus the pragmatic notion of being a direct answer falls out as a necessary component of being an explanation. In the next chapter I will begin to flesh out further constraints to arrive at something approaching sufficiency conditions.

¹³Every account of explanation has a point where the account of explanation stops, and an account of good explanation starts; the borders are often murky. The full account of selection given in the chapters to follow might be a good place to draw the line, in which case selection would constitute explanation (so I suppose it would constitute bona-fide explanation), but not necessarily *good* explanation.

6.4 Choice Class of Explanandum and Explanans

There is an important difference between choice class of explanans and choice class of explanandum. While all the other members of the explanandum choice class other than the explanandum itself must be false, this need not, at least *prima facie*, be the case with the contrast class of explanans. There may, after all, be more than one factor of the right type which is true. If choice class of candidate explanantia contains two members which are true, and—to make the case completely unproblematic—they both select the explanandum from the rest of the contrast class, then the choice class will have yielded up two equally good explanations. It is easy to see how this happens. Suppose we were searching for physical factors about the construction of the Challenger which explain the disaster, and included in our list of candidate explanantia are strong local density variations in the solid fuel. Suppose there are in fact two of these; each sufficient for the disaster. This straightforward case of overdetermination had better not be ruled out by our theory.

I have two suggestions.

Firstly, we could formally define a *unique* direct answer as the primary notion in a schema not unlike van Fraassen's. If we call the choice class of explanans *A*, we could say that:

Given *K*, *P* is a unique direct answer to *Q* iff

(2) *P* is true only if

(a) b_k is true

(b) no other member of *B* is true

(c) a_j is in *A*

(d) a_j is true

(e) no other member of A is true.

Non-unique answers can be simply dealt with by leaving out the stipulation (e) above.

Alternatively, the model of unique answers could be constructed so as to make the unique answer the conjunction of the true elements of Y . If we allow the inclusion relation on choice classes to be such that for all a_c, a_d in A , if a_c is in A and a_d is in A then $a_{cd} = \{a_c, a_d\}$ is in A , then we can get the desired result by changing the last stipulation to read:

(e') no other member of A not also a member of a_j
is true.

7 Answers and Telling Answers

The apparatus so far explicated, whether in van Fraassen's canonical form or in my version designed for further constraints, gives at most an account of direct answers to why questions. It does not provide a way of *evaluating* those answers. In my own formulation there is mention of *selection* of the explanandum (from the choice class of explanandum) by the explanans (rather than the rest of the choice class of explanans). This notion of selection is crucial; without it, one merely has to stipulate that there is some true proposition in the choice class of explanans and one instantly has a direct answer to a question.

Van Fraassen does not build any such notion into his theory. Rather he has a separate account of evaluation of answers to why questions as telling or otherwise. Much of it is unclear; basically it consists of probabilification of the explanandum (P_k) by A , the *core* of the answer to the why-question. This leaves the notion of being an answer to a question a hugely

unconstrained one. In the next chapter I will deal with constraints required for my notion of selection, and the most appropriate places to place them.

Chapter Three

The Causal Constraint

Table of Contents

1	The Asymmetries Revisited and the Causal Constraint	64
1.1	An Asymmetry Revisited	64
2	The Tower and the Shadow	66
2.1	Kitcher and Salmon's Objection.....	68
2.1.1	A Version of the Kitcher	70
2.2	A Constraint Rejected.....	71
2.2.1	The Second Kitcher.....	72
3	The Causal Constraint.....	76
3.1	Constraints on the Selection Relation	77
3.2	Counterfactuals and Causal Dependency.....	77
3.3	Causal Relevance and Chains of Counterfactual Dependency.....	78
4	The First Constraint.....	78
4.1	A Counterfactual Approach to the Asymmetry.....	79
4.2	The Reichenbach-Salmon Approach.....	81
5	The Second Constraint.....	85
6	The Third Constraint.....	87
7	Some Problems of Generality.....	88
7.1	Common Causes.....	89
8	Extension to the Probabilistic Case	92
8.1	Genuine Single Case Probabilistic Explanation.....	94
9	The Limits of Causal Relevance.....	96
9.1	A Cosmological Explanation Ruled Out	97
10	What kind of explanation is a causally constrained pragmatic account?.....	100

1 The Asymmetries Revisited and the Causal Constraint

In this section I argue that, for all the benefits which the Pragmatic theory of explanation has so far, there remain problems which cannot be overcome except by appeal to a primitive notion of causation, and imposing a causal constraint on what counts as an acceptable member of a contrast class of explanans, and on the notion of selection of the explanandum. This raises the question of whether this means adopting either the spirit or the letter of Wesley Salmon's later *causal* account of explanation in his *Scientific Explanation and the Causal Structure of the World* [Salmon 1985]. I argue that, properly understood, a causal constraint still retains much of what is important in a Pragmatic account of causation, even if not in quite so liberal a form as van Fraassen [van Fraassen 1980] [van Fraassen 1985] would prefer. The last chapter has shown that pragmatic theory is not just a pragmatics of explanation (as claimed by [Kitcher and Salmon 1987]) but rather is fully part of the theory of explanation. It is concluded that what is distinctive about a causal-pragmatic account is that it describes explanation as the pragmatic selection of information which is true *in virtue of causal connexions*. For explanation, the causal regularities themselves do not have to be disclosed. Thus, we can consistently deny that there is anything pragmatic about *causation*, while insisting that what makes causal information explanatory is entirely pragmatic.

1.1 An Asymmetry Revisited

In the section above concerning choice class of explanandum, I expressed the view that this piece of apparatus does not help with some of the most notorious asymmetries of explanation. Nor is it clear that van Fraassen's relevance relation or my choice class of explanans is of much help.

Consider again the problem. There is a tower, of height h , which at time t has a shadow of length l . We suppose that the height of the tower, together with the position of the sun and our background body of knowledge about the properties of light and so on, should allow us to explain the length of the shadow; but the height of the tower ought not be explained by the length of the shadow together with the background knowledge.

Let us stipulate that there is a choice class of explanandum which includes various heights of the tower: h , $2h$ and $3h$ metres. Unless some constraint is placed on what counts as information which is *relevant* to the explanandum, then on van Fraassen's account the length of the shadow, l , then explains why the tower is h rather than $2h$ or $3h$ metres, since h can be deduced from l plus a background theory. The length l even *favours* the explanandum probabilistically over the other members of the choice class in the fashion that van Fraassen uses as a criterion for a *telling* answer to a why-question in [van Fraassen 1980 pp 146–151].

It might be objected that this just shows that there is *some* pragmatics under which the undesired direction of explanation goes through, not that the undesired explanation *must* go through. In other words, we would be advised not to ask this particular question if we don't like the answer that we would get if we did. But this misses the point; such a response is justified just when there is some independent ground for thinking the question is inappropriate. In this case, however, there is nothing odd about the choice class of explanandum. A range of different heights of the tower is *exactly* what we would expect to be in such a choice class.

The choice class of explanans does seem to provide a little more help in avoiding the conclusion that the length of the shadow explains the height of

the tower, but by no means enough. By restricting the kind of factor which we are looking for to reasonable ones, the problem might be skirted. If the choice class of explanans includes, for example, only putative facts about the intentions of the designers, then an answer in terms of the length of the shadow would be excluded. This would be equivalent to having, on van Fraassen's scheme, a relevance relation which specifies relevance in terms of intentional states of the designers of the tower in question. This leaves open, however, what grounds there could be for supposing that there is anything *wrong* with stipulating a choice class of explanans which includes lengths of the shadow. This is slightly less problematic than in the case of the choice class of explanandum; in that case there was obviously *nothing* wrong with including the heights of the tower in the choice class. In this case there seems to be something wrong with a choice class of explanans containing lengths of the shadow; but the pragmatic theory of explanation fails to specify why. No account of what makes choice classes of explanans (or in van Fraassen's account the relevance relation) reasonable is to be found in the theory.

2 The Tower and the Shadow

Van Fraassen himself is a little puzzling on this question. He seems to think that while there is in general no criterion for constraints on the relevance relation which are needed to prevent wildly counterintuitive consequences, there may be local variations of interest dependence which allow some kinds of facts to be relevant at some times, but not at others. His contribution is a little fable called 'The Tower and the Shadow' in which it is shown how the length of the shadow could indeed legitimately explain the height of the tower. A traveller learns from a libidinous ghost that a tower has been erected to a certain height in order that its shadow will be a certain length at sunset. This is because the tower's owner wanted to ensure that the

verandah on which he first declared his love for a mistress (probably the ghost) whom he murdered for her infidelities (probably committed with the traveller's father) should be in shadow at sunset.

The fable is offered with little comment; presumably it is meant to show that whether or not we take the length of the shadow to be a legitimate explanatory factor in an account of why the tower has the height that it does, depends on contextual pragmatics. In the context of the fable we will; in many other contexts we will not. What *kind* of pragmatic factors are at work, however? We have already seen that while choice class of explanans can capture a distinction between cases where the length of the shadow is relevant, it does so by mere stipulation. Further, *prima facie* it cannot even capture all such distinctions. Van Fraassen's fable is a case in point; if it were a case where the length of the shadow *did* count as explanatory; then lengths of the shadow should appear in the choice class of explanans. But this would not even formally distinguish it from the case where shadow lengths appear in the choice class but we want to rule out the explanation. In other words, the problem is not just of accounting for why it should be that shadow lengths should *not* appear in a choice class of explanans and other factors should, since on some occasions they are allowed in. Intuitively the missing ingredient seems to be causation; typically, lengths of shadows do not play any causal rôle in determining the height of the tower. This is why it seemed right that shadow lengths should not be included in choice classes of explanans. On the other hand, van Fraassen's fable seems to show us that lengths of shadows are sometimes relevant. This need not mean that the length of the shadow *now* in any way explains the height of the tower; rather, in van Fraassen's example, it seems to be the Chevalier's *belief* about the length of the shadow and its consequences which explains what is going on. Under some circumstances,

beliefs about what the length of the shadow would be do indeed play a causal rôle in fixing the height of the shadow, and so such beliefs are fit candidates for inclusion in a choice class of explanans constrained by causal relevance.

Perhaps in some cases the length of the shadow could itself cause the actuating belief—though not in the example as van Fraassen presents it. Suppose that the Chevalier had built his tower one layer of bricks at a time, and waited until sunset each day to see how long the shadow was. On the day when the shadow reached the required length, he stopped building. This might be a case where the length of the shadow itself (though not the *present* length of the shadow, or indeed the length of the shadow on any day but this crucial one) might play an explanatory rôle, but not for any mysterious and unspecified contextual reason, but rather because of the causal relevance of the length of the shadow on that day. In section 3, I will consider formulating a constraint on choice classes of explanans involving causal relevance.

2.1 Kitcher and Salmon's Objection

Philip Kitcher and Wesley Salmon [Kitcher and Salmon 1987] have a more generalized worry about putting constraints on what can properly count as a relevance relation in van Fraassen's sense. They show that if no external constraints are placed on the relevance relation, then for any two true propositions, there is a why-question which has one of them as the explanandum (or topic) and has the other as the explanans (essential part of the core of a perfect answer).

They start with van Fraassen's *unevaluated* theory of why-questions; that is the theory without any consideration of the criteria for answer's being *telling*. They take some arbitrary true proposition P_k , and some

contrast class X of which P_k is the only true member, and then define a relevance relation between some further arbitrary proposition A , and $\langle P_k, X \rangle$ such that A is the core of the only explanation of P_k .

They define the relevance relation, R , as $\{\langle A, \langle P_k, X \rangle \rangle\} \cup S$, where S is any set of ordered pairs $\langle Y, Z \rangle$ such that Y is a proposition and Z is $\langle V, W \rangle$ where V is a proposition and W a set of propositions, one of whose members is V . [Kitcher and Salmon 1987 p.319].

Let us consider the first half of their disjunctive definition first, $\{\langle A, \langle P_k, X \rangle \rangle\}$. From this part of the definition, A bears R to $\langle P_k, X \rangle$ by definition since R just is $\langle A, \langle P_k, X \rangle \rangle$. Given that A and P_k are true, and no other member of X is true, A is therefore the core of a direct answer to the why question $\langle P_k, X, R \rangle$. It remains to show that R can be more general, and it can still be the case that A is the only such direct answer. This is the function of S .

This can be guaranteed by placing restrictions on the other part of the definition, S . From above, S is any set of ordered pairs $\langle Y, Z \rangle$ such that Y is a proposition and Z is $\langle V, W \rangle$ where V is a proposition and W a set of propositions, one of whose members is V . If any of the Z were $\langle P_k, X \rangle$ (i.e., if for some Z , $V = P_k$ and $W = X$), then that particular Y would be relevant to $\langle P_k, X \rangle$. If, further, Y were both true and not identical to A , then there would be some other proposition which was a direct answer to $\langle P_k, X, R \rangle$. To avoid this possibility, S must be constrained so as not to allow that there is any $\langle Y, Z \rangle$ such that Y is true and Z is such that $V = P_k$ and $W = X$. This guarantees that there will be no other true proposition relevant to the explanandum and the contrast class, and thus that A is a unique direct answer to the why-question.

Kitcher and Salmon take this to be almost a *reductio* of the van Fraassen position. I find it, rather, to be quite unsurprising. Not even van Fraassen supposes that the relevance relation should be *totally* unconstrained, although what the constraints are supposed to be is not clear, as I have argued above. After all, it is called a *relevance* relation; and so it is surely not just *any* relation R that holds between some true proposition and $\langle P_k, X \rangle$. If no constraint is placed on relevance then of course given a pair of true propositions, it should be possible to dream up *some* question to which one proposition is the only answer to the other.

Though it may be unsurprising it does show one thing, which is that the van Fraassen model is no sufficient condition for explanation unless the notion of relevance is taken to be substantial; and it is no *account* of explanation unless that substance is spelled out.

2.1.1 A Version of the Kitcher–Salmon objection for Choice Class of Explanans

A benefit of the choice class of explanans model over van Fraassen's relevance relation is that the kind of complaint that Kitcher and Salmon make is seen much more easily for what it is. On the choice class of explanans model, (see Ch. 1 sec. 6) some true proposition in the choice class of explanans has to *select* the explanandum from the choice class of explanandum. The Kitcher–Salmon objection would amount to saying that, in the absence of any constraint on what counts as selection, all that was required to make one arbitrary proposition the only explanation of another, would be to make one the only true proposition in the choice class of explanans, and the other the only true proposition in the choice class of explanandum. This would certainly be true, and in this case transparently so. Selection of the explanandum is an essential feature of this version of a

pragmatic theory of explanation; and the model will not be complete until I have given an account of it.

2.2 A Constraint Rejected

One candidate for an account of the selection relation is to apply van Fraassen's own criteria for a *telling* answer. A true member, *A*, of the choice class of explanans *C*, selects the explanandum *B* from the choice class of explanandum *D* exactly if the purported explanans would be a *telling* answer to the a why-question $Q = \langle P_k, X, R \rangle$ where $P_k = B$, $X = D$, and the relation *R* holds and only holds between each member of *C* and $\langle P_k, X \rangle$.

This constraint fails because of the inadequacy of van Fraassen's account of tellingness. There are at least three criteria for tellingness; the answer must be probable in the light of the background theory, it must probabilistically favour the explanandum over the contrast class and it must fare well against competing answers judged by the first two criteria.

We can dispense with the details of van Fraassen's account of probabilistic favouring by dealing with what he takes to be the best case of probabilistic favouring: if the probability of the explanans given the background theory and the explanans equals 1, and the probabilities of the other members of the contrast class (of explanandum) given the background theory are 0 [van Fraassen 1980 p. 148]. If this condition is met, then no explanans can do better by the second criterion for tellingness. Further, van Fraassen says that this condition is met when the background theory plus the explanans *implies* the explanandum and the negation of each of the (other) members of the contrast class (of explanandum).

What kind of implication is meant here? I take it that entailment is meant; material implication would not be consistent with the probabilistic

reading, for A and B could both be true, and therefore $A \supset B$, without $P(A/B) = 1$. Suppose I throw a die, and a six comes up. The proposition that the die was thrown materially implies the proposition that a six comes up, but surely the *conditional* probability of the die coming up six given that it was thrown is $1/6$ (assuming a fair die).

To overcome this objection, let us suppose that the kind of implication required is entailment. So if the explanans plus the background theory entails the explanandum and the negations of each of the other members of the contrast class, then we will say that the explanans does as well as it can on this criterion.

2.2.1 The Second Kitcher-Salmon objection

Kitcher and Salmon have a second argument to show that the constraints on tellingness reduce to triviality. Their strategy is to show that there is a way of generating an answer B which is an answer which scores maximally well on van Fraassen's criteria of tellingness to any why-question $\langle P_k, X, R \rangle$, where some proposition A, the explanandum P_k , and the negations of each of the other members of the contrast class (henceforth $\sim Z$) are all included in K.

They define a relevance relation R:

- (1) R holds between some proposition, B, and $\langle P_k, X \rangle$ iff $B \vdash P_k$.

They then allow B to be the proposition $A \& (A \supset P_k) \& \sim Z$.

Now it is clear that B is relevant to $\langle P_k, X \rangle$, since both A and $A \supset P_k$ are conjuncts of B and hence $B \vdash P_k$.

The next question is whether B is probabilified by the background theory; in fact it scores perfectly here since A, P_k and $\sim Z$ are all in the background

theory, and since the 2 is the material hook, $A \supset P_k$ is a consequence of the background theory and therefore so is $A \& (A \supset P_k) \& \sim Z$.

The tower and the shadow is just such a case. Given our background theory, we may well know that the shadow is l metres long and the tower is t metres high. All that is required is a relevance relation which allows lengths of the shadow to be relevant, and a contrast class that contains heights of the tower that are known to be false, and the criteria for tellingness will read out that not only is the length of the shadow an answer to the why-question about the height of the tower, but it is a maximally telling one as well. In my version of the pragmatics, if the only constraint on what can be contained in the choice class of explanans is that it should contain the kinds of things which can potentially redistribute the probabilities in the choice class of explanandum, and if the only constraint on the relation of selection is that the explanans should meet the three van Fraassen criteria for tellingness, then the same asymmetry remains.

Van Fraassen does have a response to this, of course. Although material implication is the only kind of implication required to generate top marks for tellingness in the case that the explanans implies the explanandum and the negation of each of the other members of the contrast class, there are restrictions on what knowledge we can use to determine this implication. So it is not whether the explanans *actually* classically implies the explanandum that counts; it is whether we would *believe* that it does, given some suitable restriction on our knowledge base. He dismisses the full background theory K as being too strong; presumably because of just the problem above. Instead [van Fraassen 1980, p. 147] he introduces some restricted subset of our knowledge base, $K(Q)$, in the light of which the explanans is supposed to probabilistically favour the explanandum over the contrast class, the best case of which would be if $K(Q)$ implied the

explanandum and the negations of each member of the contrast class. This move is no defence against the Kitcher-Salmon second objection, however, since *whatever* the restrictions on K which yield K(Q), B (which is the putative explanans) plus K(Q) will yield the desired scores on the test of tellingness, just because B alone does.

Perhaps some further constraint would block the move, such as removing P_k from K, removing $\sim Z$ from K, or requiring that the probability of propositions (such as A) which are included in a putative explanans depend on K(Q) or some other kind of carving up of K.

Prima Facie these seem like reasonable suggestions—particularly the first; is it not consistent with the rest of the pragmatic theory that what will count as an answer—even a good answer—depends on some pragmatic carving up of the question and our prior knowledge? Is that not, after all, what the contrast or choice classes do? No they do not. The crucial difference is this; in the case of the contrastive analyses they were able to show how *given* a formulation of a why question pragmatically disambiguated, a certain answer was favoured. Certainly some questions might seem peculiarly uninteresting, and the interests underlying some questions might be rather odd—but it was always clear *given* those interests, how an answer might be appropriate. The odd seeming answer ‘because he was momentarily distracted by a cloud formation’ to the question ‘why did the chicken cross the road?’ becomes less odd when we realize that what was meant was ‘why did the chicken cross the road at *t*, rather than cross it seconds earlier?’ No matter how uninteresting such minutiae of chicken behaviour might be, or how odd the explanatory interests of someone investigating it, given those interests, there is nothing peculiar about the answer.

The problem with the constraints on K or K(Q) on the other hand, is that there does not seem to be any natural way to limit what is contained in the background theory. When the background theory is exhaustive, then often just in virtue of the fact that the explanandum occurred, we can be sure that the background theory will entail the explanandum, and just in virtue of the fact that the other members of the contrast class did not, we can be sure that the background theory will entail the negations of the propositions that *they* occurred. In which case, on van Fraassen's account, *any* true proposition A is going to count as a perfect answer to the why-question. So as knowledge becomes more complete, and background theories more exhaustive, we get closer to the paradoxical result that *anything* is an explanation. There are, of course, even cases where the knowledge does not have to be unrealistically complete to make entailments with undesirable consequences loom large; the causal asymmetries are such cases. Does, for example, the length of the shadow plus the background theory entail the height of the tower? An answer to this depends suspiciously on the details of an account of the required kind of entailment; and an account which comes up with the right constraint, I conjecture, will do so in virtue of mimicking an account of causal consequence.

If this were not the case it would be hard to see how these are supposed to be motivated; surely no matter *what* the motivation for some restriction on K that gives a K(Q) which contains information about the length of the shadow and the height of the tower (and hence generates the problems above), we want to reject the explanation of the height of the tower in terms of the length of its shadow. And even if one did think that there were some contexts in which such a direction of explanation were legitimate—suppose one thought that the *Tower and the Shadow* were a case of an explanation of the height by the length (rather than by a belief)—then

there is still nothing about the case which seems to impose a natural constraint on K in the way that the contrastive pragmatics naturally disambiguate why-questions.

3 The Causal Constraint

So what is required are further constraints on a model of explanation. In particular, the notion of *selection* has to be spelled out. I will do this for the case of the explanation of events; I do not have an apodeictic argument that shows that this is the only kind of explanation, though I will later argue that most other kinds of explanation are generalizations about event explanations. As I foreshadowed above, what I take to be intuitively missing is that a potential explanans has to be in some way *causally relevant* to the explanandum. In this section I shall develop an account of causal relevance which will be used as a constraint on the selection relation. Some of it follows the spirit of David Lewis' analysis of causation in [Lewis 1986b] and causal explanation in [Lewis 1986d]. I agree with Lewis that it is a necessary condition of a successful explanation of an event that the explanans be what *he calls* a cause of the event. I disagree in that I do not think that it is a *sufficient* condition, and in so far as I do not think that, typically, what he calls causes are in fact causes.

3.1 Constraints on the Selection Relation

For the case of *event*¹ explanation, I define selection as follows: for some choice class of explanans $A = \{a_1, \dots, a_n\}$ and some choice class of

¹The kind of events I mean here are the fine-grained kind, similar to Kim's [Kim 1973]; with suitable restrictions (I will discuss events in more detail in Ch 4 and 5) the realization of any property of an object will count as an event.

explanandum $B = \{b_1, \dots, b_k, \dots, b_n\}$, some member a_j of the choice class of explanans will be said to *select* the explanandum b_k iff

- (i) a_j is *causally relevant* to b_k (in Lewis' analysis a_j is a cause of b_k).
- (ii) for all b_i in B : $i \neq k$, b_i depends causally on $\sim a_j$.
- (iii) the selection of b_k by a_j depends, in some way, on real causes in the microstructure.

3.2 Counterfactuals and Causal Dependency

These constraints depend on the notions of *causal dependency* and *real causal relations*. Causal dependency is the same notion as in [Lewis 1986b]. It in turn is defined in terms of counterfactuals. Counterfactuals take as primitive propositions, analysed in terms of truth, worlds and relations of similarity between worlds. The proposition that if A were true, so would C is the counterfactual of A and C , $A \Box \rightarrow C$. It is true at a world w iff either there are *no* worlds at which A is true (in which case Lewis calls it *vacuous*) or else there is some world at which A is true and C is true which is more similar to w than any world where A is true and C is not.

Counterfactuals are supposed to provide the basis for the account of causation that Lewis goes on to give. Causal dependency among *events* is then defined roughly as follows. For any possible event, e , there is a *proposition* that e occurs, $O(e)$, which is true at all those worlds where e occurs. Thus we are able to forge a link between counterfactual relations between *propositions*, on the one hand, and relations between events on the other. Some event e *causally depends* on some event c iff $(O(c) \Box \rightarrow O(e)) \ \& \ (\sim O(c) \Box \rightarrow \sim O(e))$. The conjunctive formulation is to allow the notion to remain defined for both cases where c and e and the case where they do

not. In the case where c and e actually occur, $O(c)$ and $O(e)$ are both true, so the first conjunct is true. Therefore the truth of the conjunction depends on the second conjunct; ie on whether if it were *not* the case that $O(c)$ were true, *nor* would it be the case that $O(e)$ were true. In the case where c and e do *not* actually occur, the second conjunct is true since $\sim O(c)$ and $\sim O(e)$ are both true. So in this case the truth of the conjunction depends on the truth of $O(c) \Box \rightarrow O(e)$; ie on whether if $O(c)$ were true, so would be $O(e)$.

3.3 Causal Relevance and Chains of Counterfactual Dependency

A *chain of counterfactual dependency* is any string of events $c, d, e \dots$ such that d causally depends on c , e causally depends on d and so on. An event, a , is *causally relevant* to an event, b , just if there is a chain of counterfactual dependency from a to b . Being causally relevant in this sense is roughly what Lewis takes to be the property of being a cause. For reasons which will become apparent in the discussion of real causal relations below, and which are elaborated more fully in chapter 5, I do not use that terminology here. But I do allow that causal relevance captures roughly what is meant in the Folk conception of causes; if bombs cause explosions, droughts cause starvation, cups of coffee cause alertness and so on, then causally relevant events are causes. It is because I will argue later that there is much to be said for allowing real causal relations to occur only at some as yet not determined microstructural level of causation, that I do not allow causally relevant events to necessarily be causes.

4 The First Constraint

The first constraint is that of causal relevance; a_j must be causally relevant to b_k . The point of this constraint is to deal with the asymmetries of explanation; and it answers to the intuition that in most of these cases it is a causal asymmetry which underlies the explanatory one. In the case of the

tower and the shadow, the idea is that the length of the shadow is not causally relevant to the height of the tower; although *beliefs* about the length that shadow will be are causally relevant to the creation of the tower in van Fraassen's *Tower and Shadow* fable.² The idea is that any successful explanation of an event is successful in virtue of causal relevance, though not every event causally relevant to an explanandum is a successful explanation.

Expressed in these terms, though, the requirement looks like victory by *fiat*. Does not the problem of asymmetry simply reoccur with just as much force at the level of *causal* asymmetry? If this were true, I would still think that progress had been made. Two problems are reduced to one; and that is some kind of progress. Even in van Fraassen's unmodified model, one possible relevance relation is that of figuring in the *ætiology* of an event, and for it to be determinate whether some event passes this test of relevance, some kind of solution to the problems of causal asymmetry is required. Given that solution, the constraint on selection above applies it to event explanation in general.

4.1 A Counterfactual Approach to the Asymmetry

In fact, I think the prospects are better than this. Explanatory asymmetry remains mysterious; if it depends on context, no story has been offered about what the contextual constraints might be that rule out the cases that go in the wrong direction. Causal asymmetries, while difficult, remain a

²Of course, as mentioned above, a slight change to the story could make the length of the shadow at some time causally relevant to the beliefs. But this would not be a case where some mysterious contextual factor makes makes some factor which is *not* causally relevant explanatorily relevant.

problem that needs solution regardless of one's particular views on explanation. In the case of the criterion of causal relevance outlined above, because it is equivalent to Lewis' account of what causes are, we have available his own solutions to the problems of effects, epiphenomena, preemptive and redundant causation. The causal asymmetry in the tower and the shadow case seems to be an instance of the problem of effects. The height of the tower is causally relevant to the length of the shadow, yet we want to be able to say, *ceteris paribus*, both that the length of the shadow is not causally relevant to the height of the tower, and that if the length of the shadow had been different, then so would the height of the tower. Lewis's response in cases of the problem of effects is simply to deny the troublesome counterfactual; if we were to apply it to this case it would mean denying that if the shadow were of a different length, then so would the height of the tower. This is explicated in terms of similarity between worlds; it makes more of a difference to the world to hold the height of the tower fixed and have the different length of the shadow otherwise caused, than to change the length of the shadow and go back and abolish *its* cause in the actual world as well.

I offer the following as an argument in the spirit of Lewis for the view that the world in which the height of the tower remains fixed but the length of the shadow varies, is more similar to our world than the world in which both the length of the shadow and the height of the tower are different.

Presumably the motivation for supposing that the world in which the height of the tower is also different is more similar to the actual world, is that in that world the laws which relate the height of the tower and the length of the shadow are preserved; roughly—without begging the question about causes—the intuition is that a world in which the invariant causes of effects are changed as well as the effects is more similar to an original world

than ones in which just the effects vary. So on this intuition, some world w' where the height of the tower and the length of the shadow are both changed is more similar to the world with the actual tower³ than some world w where only the length of the shadow is varied. But if the motivation for this is respected, then surely some other world w'' where the proximate cause of the change in the height of the tower is also changed will be closer to the original world than w' or w , since it will not have some dangling change of effect without a corresponding change of cause. These considerations may push us back through the causal (or nomic) ancestry of the world until we are committed to claiming that some world $w''''''...$ is more similar to the original world than w , despite the fact that w shares the entire history of the original world up until the change in the shadow length, and $w''''''...$ is entirely different in its early history and only gets reasonably similar at the point of changed shadow length—and there are no guarantees that it will stay similar afterwards. In the light of this it is perhaps tolerable to put up with any violation of the original world's laws in w , and allow that w is after all more similar to the original world than w' , and thus that the counterfactual, (the length of the shadow is Y) $\Box \rightarrow$ (the height of the tower is X) is false.

4.2 The Reichenbach-Salmon Approach

It is, of course, not *necessary* to buy into a whole story about counterfactual dependence to have the benefit of reducing the problem of the asymmetry of explanation to asymmetry of causation. With suitable constraints Reichenbach's and Salmon's work on causation [Reichenbach 1956] [Salmon

³I take it that there is no such tower (or Chevalier)!; this is the 'actual' tower in the dialect of the inhabitants of the world where the tower of the story exists.

1984] can be used to define a notion of causal relevance which will do the required work.

Salmon takes *causal processes* as the basis of a theory of causation. Causal processes are distinguished from *psuedo processes* by Reichenbach's principle of mark transmission: roughly, that a change to a genuine causal process will be propagated by that process, whereas a change to a psuedo process will not. For example, if a white spotlight is played across a field, and if a red filter is placed across the light, then the spot of light moving across the field will, (assuming standard coloration of the field), *ceteris paribus*, remain red until some further intervention in the causal process. If, on the other hand a red card is placed on the field to colour the white spot red at a point, as soon as the spot passes the point it will revert to white. Salmon formulates the principle of mark transmission (MT) as follows:

MT: Let P be a process that, in the absence of interactions with other processes, would remain uniform with respect to a characteristic Q, which it would manifest consistently over an interval that includes both the space-time points A and B ($A \neq B$). Then, a *mark* (consisting of a modification of Q into Q'), which has been introduced into process P by means of a single local interaction at point A, is *transmitted* to point B if P manifests the modification Q' at B and at all stages of the process between A and B without additional interventions. [Salmon 1984 p.148].

As Salmon immediately notes, this principle is counterfactual⁴, as it relies on the claim that P is a process that *would* remain uniform in the absence of

⁴It has to be to rule out the following kind of objection to an actualist version (delivered in conversation to Salmon by Nancy Cartwright). Suppose that an intervention in the causal process happens simultaneously with an intervention in the purported psuedo process; then it may look as though the intervention in the psuedo process is being propagated though we think that it is due to the intervention in the causal process.

interactions. So if the counterfactual account of causal relevance above did not appeal simply because of its use of counterfactuals, then a version descended from Salmon will have no more appeal.

The MT principle is used as the basis for a distinction between real causal processes and *psuedo*-processes, like the movement of a spot of light. One additional motivation for denying that processes like moving spots of light are real causal processes, is that they can propagate faster than light; consider a vast Galactic Lighthouse sending a laser signal out from Earth. Allow it to rotate reasonably fast. By the time the signal reaches the furthest parts of the universe, if the distance is far enough and/or the lighthouse rotation speed is great enough, the play of the beam will move across intercepting surfaces faster than light, even though the beam itself—the real process—is propagated only at lightspeed.

In addition to the MT principle to distinguish real causal processes, Salmon has a Reichenbachian apparatus of *forks* to describe the interaction and intersection of genuinely causal processes. The kinds of forks which order causal processes are *interactive forks*. The idea is that causal processes may intersect at a point, and transmit a difference to the processes that continue past that point. Salmon formulates this principle of causal interaction (CI) as follows:

CI: Let P_1 and P_2 be two causal processes that intersect with one another at the space-time point S which belongs to the histories of both. Let Q be a characteristic that process P_1 would exhibit throughout an interval (which includes subintervals on both sides of S in the history of P_1) if the intersection with P_2 did not occur; let R be a characteristic that process P_2 would exhibit throughout an interval (which includes subintervals on both sides of S in the history of P_2) if the intersection with P_2 did not occur. Then, the intersection of P_1 and P_2 at S constitutes a causal interaction if:

(1) P_1 exhibits the characteristic Q before S , but it exhibits a modified characteristic Q' throughout an interval immediately following S ; and

(2) P2 exhibits the characteristic R before S, but it exhibits a modified characteristic R' throughout an interval immediately following S. [Salmon 1984 p. 171]

This principle does not have inherent direction; if two processes interact and change their 'characteristics' Q and R to Q' and R' respectively, there is nothing in this definition to preclude saying that Q' is *erased* by the causal interaction. But in the event that better progress is made on the question of thermodynamic asymmetry, perhaps in its guise (as favoured by Salmon) of concern with irreversability of changes, then the forks could be directed forks. Causal relevance (for the purposes of my *first* constraint) could then be defined as follows (setting aside problems of overdetermination or preemption): an event, *a* is causally relevant to an event *b* iff both *a* and *b* are *located on* causal processes such that there is an unbroken chain of causal processes and interactive forks between *a* and *b*, and every fork between *a* and *b* is oriented so that *b* is *temporally downstream* from it. Some event is located on a causal process just if it supervenes on some segment of that process.

This version of the constraint on the selection relation similarly helps with the asymmetries. The causal process requirement eliminates explanations of the position of the shadow at time *t* in terms of the position of the shadow at time *t'*, since that will turn out to be a psuedo process. The explanation of the height of the tower by the length of the shadow will be ruled out by the causal process requirement together with the temporal asymmetry requirement, since the shadow's length at any time will turn out to be caused by the height of the tower (and background conditions) at an earlier time.

It is perhaps appropriate to note here that Salmon's justification of the counterfactual elements in the principles MT and CI above, is in terms of

experimental interventions. What to keep fixed in establishing counterfactuals is just what an experimenter working in a particular science would do; and the *truth* of such counterfactuals (though not an account of them) is supposed to be obtainable from controlled experimentation. Such an approach can be applied to the case of the tower and shadow to support the elimination of the troublesome counterfactual for which an argument in the spirit of Lewis is given above. In determining whether the tower causes the shadow, no matter how often we shine lights into the shadow to eliminate it, or reduce its length, we will produce no change in the height of the tower. Of course we *could* change the height of the tower in order to change the length of the shadow, but the methodological intuition that that is no fair test of whether the resultant change in the length of the shadow causes the change in the height of the tower is a strong one; and, I suspect, is closely related in counterfactual terms to a stipulation against backtracking counterfactuals for causal purposes.

5 The Second Constraint

The second constraint on the selection relation is that for all b_i in B : $i \neq k$, b_i depends causally on $\sim a_j$. For this constraint, only the counterfactual formulation of causal relevance will do. This is because every b_i in B such that $i \neq k$ does not occur; these are the events in the choice class of explanandum other than the explanandum itself, and none of them occurs. So no story about causal processes will capture the kind of causal dependence required, since no actual causal process provides the supervenience base for any of these events.

The motivation for this constraint is that it does the work of the 'rather than' clause in the informal expression of a why question as a question about a rather than b . It provides a way of selecting from all the events which are

causally relevant to the explanandum, those which select the explanandum rather than the rest of the choice class of explanandum. The causal dependence of each member of the choice class of explanandum on the non-occurrence of the explanans expresses the ideal that the explanans is also causally relevant to the failure of the non-occurring members of the choice class to occur.

Let us consider an example. The choice class of explanandum contains two events; my walking to work and my driving to work. In fact I walked to work. The choice class of explanans contains two events—my car's having broken down and my desire to attend a seminar. Both my desire to attend a seminar and my car's having broken down are causally relevant to my walking to work—the explanandum event. My driving to work, however, causally depends on the non-occurrence of my car's breaking down, and thus my car's breaking down *selects* my walking to work from the choice class, whereas my desire to attend a seminar does not.

To see how the constraint works, let us consider again the definition of causal dependence. Some event e *causally depends* on some event c iff $(O(c) \square \rightarrow O(e)) \& (\sim O(c) \square \rightarrow \sim O(e))$. In the case of the members of the choice class of explanandum other than the explanandum itself, each of them does not occur so each proposition about their occurrence $O(b_i)$ is false. The explanans, *ex hypothesi*, occurs so the proposition $O(\sim a_j)$ that expresses its non-occurrence is false. Thus the second half of the conjunction is satisfied, so the truth of the causal dependency depends on the truth of the first conjunct. This will be true just if, for each b_i $i \neq k$, there is a world in which the proposition $O(\sim a_j)$ is true (because a_j does not occur in that world) and $O(b_i)$ is true, which is more similar to the actual world than any world in which $O(\sim a_j)$ is true and $O(b_i)$ is false. Note that the possibility that there is no *closest* world in which $O(\sim a_j)$ is true is allowed: in the case where there is

a choice class of explanandum of more than two members, we must allow that for each b_i other than the explanandum, there is a world in which the explanans does not occur and that b_i does, and that world is no more or less similar to the actual world than any other world in which the explanans does not occur and some *other* b_i other than the explanandum does.⁵

6 The Third Constraint

The third constraint is required if you think that the level at which causation operates in the universe is somewhere deep in its microstructure, and that causation may eventually be able to be *identified* with certain physical interactions in that microstructure. My purpose here is not to defend this view—though I am somewhat sympathetic to it and offer some considerations in its favour in ch. 5—but rather to examine what the consequences of this view are for a view of explanation which takes seriously the requirement that explanation depends on there being causal connexions between the explanans and the explanandum.

Causes, then, are microstructural features; but if selection of the explanandum by the explanans is supposed to depend on these real causes, it is not immediately obvious that this can be got from the constraints imposed so far. For there is nothing to say that the kinds of events which we can take as *relata* in this account are microstructural events. Nothing excludes explosions, crashes and so forth from featuring as explanantia or explananda. Nor should anything; crashes surely count as paradigm potential explanations of explosions. Yet if a crash is not *identical* with some

⁵This applies to the limiting case where the $b_i : i \neq k$ are not compossible; the situation is slightly more complicated if some or all of them are, for we have to allow for equi-similarity of worlds in which a_j does not occur, and various combinations of the b_i do.

microstructurally specified event, and similarly the explosion, then on the microstructural account of real causation no crash was ever the cause of any explosion. If we wish to preserve all these intuitions, we will need to have an account of how an explanans can select an explanandum, while not being a cause of it (though being causally relevant to it) and yet having this explanation depend on real microstructural causes. That project is the work of the next chapter.

7 Some Problems of Generality

So far I have only talked about the explanation of particular events. Obviously a lot of requests for explanation are more general than that. We ask why it is that trees drop their leaves in winter, or that salt when placed in water dissolves, or that economies whose money supply increases suddenly exhibit inflation. Similarly to Lewis [1986d p.225] I think that what is happening here is that answers to these kinds of requests for explanation provide information about the similarities between chains of causally relevant events in the causal histories of individual events which fall under generalizations. The causally relevant events in each case of salt dissolving are similar in various ways, and providing this kind of general information is just providing an answer to general explanation requests. The second constraint above works in the same way as in the single case; to explain in general why salt dissolves in water rather than hangs in suspension is to provide different information about the similarities in the chain of events causally relevant to the various token events of dissolving salt than would be required if we wanted to know why salt dissolved rather than precipitated out.

I make no attempt to spell out a formal model of such generalizations, but I do not think that it will be out of the spirit of the model of event

explanation. What will still be required is that information is given in virtue of which we have reason to believe that events of a kind constrained by the pragmatics occur in each case. If one had to make this into a motto it would just be "if you ask a general question, you get a general answer."

The fact that such generalizations are not always perfectly true may underlie one kind of probabilistic explanation (for a discussion of more kinds, see below). If events of a certain kind are said to be probabilistically relevant to events of another, it may just be that some but not all of the tokens are (deterministically) causally related.

7.1 Common Causes

A special case of such generalizations is that of common cause explanations. A common cause explanation is where two events, the conjunction of which is fairly improbable, is explained in terms of a common cause, and the probability of the conjunction of the events given the common cause is greater than the probability of the conjunction of the events *simpliciter*. Reichenbach, in [Reichenbach 1956 sec. 19] describes a statistical structure he called a *conjunctive fork*. It is defined by the following equations and inequalities:

- (1) $P(A.B | C) = P(A | C) \times P(B | C)$
- (2) $P(A.B | \sim C) = P(A | \sim C) \times P(B | \sim C)$
- (3) $P(A | C) > P(A | \sim C)$
- (4) $P(B | C) > P(B | \sim C)$

One consequence of this definition is that $P(A.B) > P(A) \times P(B)$ ⁶, i.e. that the probability of A and B *simpliciter* is higher than the product of the probabilities of A *simpliciter* and B *simpliciter*. Thus A and B are not probabilistically independent of one another. The idea is that when we come across two explananda, A and B, which are not probabilistically independent of one another in this way, and we have reason to suppose that there is no interaction between them, we should hypothesize a *common cause*, C, which is related to A and B as set out in the definition above. It must both probabilify A and B separately ((3) and (4)), and the probability of A and B *given* C (or $\sim C$) must be independent of the probabilities of A given C (or $\sim C$) and B given C (or $\sim C$).⁷

Salmon and van Fraassen take it that such a common cause (construed as the set of probabilistic relations themselves; the conjunctive fork relation, rather than the actual cause *in re*) *explains* the apparent statistical dependence of A on B. On such a reading of common cause explanations, it seems as though it is the statistical correlation that is the explanandum, and the statistical relation of the common cause to the explanandum that is the explanans. Salmon, though, usually keen to keep things causal, has an extra constraint that in addition to meeting the statistical definition, a *bona fide* common cause must be in fact related along lines of causal process. It is hard to reconcile this with the idea that the conjunctive fork is an *explanatory* notion.[Salmon 1984 p.168].

⁶For a proof of this see [Salmon 1984 p. 160] or [Reichenbach 1956 pp 160-1].

⁷A consequence of this is that C will *screen off* A from B and vice versa, i.e. in the presence of C, A is statistically irrelevant to B and B is statistically irrelevant to A.

I do not think that such statistical relations *have* to play a crucial rôle in explanations themselves. It is clear that reasoning from such probabilistic dependencies to the existence of common causes which have the appropriate statistical relations, constitutes a valuable route to explanatory discovery; but I think its rôle can be described as an *evidential* one. In a common cause explanation what *explains* A is C—if it is causally relevant to it—and similarly what explains B is C. The role of the statistical correlations is to provide evidence that, whatever the explanations of A and B, they are in some respects similar. In other words, it provides *evidence* that there is some explanation of A which is also an explanation of B. The scientific task is then to find this explanation; but there is nothing that C explains over and above explaining A and explaining B.

Salmon himself admits that in some cases we need to add the stipulation that not only must three events be linked by the four probabilistic relations above, but there must be *bona fide* causal processes linking them. He gives the following example, due to Ellis Crasnow:

Consider a man who usually arrives at his office about 9:00 A.M., makes a cup of coffee, and settles down to read the morning paper. On some occasions, however, he arrives promptly at 8:00 A.M., and on these very same mornings, he is met at his office by one of his associates who normally works at a different location. Now, if we consider the fact that the coffee is already made when he arrives (*A*) and the fact that his associate shows up on that morning (*B*) as the coincidence to be explained, then it might be noted that on such mornings he always catches the 7.00 A.M. bus (*C*), while on other mornings he usually takes the 8.00 A.M. bus ($\sim C$). In this example, it is plausible enough to suppose that *A*, *B*, and *C* form a conjunctive fork satisfying (1)–(4), but obviously *C* cannot be considered a cause either of *A* or of *B*. The actual common cause is an entirely different event *C'*, namely, a telephone appointment made the day before by his secretary. *C'* is, in fact, the common cause of *A*, *B*, and *C*. [Salmon 1984 p 168]

It is to cope with cases like this that Salmon introduces the extra requirement that for a *bona fide* common cause to exist there must be the right kind of causal mechanisms. Having admitted that, though, why not allow that what does the explaining is information about those causes—including evidential information, which is what the conjunctive fork would amount to on this reading. So finding such a statistical relation should not *of itself* explain the surprising correlation; it explains insofar as, and only insofar as, it amounts to information about causes.

The selection of the common causes as being the causes of interest in particular cases is perhaps a largely pragmatic matter. If we are seeking an explanation of an A, and note that despite our conviction that As and Bs do not interact, there is in general a statistical dependency between A and B, then we may suspect that there is a cause in common between this A and some B that presents itself. This directs our explanatory interest towards a certain subclass of events causally relevant to A. If we include in the choice class of explanandum $\sim B$, then of the events causally relevant to A we will be concerned only with those whose non-occurrences are causally relevant to $\sim B$. Such events will be causally relevant to both A and B.

8 Extension to the Probabilistic Case

The primary account of causal relevance which I have offered has been a modal one, despite the additional requirement that actual physical causal relations in the microstructure be involved, as well as the more abstract criterion of causal relevance. In the case where two events *a* and *b* actually occur, and *a* is causally relevant to *b*, then not only must the counterfactual $O(a) \Box \rightarrow O(b)$ be true, but so must $\sim O(a) \Box \rightarrow \sim O(b)$, so in the possible world most similar to the actual world in which *a* does not occur, nor does *b*. This does not mean that *a*'s occurrence is strictly *necessary* to *b*'s occurrence—

there will be worlds where *b* occurs without *a*, but it does constitute the claim that had *a* not occurred, nor would have *b*.

This is enough to bring on the worry that Salmon has about modal accounts of explanation. In his strictures against von Wright [von Wright 1971] (for ruling out probabilistic explanation) and Mellor [Mellor 1976] (for failing to find a satisfactory way to rule it in), he is concerned that it is not easy to provide an account of probabilistic explanation if one is committed to modal claims like had *a* not occurred, nor would have *b*. This is because there is no mention of probabilities, just of determinate occurrence or non-occurrence of events.

I do not have a full account of probabilistic explanation, but I will offer some directions in which it might go, which would account for such explanatory practices. Probabilistic explanation needs to be divided into at least two categories; ineliminable probabilistic processes which it seems likely occur in the microstructure of the world, and macroscopic everyday probabilistic explanation which is meant to be consistent with micro-determinism, and is in any case not dependent on micro-indeterminism for its probabilistic character.

In the first case I see no reason why an event could not be causally relevant to the ineliminable probability—henceforth chance—of another event's occurring. So if an atom of Uranium has a chance *C* of decaying during some interval, there may be events which are causally relevant to

this chance. In the allegedly objectionable modal formulation, if *e* had not occurred, the atom would not have had chance C of decaying.⁸

The macroscopic case is very different. One kind of macroscopic statistical explanation is that which I have already mentioned: causally relevant generalizations which are not completely accurate. The more accurate the generalization, the better the probabilistic explanation of the classes of events. In the single case explanation, it would be possible to reconstruct an allegedly probabilistic explanation as *evidential*. The generalization about causally relevant events provides evidence for their being an explanation of a certain kind in the single instance. Suppose people who eat tainted seafood often suffer stomach complaints to which their eating of seafood is causally relevant, and someone who eats seafood indeed gets sick. The statistical correlation is not itself an explanation; it is rather evidence that, with a certain (subjective) probability the eating of seafood could be found to be explanatory in this case. This kind of case is not, strictly speaking, single case probabilistic explanation. The probabilistic correlations are evidence that there is a single case *deterministic* explanation of a certain kind to be found.

8.1 Genuine Single Case Probabilistic Explanation

There is another way to go in the single case probability. We may wish to accommodate the intuition that although, say, the eating of seafood was in

⁸If the language sounds a little unnatural here it is because of the very liberal notion of events at work, the atom's having chance C of decaying is an event, and the atom's being in a certain configuration of particles is also an event. Of course its being in a certain configuration of particles is also what makes it a Uranium atom of the type that it is, which is why we are able to cite the fact that it is a particular isotope of uranium as an explanation of its chance of decaying in the interval.

fact causally relevant to a particular person's illness, it could have been the case that that *very same* eating of seafood was not causally relevant to the illness, so that the explanation is somehow probabilistic, even though under deterministic laws we have a full account of what happened.

Even this intuition can, if necessary, be met. Suppose that macroscopic events like eating seafood are *multiply realizable* at some microstructural level. Not every feature of the event at this level is necessary to them, so the very same event might have been different.⁹ We might want to say that had my lunch companion dropped her knife, momentarily distracting me, it might have made a difference to my eating of seafood—perhaps I would have distractedly eaten the crucial extra milligram of oyster that made me sick—while insisting that it was the very same event that it made a difference to. To deny this is to deny the possibility that anything can make a difference to an event, rather than bring a different event into being. Perhaps we might want to say that an explanandum event is only probabilistically explained when, although there is an explanans whose realization is causally relevant to the realization of the explanandum, it is not true that *all* pairs of realizations of the two events that would count as the same (token) events would be so related. If we took this line, then single case probabilistic explanation would still obey the causal constraints, but for *realizations* of events.¹⁰ This should become clearer in the light of ch. 4 on levels of organization and explanation.

⁹Pace Davidson [Davidson 1980a] For a fuller discussion of events see Chs. 4 and 5.

¹⁰There is an alternative. Suppose we go with Davidson and require that events are individuated by all of their causes and effects, and in addition, that causes operate at the microstructural level. Coupled with the doctrine that events are unrepeatable particulars,

9 The Limits of Causal Relevance

Perhaps it seems as though there are kinds of explanation for which causal relevance plays no part. One good candidate is mathematical explanation. If we are trying to explain why a theorem is true by pointing to various lemmas and arguing, for example, that theorems of that kind are always true of systems of a certain kind, then it seems that something which is at least *called* explanation is going on.

I do not want a fight about terms; let it be called explanation if necessary. But it should be clear that it is a very different enterprise from the task of finding out why the world is as it is. It may well be captured by abstract epistemic conceptions of explanation like Achinstein's [Achinstein 1983] in which explanations are acts designed to bring about understanding. But we must remember that we were brought to the causal constraint by the observation that, in many cases, only by imposing such a constraint could

every feature of which is essential to them, we would have an out and out denial of the possibility of single case probabilistic event explanation in cases other than those of irreducible probabilistic relations. But those intuitions could be accommodated by observing that in most of our dealings in the world we are not concerned with particular events at all, even when we think we are, but rather with relevantly similar *classes* of events—including non-actual events in worlds similar to our own. The so-called single case probabilistic explanations would then be ones where although in fact the explanans was causally relevant to the explanandum (and indeed, in virtue of the microstructural specification of the event, a cause of it) it is not the case that all the other, non-occurrent members of the interesting class of events would be so related if they occurred. The strength of the probability would depend on the proportion of relevantly similar non-actual events which were causally related in the right way. This sort of vaguely Davidsonian account is, I think, consistent, and has what might be a benefit in allowing high-level events to be *causes* (in virtue of their univocal realization in the microstructure). However the price that has to be paid is denying that it is usually particular macroscopic events that we are concerned with, and in particular which usually constitute explanatory *relata*.

we eliminate *completely unacceptable* 'explanations'. It is not that there is some true, general, account of explanation, with causal relevance as just one interest which could inform it. As we saw in our discussion of unconstrained pragmatic accounts, there are some explanations which are not good given *any* interests, and a way to eliminate them is by invoking causal relevance. So an account of explanation which is so general that it does not have in it some constraints which eliminate non-explanations is not a complete account. And any account which does not assign to mathematical and causal explanations quite disjoint constraints will be incomplete in this sense. So even if mathematical 'explanation' shares some epistemic features with causal explanation, this is not because they are both part of some true and complete theory at a higher level of generalization than the causally constrained theory I have given (unless merely the union of the theories and an account of the similarities between them). Rather, any account of mathematical 'explanation'—and I do not offer one—must impose its own constraints on a theory as abstract as the unconstrained pragmatic theory, or an understanding theory, which will then be at the same level of generality as causally constrained explanation. So mathematical explanation, if it is to be called that, can, with some motivation, be consigned to a quite different domain, rather than seen as part of a comprehensive account of explanation of which causal explanation is only a part—despite the fact that the giving of mathematical explanations may have similar psychological effects to the giving of causal explanations.

9.1 A Cosmological Explanation Ruled Out

Another limitation of causal relevance is that there are explanations in physics which, although they look *prima facie* as though they inhabit the same domain as those which depend on causal relevance, turn out not to be

capturable within a causal model. One such example is due to Huw Price [Price 1988].

One difficulty which besets theories of causation is determining the *direction* of causation. If we observe a body of gas diffusing out of a flask, there seems to be nothing about the causal processes which is inherently directional; we may as well reverse the process from the perspective of the causal interactions and describe what is happening as the gas going *into* the flask. Yet we think that an explanation of the gas's movement in terms of *initial states* of the gas is appropriate, and that an explanation of the gas's movement in terms of the *final* state of the gas is improper. The state of the gas at t (plus background) explains the state of the gas at $t+n$, but not vice-versa.

A usual approach to this is to observe that it is a fact about this universe that it has certain *thermodynamic asymmetries*. In particular, the Second Law of thermodynamics holds, and there is a global increase in entropy. The direction of this increase can be used to give the direction of time, and hence the direction of causation (or, for example, in the case of [Mellor 1981] the direction of causation and hence of time). Having established this direction we are in a position to say why we prefer an explanation which follows the causal constraint to be in terms of initial conditions rather than final conditions.

Suppose, however, we are trying to explain the fact about the Universe that there are such thermodynamic asymmetries. One approach popular among cosmologists is an explanation in terms of initial conditions; (see, for example, [Davies 1974]). The idea is that there was a state of the universe close to the big bang such that, in virtue of the universe's having been in that state at that time, the thermodynamic asymmetries hold. This is taken

to be an explanation of the thermodynamic asymmetries in terms of initial conditions. The justification for preferring initial condition explanations, however, rests on the thermodynamic asymmetries themselves, because the direction of causation and hence of causal relevance depends on them.

So the thermodynamic asymmetries are required in a justification of the explanatory practice which prefers initial condition explanations because of their causal relevance. Is there not then a vicious circularity in proffering an initial condition explanation of thermodynamic asymmetries? *Given* the thermodynamic asymmetries, we can rightly allow an explanation of the gas diffusing from the flask in terms of initial conditions, but an initial condition explanation of the thermodynamic asymmetries themselves is more problematic. So either we will have to object to these apparently good scientific explanations of the thermodynamic asymmetry of the universe—something to undertake cautiously on philosophical grounds alone—or else revise the account of causal relevance.

I do not think that such revision is called for by these cases; I stick by the causal relevance requirement. Perhaps certain initial conditions entail thermodynamic asymmetries, but then so do certain *final* conditions. Indeed, even calling the conditions initial or final depends on the asymmetries. It seems that the conditions and the asymmetries are both aspects of the same fact about the universe. Perhaps what is going on here is very much like what happens in mathematical 'explanation', seeing the (truth-conditionally) same fact described in various ways has a psychological effect—it brings about understanding—but it is no explanation about why it is the case. To say this is not to deny that the development of initial condition stories about the universe is good *science*; just that interpreting them as *explanatory* is inappropriate. They may well be

perfectly good (and illuminating) scientific descriptions of the phenomena. So perhaps we can be forgiven a little philosophical *hubris*.

This leads to a conjecture. The alleged explanation of the thermodynamic asymmetries by initial conditions is only a special case of the fact that *ultimate* cosmological explanation in general has to be outside time, because in part what is to be explained is why there is a system with temporal direction. Part of the unease which has greeted super-cosmological explanation in general may be accounted for by the fact that no proffered explanation of universes could be causally relevant in the way that causes work in the actual world. This is not to rule out the possibility of a more general notion of cause emerging—perhaps with some higher dimension in a higher order topology playing the rôle that time plays for our everyday notion of cause—but merely to postulate that much of cosmology is best postulated as descriptive. This also has the desirable consequence that, for example, wildly speculative metaphysics like some of the stronger versions of the Anthropic principle can be ruled out since, if there are no *explanations* to be had, it is no evidence for the truth of a description that such and such an explanation is called for, where that explanation is modeled on causal ones.

10 What kind of explanation is a causally constrained pragmatic account?

It is time to see how the pragmatic account of causation fits into the general intellectual terrain surrounding theories of explanation. It falls, I think, between two broad schools: the logical or epistemic on the one hand, and the purely causal and naturalized on the other.

Logical accounts of explanation—such as the D-N account discussed in ch. 1—concentrate on logical relations between propositions. Epistemic

theories of explanation may keep some of this component, but they are also concerned with the interest relativity of various kinds of logical or probabilistic components. Extreme cases of epistemic accounts are concerned with the psychological *effect* certain information has. Accounts which stress the rôle of understanding in explanation are typical of this type. Pragmatic accounts of explanation—such as van Fraassen's [van Fraassen 1980] or Garfinkel [Garfinkel 1981] rightly stress the centrality to the notion of explanation of interest dependence in eliminating from consideration vast areas of information about explananda, but do not have constraints which prevent bad explanations which meet the pragmatic criteria from being allowed.

On the other hand, there are those accounts of explanation which take explanation to be the uncovering of the causal structure of the world. David Lewis [Lewis 1986d] takes explanation to be just the giving of information about causes; everything else is incidental. Perhaps there might be criteria for explanations being good or bad, such as giving not enough information, or too much information, or providing information of a certain kind—but this is all true of description as much as explanation. What is distinctive about explanation is *solely* that it provides information about causes. Wesley Salmon in these latter days, [Salmon 1984] [Salmon 1985] argues that explanation is the process of laying bare the causal structure of the world. He even is tempted to (or does—this is not always clear) abandon events as causal *relata*, and he certainly does not see event explanation as central to an account of explanation. Explanation becomes programmatic: in fact explaining becomes more or less the process of doing science. To explain a particular event we have to lay bare the causal structures that surround it; and the more thoroughly we have done this, the better we have explained it.

My account rests between these two broad schools in this way. It is concerned with the causal why, the why of being concerned with how something came to be the case. In order to exclude crazy explanations that would have to be allowed if an unconstrained pragmatic account of explanation were adopted, I have constrained explanation with a requirement of causal relevance. But it does not consist just in the giving of *any* causal information; nor is it identified with the scientific program of describing more and more of the causal structure of the universe that surrounds the explanandum.

Explanation is the pouring of information about causal structures through a pragmatic net. It is concerned with *particular* pieces of causal information. To reply to a question about why there is a war memorial in almost every small Australian town with a litany that begins with a description of the big bang is not to provide a poor explanation; ignoring the pragmatics of such a question is to provide *no* explanation. On the other hand providing information which renders the event probable to an individual, or gives them some sense of understanding, while it might satisfy the individual more than the cosmological account neglects, *inter alia*, the possibility that these might be deeply inappropriate psychological responses. And an account of what makes them appropriate will bring in a causal requirement.

So explanations arise in various contexts. What is required for explanation in any contexts varies; this much is right about the Pragmatic account. Information which is irrelevant to these interests is no explanation; merely description. On the other hand, in order to be explanatorily relevant, it had better be causally relevant information. Information can meet the interests while not being explanatory because it just isn't that it is causally relevant. This is why it is right to view my account as a causally constrained

pragmatic account, rather than a theory of causal explanation with a pragmatics to match. It is not as though in the seeking of particular explanations we need a pragmatics to choose among the many explanations, each of which is properly an explanation, the one we are interested in. Rather, the pragmatics determine which causally relevant information is also *explanatorily relevant*.

We are left with the task of seeing how causal relevance as I have defined it here is related to actual causes, and what the relationship is between information about actual causes and about counterfactual causes; and thus how there can be explanations at various levels of nature, even if we do not think that there are causes at these levels. That is the task of the next chapter.

Chapter Four

Levels and Explanation

Table of Contents

1	Introduction	105
	1.0.1 Preliminary Notions.....	105
	1.1 The Return of the Third Constraint.....	106
2	Levels of Description	108
	2.0.1 Ordering of Levels of Description.....	109
	2.1 Events and Physical Objects.....	110
	2.2 Gerrymandered Levels of Description	113
3	Structural and Functional Levels of Description.....	114
	3.1 Structural States.....	116
	3.1.1 New Criteria of the Physical	117
	3.2 A Second Kind of Function/Structure Distinction	120
	3.2.1 Richardson's Reductionism and Structural Realization	123
	3.2.2 Structural Realizers and Type-Type Identities.....	125
	3.2.2.1 Weak Type-Type Identity.....	125
	3.2.2.2 Strong Type-Type Identity.....	127
	3.3 Strongly Functional States	128
	3.4 Explanation at Higher Levels.....	131
4	Program and Process Explanations.....	132
	4.1 Is There Really A Program/Process Distinction?.....	134
	4.2 The Weak Objection to the Program/Process Distinction.....	135
	4.3 Program Explanation and Choice Classes.....	136
5	Causally Efficacious Properties.....	141
	5.1 Why Does The Program Process Distinction Look so Plausible?	145
	5.2 Causal Efficacy at Higher Levels.....	146
6	The Constraints Reformulated.....	149
	Summary of Part One.....	153

1 Introduction

This chapter adds the last ingredient to the account of explanation on offer; that explanations might offer *counterfactual* information about causes, as well as actual information about causes. The motivation for this is to provide a satisfying account of the explanation of multiply realizable events at higher levels. Roughly, the line will be that in these cases the explanation should have more to say than adverting to the actual causes, since a story solely about the actual causes of the *realizations* of higher level events neglects the generalizations which are the motivation for having higher level events in the first place; namely that there are counterfactual states of affairs that, had they occurred, would have realized the very same events.

I start by offering an account of events and objects at different levels, so as to be able to set up the problem of how to explain events at higher levels (or to explain *with* such events). I also provide a criticism of the traditional function/structure distinction, and introduce special notions of structural and strongly functional levels of description, which will play a rôle in later chapters. Only then do I approach directly the question of explanation at these levels.

1.0.1 Preliminary Notions

Explanations are regularly offered at all sorts of *levels of description*; we offer, *inter alia*, physical, chemical, biological and psychological explanations—and it seems as if, at least in principle, it might be possible to offer these different kinds of explanations of the very same phenomena.

It is not hard to see how explanations at these levels meet the criterion of causal relevance outlined in ch. 3. Some chemical event, for example, is causally relevant to some other chemical event iff there is a world in which

the first event does not occur, and nor does the second, which is more similar to the actual world than any world in which the first event does not occur but the second does. My adding copper sulphate solution to the sodium hydroxide solution is causally relevant to the formation of a precipitate, because in the nearest world in which I do not add the copper sulphate solution, no precipitate is formed. For that matter, straightforwardly disjunctive events can be causally relevant in this way; if we allow my adding copper sulphate *or* aluminium sulphate solution to the sodium hydroxide to count as an event, then that event will be relevant in the same way.

1.1 The Return of the Third Constraint

What is more difficult is to see how this meshes with the *third* constraint on selection of the explanandum by the explanans—the requirement that, somehow, real causal relations be involved in the explanation. This is not only a problem for someone who believes that causation occurs at some specified deep microstructural level; even if you believe that macroscopic events or objects causally interact at that level (gravitational interactions between massive bodies for example), this constraint seems to be exhausted by the *actual* processes. Surely, at whatever level causal processes take place, in the example above about the event of adding copper sulphate or aluminium sulphate, the causal interactions are exhausted by whatever actually takes place between the chemical events which *are* involved, and the reformulation of the explanation with a disjunctive explanans seems to add nothing, and to take away something, from our *causal* information. And similarly with the domains of organization; if you thought that the *causal* interactions in, for example, brains were exhausted by (at most) biochemistry, then it would be hard to justify the explanatory rôle of psychology. It would pass the first two tests, but not, perhaps, the third. A

theory of explanation that had this feature would throw out most explanations as ultimately merely enthymematic; they would add nothing to explanations which cite the particular causes that are at work. This is especially bad news if you think that what justifies higher levels of description is that there are distinctive explanations to be had at such high levels.¹

This is hardly satisfactory. It is the task of this chapter to examine the notion of levels of description and organization, give some account of functional and structural properties at levels of descriptions, and to formulate the third constraint on selection so that higher-level properties and events can indeed be used in explanation. I begin with a discussion of levels of description, and develop an account of object supervenience and event supervenience based on it. I then discuss kinds of levels of description, and how to distinguish the scientifically interesting ones from the arbitrary and gerrymandered ones. This leads to a discussion of the function-structure distinction, and an account of functional and structural levels as the interesting ones. Then, using Jackson and Pettit's notion of *program explanation*, I show how the third constraint can be formulated so as to allow explanation at these higher levels.

2 Levels of Description

The most neutral way to understand the talk about levels of organization is to understand them as levels of *description*. The idea is simple; there are many ways in which the world can be described. We can describe it at the level of particle physics, we can describe it at the level of physical chemistry,

¹ It is considerations like this which, I think, motivate the insistence of many on the *causal* efficacy of high-level states. See Ch. 7 on Fodor and the Language of Thought.

we can describe it at the level of biology, and so on. And, in some cases, these descriptions can seem to be descriptions of the same *region* of the world. A comprehensive description of a plant in the language of particle physics, while boring and just about impossible to produce, is nevertheless a description of the same region of the world as a description of it in the characteristic vocabulary of biology. We even have intuitions about how these descriptions are related; perhaps you might think that a description in terms of particle physics which is complete may tell you all there is to know about the actual intrinsic properties of the region, whereas the probably much more useful description in biological language, no matter how complete, leaves some things unsaid—since perhaps not everything about the plant is relevant to it biologically

We need, then, some account of what levels of description are, and how they are related to each other. I take a *level of description* of a world (or region of a world) to be a characteristic vocabulary. So the biological level of description is the vocabulary characteristic to biology, the physical level of description the vocabulary characteristic of physics and so on. I do not rule out contrived or gerrymandered levels of description; if one puts together a vocabulary which includes terms like Macinprof (there is a Macinprof whenever there is either a Macintosh computer or a Professor of Philosophy) then it too constitutes a level of description.

2.0.1 Ordering of Levels of Description

Roughly, the idea behind ordering of levels of description is that they supervene on each other: if one level of description is at a higher level than another, then there should be no change in what descriptions are true of a world at the first level without a change in what is true of the world at the other, and there should be some changes possible at the second level

without change at the first level. The first clause corresponds to the formulation of supervenience as 'no change at the supervenient level without change at the subvenient level'² and the second clause to a *multiple realization* constraint. Unless there is more than one way for a subvening level to realize the supervening level, then the supervenience relation in such a case will be symmetrical, in which case there is no motivation for regarding the putative levels as distinct.

I shall call a class of all worlds at which the same descriptions³ at a level L is true an L -equivalence ($L \equiv$) class. Now suppose we are interested in whether L is a higher level of description than some level L^* . Consider each of the worlds w in an $L \equiv$ class. Take the descriptions at level L^* which are true at each of these worlds, and consider, for each of these worlds, the classes of worlds at which the same descriptions at level L^* are true (call these the $L^* \equiv$ classes of the $L \equiv$ class)—there will be as many of these as there are worlds in the $L \equiv$ class. Call the union of these classes Λ . Now, a level of description L will be *at least as high a level as* L^* iff, for every $L \equiv$ class E , any world w which is not in E is also not in Λ . L will be *strictly at a higher level than* L^* iff, in addition, there are at least two distinct $L^* \equiv$ classes of E . L

²See, e.g., [Kim 1984].

³This is formulated in terms of descriptions to make the weakest kind of assumptions about levels of organization: most actual science involves the production of theories which describe the world, and an account of levels in terms of the truth of those descriptions should be adequate. But I mean to be as ecumenical as possible; an account of the levels of organization in nature as they actually are, whether actually described or not, can be had by substituting realized properties for true descriptions in the definition above. Indeed I take it that the properties that a world realizes are the truth makers of the descriptions.

will be at the *same level* as L^* iff it is at least as high level, but not strictly at a higher level.⁴

2.1 Events and Physical Objects

I have talked of levels of description because this is as neutral as possible about what ontological commitments are being made. I hope that there is nothing contentious in an analysis of levels of description.⁵ My account of explanation, however, involves events, and events on some accounts are

⁴This leaves levels of description partially ordered. They could be totally ordered by stipulating that if some total description D^* at L neither supervenes on, nor is supervened on by, some total description D^* at L' , then L and L' are in fact the same level. There are problems with this. Consider cases where there are different levels of description which have different domains, such as the plasma theoretic level of description and the psychological. Presumably neither of these levels supervenes on the other one, though they both may supervene on some ultimate microstructural level(s) of description. This extra well-ordering constraint would put psychology and plasma physics at the *same level of description*; a consequence which is wildly counterintuitive to say the least. More plausibly, you *might* want to build in a well-orderedness between descriptions of the same domain (where a domain is, perhaps, a spatio-temporal region of the world) so that there could be no pair of true descriptions of that domain where one is not either at the same level, at a higher level, or at a lower level than the other. I take this to be equivalent to ruling out some very strong sense of their being incommensurable conceptual schemes.

⁵One candidate for something contentious: in this chapter and elsewhere I have assumed that there is available to us some notion of trans-world *identity*. The very same thing can exist in two worlds. I do this because the most direct understanding of the notion of multiple realizability of, for example, events, is that the very same event could occur in different worlds and yet be realized differently. I do not think that this is a crucial feature of my account; for my purposes those who are sceptical about transworld identity are welcome to reformulate what I say in a counterpart-theoretic way. I *think* this can be done with no loss.

property-instances—properties of physical objects.⁶ So a definition of what it means for events and physical objects to be at levels of description is called for. I propose the following.

An object, O (or an event, E), is at a level, L, iff there is in the characteristic vocabulary of L a description dO which describes O as existing (or a description dE which describes E as occurring) and for every level of description L' lower than L, there is no non-disjunctive description dO' (or dE') such that every world which satisfies dO' (or dE') also satisfies dO (or dE).⁷

In what follows I shall say that when a certain description is true, then a *property is realized*. When a true description of a world is a description of it as containing an object or event, then I shall say that that object (or event) is realized in that world. Where some object or event at a level supervenes on some properties at a lower level which are realized, I shall say that it is *realized by* the realizations of the property at the lower level. Where the realized properties at the lower level are those of an object or event existing or occurring, then I shall say that the object or event at the higher level is

⁶As in, for example, Kim 1976. For more on events, see ch. 5.

⁷The second clause is required because there is nothing in the definition I have given of levels of description to prevent a level L which is higher than L* containing part of the vocabulary of L*. The second clause specifies that if O is at L, then dO should be in that part of the L vocabulary which is multiply realizable at a lower level. This captures the notion that it is what is *distinctive* about the vocabulary that counts. There are terms in biochemistry that are also in physical chemistry. This definition allows that they remain *terms* in biochemistry, but does not allow that events or objects picked out by them are events or objects at that level.

realized by the object or event at the second level.⁸ This allows different objects to occupy the same spatio-temporal region—perhaps a cloud of gas is not the same object in this sense as the collection of gas molecules that realize it. The cloud of gas could have had its molecules arranged somewhat differently and remain the same cloud, whereas the collection of molecules, perhaps, could not have done so.

2.2 Gerrymandered Levels of Description

The definition of levels of description which I give above allows any true description to have a level—even useless or gerrymandered ones—so there could be levels of description consisting entirely of arbitrary and unnatural

⁸This formulation does not exclude the possibility that an object could be realized by an event; which is surely a bizarre possibility. It does not exclude them because in this formulation I do not make any distinction between objects and events, since it is unnecessary to do so at this stage. Objects and events are both instantiations of properties of regions (four dimensionally) and this much is all that I require here. Of course they are instantiations of rather different sorts of properties. Objects are, on the usual account, temporally extended, and in order to capture usual folk intuitions, spatio-temporally continuous. Perhaps we may even have to limit ourselves to certain kinds of *natural* properties (see [Lewis 1983] for two ways to get a notion of naturalness of properties). Events, on the other hand, are not usually taken to be temporally extended. On Kim's account [Kim 1973, 1979], for example, they are properties of objects at times (though I have doubts about the event-process distinction; most of what we take pre-theoretically to be paradigm events—races, demonstrations, explosions and so on—are in fact processes on an account like Kim's). Thus, if objects are realizations of properties of regions, events might seem to be realizations of *second-order* properties of regions. This baroque doctrine may not be necessary, however. If an event is realized by an object's being thus-and-so at *t*, surely there is a straightforward first-order property which a region must realize for an object to be thus-and-so at *t*.

descriptions.⁹ The description of Australia as having a grain of sand of mass X at some grid reference and 1000 litres of water in a tank just south of Tuross, is of a higher level than the description of Australia as having a certain molecular configuration at these places. There are lots of molecular configurations that would make it true that there was a grain of sand of the right mass and a tank containing the right volume of water. In realizability talk, the property of having the grain of sand and the water in the tank is *multiply realizable*; there could be other ways for Australia to have the grain of sand and the water in the tank. The tank could be made of titanium; the sand could have a different surface geometry.

This is all very well as raw ontology; it should not concern us if some object or event whose identity conditions are unmotivatedly disjunctive and which is realized by, say, a hugely distributed and non-continuous region of the world, merely exists or occurs. Consequently it should not worry us if there are hugely many levels of description brimming with such descriptions. But surely events or objects at these levels should play no part in good explanations.

One approach would be to say that what is required is a notion of natural properties.¹⁰ Natural levels of description could be the ones which describe the world as realizing only natural properties. This is all very well, and Lewis in [Lewis 1983] gives us a mechanism for adding naturalness to an account of properties. But it is, of course, not enough; it is one thing to have a conceptual apparatus to accommodate naturalness of properties, and

⁹It does not even exclude so-called unprojectible predicates (predicates which it is not rational to perform inductions on) such as 'grue'.

¹⁰Or in the more parsimonious vocabulary, natural descriptions.

hence levels, and quite another to attempt to have some principled way to pick out the natural properties. It is not even absolutely clear what the pre-theoretic intuitions are about naturalness. Are artifactual properties natural? Presumably so in some sense, but if so a lot of very gerrymandered properties will slip in.

I am not clear as to what an exhaustive characterization of the levels of description which are explanatory should look like. There is, however, a widespread practice of explaining at *functional* and *structural* levels of description. These are often at sufficiently high levels to make it unclear how explanations at these levels obey the Third Constraint, since the properties they involve are not always obviously causal ones. I shall be particularly concerned with functional and structural levels of description.

3 Structural and Functional Levels of Description

The function/structure distinction makes its first significant modern appearance in [Putnam 1975a] and [Putnam 1975b]. This distinction first appears in [Putnam 1975a] as a distinction between *logical* states and structural states; and with it appear a number of confusions which have survived in the field to the present day. In § 3 of [Putnam 1975a] Putnam talks of both logical *states* and logical *descriptions*: and does not appear to distinguish between them, and so the ontological status of the states is very unclear. At times it seems as though logical states are things in themselves that have existence even prior to being 'realized' by a structural state; at other times it seems as though they are just descriptions, in which case realization is just bringing it about that there is something of which that description is true. It can even seem as though there can be descriptions of logical states, though. In which case it would seem that there *is* a

state/description distinction. In any case, the gist of the structural/logical distinction comes in this famous passage:

In Particular, the 'logical description' of a Turing machine does not include any specification of the *physical nature* of these 'states'—or indeed, of the physical nature of the whole machine. (Shall it consist of electronic relays, of cardboard, of human clerks sitting at desks, or what?) In other words, a given 'Turing Machine' is an *abstract* machine which may be physically realized in an almost infinite number of different ways.

As soon as a Turing Machine is realized, however, something interesting happens. Although the machine has from the logician's point of view only the states A, B, C etc., it has from the engineer's point of view an almost infinite number of additional 'states' (though not in the same sense of 'state'—we shall call these *structural* states). [Putnam 1975a p. 371]

In [Putnam 1975b], the logical states get called *functional* states. In the 1960 paper [1975a] he emphasizes that it is *abstractness* which characterizes functional states. A functional state of a Turing machine is just its being organized in a certain way; the structural state is what *physically realizes* the functional state. So, in his example, the *program* of a Turing machine might be its functional state, and a configuration of vacuum tubes, for example, might count as a structural state which could be the physical realization of the functional state.

The two central distinctions between functional and structural states, then, are meant to be the abstract/physical distinction, and the multiple realizability/non-multiple realizability distinction. A functional state is abstract and is supposed to be able to be realized in any of a number of ways; the states which realize them are the structural states, which are token physical states of concrete entities. So, using the analogy with computers that pervades the field, the functional state of the computer might be realized by an array of vacuum tubes (or these days, VLSICs). But it need

not be realized by it; some other array of tubes would do, as indeed (at least in theory) would the array of clerks.

The problem with this distinction is that if multiple realizability is what distinguishes functional states from structural states, then the structural states so-called had better not be themselves multiply realizable; but this is not at all clear in these and other examples of structural and functional states. An array of vacuum tubes can, in some sense, be realized by different token vacuum tubes: it remains the same array if you replace the tube. Vacuum tubes themselves are presumably multiply realizable: there is more than one way to build a thermionic valve. Similarly for transistors, and perhaps also for neurons. And if this is so, then the putative structural states are just as abstract as the functional state they realize. If there is a distinction here it is between a *realizer* state and a *realized* state; but these are not *intrinsic* properties of states, rather they are relational properties. A realizer state realizes the realized state (I shall sometimes call realized states *rôle* states. The same state (or description) may realize one state while being realized by another. The word processing program I am using is realized by the machine's architectural structure, which is realized by the CPU, which is realized by a structure of conductors, insulators and transistors, which are realized by particular metals and plastics and so on. So the realizer/realized distinction is one which points us to a *relation* which may hold between states. I take this kind of talk about the function/structure distinction as being about nothing more or less than levels of description couched in language of states.

3.1 Structural States

My dismissal of the distinction between functional and structural states does seem in some ways a little unsatisfactory, however. Even if all you can get

from the multiple realizability story is a hierarchy of levels, there do seem to be intuitions at work in the talk of structural states which demand more. I think there are at least two. The first is that what realizes functional states must, if physicalism is to be preserved, be a physical state; and if structural states are what realize functional states, then they must be physical and particular in a way that the more abstract functional states are not. The second is that there should be some difference between *merely* functional states—ones which specify in terms of inputs and outputs *what* it is that a system does, and states which specify *how* it is that a system has those properties in terms of the way it is structured. I shall deal with question of the physicality of realizations first.

3.1.1 New Criteria of the Physical

Sometimes we are concerned not with *how* some functional state supervenes on another, but rather in some sense with *what* actual thing they are states of. The idea is that there is surely something which is an unrepeatable particular, which is the thing in the actual world which realizes the state. This concern is evident in the fact that both Putnam [Putnam 1975a] and Fodor [Fodor 1968 onwards] use the term '*physical state*' to mean the same thing as 'structural state'. There is something disquietening in this talk of levels of description all the way down. Surely there *is* something which does the realizing, and in this sense whatever it is could not be multiply instantiated since it is a concrete particular.

A tempting suggestion is that this could be a concern about what *physical object* actually realizes some functional state. This option is not really available, however, given my account of objects at levels, for objects on this account are multiply realizable; the same mind, for example, could be realized by many brains, or the same house by many collections of bricks

and mortar. Indeed there is nothing in this account of objects which requires that an object supervenes locally on the region it seems to be in. So, for example, a coin might be taken to supervene both on local properties—such as its mass and topological structure—but also on other properties, such as the fiscal and legal institutions of the times.

One possibility is to make physical objects a special case of objects. They are the objects which supervene only on local properties (and any other qualifications called for by one's preferred account of objects—see footnote 8). So corresponding to a coin, there might be a lump of shaped metal which is a *physical* object. The price of this distinction is that the physical object does not wholly realize the coin, if supervenience is required for realization, since there *could* be changes in the coin without changes in the physical object. It could, for example, cease to be a coin because of changes in the fiscal system.¹¹

I think this distinction is a useful one, but it still fails to capture the intuitions behind some strong notion of physical realization, because there is nothing stopping physical objects from being multiply realizable. Perhaps, depending on particular views about essential properties, a given lump of metal could be realized by different microstructural states.

A better suggestion is that it is a region of a world which realizes most of these states.¹² The identity of a region is given by *all* its properties, at

¹¹Though perhaps it remains a coin in another sense; people collect Roman coins after all.

¹²A strong essentialism about physical objects could be captured by insisting that physical objects just are regions of worlds, particular and non-multiply instantiable. I leave this as a move to be made by those of a Kripkean (or stronger) bent.

whatever level, including the property of being in a particular world. This of course leaves it acutely context sensitive. Any change in the environment will locate it in a different world, but that is the price of saving the intuition of actuality: the intuition that it is something *actual* which realizes the states. I shall say that a region of a world is the *strong physical realizer* of a property. Sometimes a weaker notion of physical realization is called for. When, for example, we claim that a mental state is realized by some physical state, we may want to claim more than that it is realized by a neurophysiological state (because that would be multiply realizable by different chemical states, and we are concerned with anything physical) but less than that it is realized by a region (because we may want our physical state to count as the same physical state in neighbouring worlds: surely whether a sparrow drops from Heaven should not affect what physical state I am in). To capture this notion we can introduce the notion of intrinsic similarity at *all* levels of description lower than the one which is being examined. The "physical realization" of something in this weaker sense just is intrinsically similar at *all* the lower descriptive levels.¹³ Of course neither the strong sense nor the weak sense is

¹³There is a middle road here; that would be a notion of physical realization which invoked all the realized intrinsic properties. This would be weaker than strong physical realization as it would not be context sensitive in the same way. But it would be stronger than weak physical realization, since it would not just involve intrinsic similarity at every level, but involve every intrinsic property at every level. This is perhaps too strong for some purposes. Suppose that the state of storing the number 37 in a pocket calculator is a functional state, and we are concerned with the physical state which realizes it. It seems clear that if the state were completely divergent at any level it could not count as the same physical state; it could not, for example have been made of anti-matter. On the other hand, *some* divergence at *every* level may be acceptable—especially in the categorization of physical state types, but also perhaps in the formulation of identity conditions for physical state tokens—just so long as the scientifically interesting properties at every level remain the same.

often *needed*, most talk of how something is realized is in terms of some set of relevantly interesting lower level properties which are either jointly or severally sufficient for having the higher level property. Which lower level properties are relevantly interesting will depend, naturally, on the explanatory interests of the inquirer. For example, a computer repair engineer whose job is plugging in replacement boards is concerned with the machine's instantiation as a set of interconnected boards; the board designer who looks at faulty boards which have been replaced to see how hardware problems might be avoided in the future will be concerned with the actual details of how boards are instantiated, and so on. A heart surgeon who plans to remove a patient's faulty heart and replace it with a prosthetic one may well not be concerned with which of the different models of prosthetic heart (they operate differently, say) she will use, but another surgeon who plans to *repair* the heart will certainly be concerned with lower level properties. Not all such properties, though. If the repair is effected by inserting a piece of plastic tubing, it may not matter that it does not share many lower level properties with the original pieces of bodily plumbing. In any event, for those occasions on which it is required, we have a notion of structural state which, while different from Putnam's or Fodor's, captures some of the intuitions which motivate it. A structural state in this sense is a state which can be either the weak or the strong physical realizer of another state.

3.2 A Second Kind of Function/Structure Distinction

So much for the sense of structural realization where what is meant is *physical* realization. There is a second sense of structural realization of functional properties which is of interest. Henceforth I shall use strong or weak physical realization to refer to the first senses, and structural realization to refer to this second sense. This is where we want to claim that

there is a scientifically respectable description of whatever has the functional property (specified, say, by inputs and outputs), at a *lower* level of description than the one at which the functional property is located, which *accounts for* its having those properties.

Consider the case of bushes which are trimmed so as to look like giraffes. The giraffe-likeness supervenes on the structure of the twigs, and there may be an explanation in terms of the design of the gardeners. But there is no account of giraffe-likeness in general in terms of the structural organization of the twigs, except at the very highest level which corresponds to the very surface geometry we are trying to explain. The lower level structure of the twigs in giraffe-bushes varies considerably. This is the kind of case where we may want to say that the functional property is, in the case of each giraffe-bush, physically realized, but that it is not *structurally realized*.¹⁴

To make this claim about the giraffe-bushes we need further constraints on what a structural realizer could be. If there are no constraints whatsoever, then there will be a hugely disjunctive account in terms of possible twig-configurations that might be taken, in the case of bushes, to be a structural realizer of being giraffe-shaped. I take it that the right constraints are whatever the best sciences tell us they are. If hugely disjunctive twig-configurations that realize giraffes appear in the best

¹⁴This is not to deny that there is something in common between each giraffe-bush; namely that it is trimmed to look like a giraffe. Nor is it to deny that this may be in ordinary parlance a structural feature of the bushes. It is, rather, to deny that the functional (or structural in the ordinary sense) feature of the bushes is realized *at a lower level* by another property featured in a good science at that level. In other words, being giraffe shaped is not a botanical property, or a chemical property, or

botany, then so be it—but I doubt it. Still less likely is that it will turn out that some notion will have a rôle in the best physics that would realize these states.¹⁵

On the other hand, sometimes there are lower level properties at an interesting scientific level of description which account for how, in a particular case, a functional property is realized. Take, for example, the functional term 'heart': functional hearts are whatever ensures that blood moves around bodies.¹⁶ Now functional hearts could be multiply realized; electric pumps can count as hearts or even some distributed fact about the muscles around veins and arteries in an odd species which might display, by analogy with a certain doctrine in psychology, a property I shall call cardiovascular equipotentiality.

As a matter of contingent fact, however, there is a science of cardiology. Mammal hearts, at least, have various scientifically interesting features in common. They have valves, aortae and so on and are made of

¹⁵Of course it is not even clear whether *giraffes* have a structural property in this sense which realizes the functional property of being giraffe-shaped. Some genetic property is probably the best candidate, but there are two problems with this. First, while genetic properties are involved in the etiology of the shape, this does not mean that they synchronically realize the shape structurally. Second, it is a preformationist mistake to suppose that a genetic property *wholly* determines the shape. In fact a complex disjunction of genetic and environmental factors (food supply and so on) does. Whether these sorts of complex disjunctions form biological kinds is an interesting question for the philosophy of biology.

¹⁶Of course, as with many functional terms, what falls under its ambit depends on how the functional term is specified. There could be many different functional accounts of hearts depending on, for example, what is meant by 'blood'. Must it be actual mammal blood? Or will a specification of it as an oxygen bearing liquid do? Or a liquid bearing substances needed for life? (in which case sap counts as blood).

muscle tissue configured in a certain way. In fact, most of the predictions and explanations we need to make about hearts can be made by reference to cardiological features, though this would not be so if there *were* in fact other realizations around. For that matter, cardiological hearts could in principle be realized in a way that did not realize functional hearts. Suppose that there were a species of ape that had two hearts; one was responsible for the circulatory system, and the other vestigial—it just moved some useless fluid in a closed loop. We might describe this as a case where the ape has two *cardiological* hearts, but only one *functional* heart.

The fact that there is an independent, scientifically interesting science—cardiology—which enables us to make all the predictions we need to make about particular realizations of functional hearts, though, is what I mean by my second sense of structural realization. A property is structurally realized just if there is a lower scientifically interesting description which, for the particular case, is sufficient to guarantee that the specification at the higher level will be realized.

There are two positions in the literature on which I think this notion of structural realization has bearing: Robert Richardson's reductionism about functional characterizations [Richardson 1979]; and Frank Jackson, Robert Pargetter and Elizabeth Prior Johnson's notion of type-type identities within functionalism [Jackson, Pargetter and Prior 1982].

3.2.1 Richardson's Reductionism and Structural Realization

Richardson is arguing for the reducibility in principle of psychology to some lower level science. Roughly, his claim is that multiple realizability is not an impediment to reduction, since there are paradigm cases of multiply realizable (and in fact multiply realized) functional characterizations which are nevertheless reducible to lower levels.

He takes genetics to be such a case, and in doing so follows such classic statements of functionalism as [Wimsatt 1976 p. 222]. The claim is that the functional properties of Mendelian genetics could in principle be realized—say by chemical properties, or electro-mechanical properties if that was how humans functioned—and in fact *are* multiply realized in molecular genetics. Genes, on this account, are realized at some level by combinations of codons [Richardson 1979 p. 545], and different combinations of codons can in fact realize the same gene. If we are to take it that the reduction of Mendelian genetics to molecular chemistry is indeed one of the great achievements of science, then in order to be consistent with the data we will have to deny that multiple realization of one level of characterization by others, precludes the reduction of one level of characterization to another.

I think that my second sense of structural realization can preserve this criterion for reduction, while preserving much of the content of the scientific intuitions about the genetic case. Presumably what must at most costs be preserved about our understanding of genetics is that it is a major and important scientific discovery that, as a matter of fact, Mendelian genetics is realized by molecular chemistry. Further, molecular chemistry is a science; the laws of molecular chemistry are sufficient to account for the characteristics displayed by actual genes, and anything which shared our molecular chemistry would, *ceteris paribus*, share our genes. So, on the second criterion of structural realization above, molecular chemical states *structurally realize* Mendelian genetic states. Perhaps this scientific discovery is enough as a claim for the importance of the work done on molecular chemistry, without the further claim of a reduction of functional genetics to molecular genetics. What perhaps makes the reductive claim plausible is that you could construe what has happened as a reduction of a sub-domain of genetics—genetics as actually realized by our biosphere—to molecular

chemistry.¹⁷ But the discovery of such a sub-domain reduction just is the discovery of a pervasive system of structural realizers.

3.2.2 Structural Realizers and Type-Type Identities

Similar remarks apply to Jackson, Pargetter and Prior's defence of the possibility of type-type identities within functionalism. They argue that the right way to construe functionalism is compatible with type-type identity claims. So, for example, it might turn out that

...a (kind of) mental state is a (kind of) brain state, in particular the kind of brain state which realizes for the organism at the time the functional rôle definitive of being in that mental state. [Jackson, Pargetter and Prior 1982 p. 212]

Let us use the example of pain. It turns out that their type-type identity theory really involves two claims. The first is that the type of pain-for-an-organism (say, pain for humans) might be type identical with a certain type of brain state (say, the hoary old *c*-fibres firing). The second is that the type of pain in general is identical with the property of having some state which plays the pain rôle for that organism at that time.

3.2.2.1 Weak Type-Type Identity

¹⁷I do not discuss here the additional complication raised by the claim that there is in *in fact* multiple realization of genetics in molecular chemistry. In the first case I am not sure that this is so; genes may be multiply realized as codons, but not, according to Richardson, as *cistrons*. It is cistrons which are multiply realizable as codons. Now if cistrons are a natural kind of molecular chemistry, then genes are not as a matter of fact multiply realized at that molecular chemical level, even if those molecular chemical realizers are themselves in fact multiply realized. If it *were* to turn out that there are no regularities at the molecular chemical level which governed the realization of genes and ensured that there were molecular chemical kinds which non-multiply realized genes, then there would be a less convincing case for the discovery of a structural realizer in my second sense.

The second of these claims I do not find so illuminating, though in one sense there is no reason to dispute it. If one is a functionalist about pain, then of course the type of pain is the property of having some state which plays the pain rôle for that organism at that time, given some functional specification of how to tell that that state is one which plays the pain rôle for the organism at the time. I call this weak type-type identity. But is this what Functionalism's rejection of type-type identity theories was concerned about? Take, for example, Dan Dennett's forthright claim in his survey article [Dennett 1978]:

Perhaps no one today supposes that *types* of mental items can be distinguished directly by purely physical features, but almost no-one any longer supposes this was a reasonable goal of physicalism. [Dennett 1978 p.254]

Jackson, Pargetter and Prior show that functionalism picks out a type, all of the tokens of which are physical tokens, but is this what is meant by *distinguishing the type directly by purely physical features*, (or, say, by Davidson's search for (and rejection of) psycho-physical laws in [Davidson 1980c])? If a physical type is a type all of the tokens of which are physical, then we can have a type-type identity theory. But the requirement that we can distinguish the property by purely physical properties seems to demand more. Certainly if we had an exhaustive list of the physical features which 'played the pain rôle for an organism at a time', then given a physical feature of an organism we would be able to determine whether it was in pain (though even then I am not sure if this would be to do so *directly* in Dennett's sense). But we do not have such a list; the type is distinguished by its *functional* features. The type-type identity theory for pain in general is just the self identity of the functional type with the functional type, since the second order property of having whatever physical state plays the pain rôle for the organism at the time is itself a functional property. The demand

which in 1978 'almost no-one' believed could be met—identifying a functional type with a physical type, where a physical type is one whose tokens are picked out by properties *natural to the relevant physical sciences*, be they biology, neurology or whatever—can indeed not be met.

3.2.2.2 Strong Type-Type Identity

The *first* claim is rather more interesting, however. There is much more reason to hope that, relative to particular organisms and canonical environments, there could be some scientific account at a lower level than that of the functional specifications which will pick out the states which are pain states. Indeed there does seem to be at least a primitive neurophysiology of pain; and some of the proto-theories of it allow some predictions to be made about the effects of various neural disorders on pain. So, perhaps there may one day be a neurophysiological (or neurofunctional) theory of pain for humans which will allow us to individuate pain states in humans by something other than just the high level functional or behavioural criteria. Such a theory *would* be a type-type identity theory of the type Dennett was so pessimistic about in 1978, but only for the sub-domain of pain-in-humans. It would, in my terminology, be the discovery of what the actual *structural realization* of pain¹⁸ was—and I think this would be a better term for it, since even if there were no neuro-physiology of pain, there would, of course, *still* be a type-type identity theory of the weak kind discussed above.

¹⁸If one had Kripkean intuitions about the rigidity of reference of terms which seem to be functional terms [Kripke 1980], or for which there are as yet only functional criteria extant, this could be captured by the linguistic doctrine that the apparently functional term—'pain', for example—refers only to things which are alike at every structural level to the actual realizers of the functional term.

3.3 Strongly Functional States

So far, the criterion of being a functional state I have used is just being a multiply realizable state, with no further constraints. Whenever a functional state is realized, it is always physically realized, in all of the senses outlined above. It may or may not be structurally realized at some levels.

This amounts to an extreme liberalism about functional states, and as such accords with the early notion in Putnam. But often what is meant by a functional state, either implicitly or explicitly, has been stronger than this. Fodor, from [Fodor 1968] onwards, seems to mean by a functional state one which does actually describe how the system which has the state *actually functions*. A functional hypothesis about a system is a genuine speculation about how it works; not just an uninteresting high-level description of the state true just in virtue of a true but possibly gerrymandered or scientifically uninteresting set of facts about its input-output relations (in the psychological case, say, true behavioural level generalizations).

So on the weak notion every multiply realizable state is a functional state, including of course the structural states (which also may realize other functional states). But our problem for explanation is not very interesting in just the weak form. It is clearly a problem for a theory of explanation if structural states are not able to be involved in explanations, but what about the functional states which are not structurally realized? If the functional states which were not structurally realized were all *just* arbitrarily specified at a level of inputs and outputs, then there would be no problem for explanation since we might be happy if such states were *not* candidate explanantia.

So, to motivate the problem for explanation, we do need to be concerned with some stronger notion of functional state. I do not propose to

attempt a thorough account of what, exhaustively, such a stronger notion will require. There are, however, a number of *sufficient* criteria in the literature. These include criteria of design, evolutionary or other ætiological significance, and dispositional properties that may guarantee future selection.

It is more or less common ground that if something is in fact designed to serve a function, then under normal conditions it has that function, though this is not sufficient even to account for the case of artifacts—other selective processes may be involved [see Millikan 1984]. A good account of an evolutionary constraint being placed on high level states in order to make them strongly functional—roughly that functional states in this sense are states which are selected for—can be found in [Lycan 1980].¹⁹ The third kind of constraint on a stronger notion of function (a generalized version of which, if accepted, could probably subsume the first two) can be found in [Bigelow and Pargetter 1987]. This notion has the very considerable advantage that it allows functional states to be strongly functional at the time they are first realized, given an environment. In the absence of this possibility, we are committed to saying that, for example, when an advantageous mutation occurs, and the sport begins to spread, the

¹⁹An important distinction needs to be made here. Lycan introduces the evolutionary constraint in a way which makes him wonder how seriously to take the notion that nature is functional at every level; that each level of organization is multiply realizable. Now some constraint on what counts as a functional property is necessary if we are interested in problems of *explanation*: since explanations relate to parts of the world which have been carved out of it by our explanatory interests. But as a matter of raw ontology there is no problem with the rôle state-realizer state going down into the foundations of nature—or in other words, even if the *strong* functional states peter out at some point, the weak ones continue.

functional feature is not strongly functional in the first examples of the sport, but when it has been selected for and spread, that it then *is* functional, despite the fact that in some intuitively compelling sense of function, it is functioning in exactly the same way in the later individuals as it did in the first.²⁰

Whatever the exact criteria for strong functionality, or for being a scientifically interesting structural property are, it is easy to see that strongly functional and structural properties feature regularly in our explanations. We explain the behaviour of a chess playing program by reference to the functional property of being a chess player. We explain the behaviour of cold blooded animals in terms of the functional property of being exothermic and so on.

²⁰I do not think that this account subsumes an ætiological account, however. Rather I think that what is required is that if a property comes out as strongly functional on either account, then it is strongly functional (though perhaps there is room for their being different senses). This combination gives the right answers in the cases where we are asked to imagine a world in which a Universe otherwise like ours sprang into being *ex nihilo* as a stochastic aberration moments ago, in the same configuration that ours was in moments ago. On the purely ætiological account, *none* of the biological functions we imagine are strong functions would be so. On the purely dispositional account, this would be a world with *no* functional differences from ours. On the combination view, however, there would be some differences and many similarities. Any of the functional states which actually played an ætiological rôle but no longer have the dispositional force required in the real world would be strong functions in the real world, but not in the stochastic universe. On the other hand, the many strong functions which still have their dispositional powers would form the basis of similarity between the real world and the new-born one. This seems the right balance; such a bizarrely different world should have *some* differences in functions, and it is right that they should be the backward looking ones. On the other hand, Bigelow and Pargetter are surely right that there should be a great deal of similarity—and since it is in the future that this world is most similar to ours it is right that these should be the forward looking functions.

3.4 Explanation at Higher Levels

This concludes the discussion of levels. We now have an account of what it is for events and objects to be at higher or lower levels, and have some idea of what the more important levels—structural and strongly functional ones—are constituted by. So our problem is set up. We have multiply realizable higher level features of the world, and at least some of these—the structural and the strongly functional—should feature in our explanations. Yet in terms of the criteria in earlier chapters for selection by the explanans (rather than the rest of the choice class of explanans) of the explanandum over the rest of the choice class of explanandum, it is hard to see what explanations invoking these sorts of properties like this add—in the case of explaining particular events—to explanations in terms of the lower level realizations of these properties; for by the supervenience hypothesis having those realizer properties tells you everything you need to know about the intrinsic states. In any case, the explanation in terms of realizer states is likely to give you a closer grip on the actual causes involved. In particular it is much more clear how the third constraint, requiring real causal relations between the explanans and the explanandum, can be met by the *realization* of a higher-level event, than by the event itself. So we must seek for a reformulation of this constraint if we are to allow higher-level events to explain or be explained.

Consider an explanation of why a lizard is sitting in the sun, rather than in the shade. We have two suggestions; one is a (fairly) complete description of the lizard's workings at a biochemical level, together with the lizard's initial states and final state. The second explanation is that the lizard has the functional property of being exothermic, together with information about the temperatures in the shade, in the sun and the lizard's initial temperature.

The first explanation satisfies the three constraints of ch. 3., and it also seems to satisfy the third constraint rather more satisfactorily: it is at least closer to what the actual causal processes involved are. The second explanation certainly cannot provide *more* information about the actual causal processes, and plausibly it provides less. Is it just an enthymeme for the first explanation? Or is it just a point of pure pragmatics as to what is included in the choice class of explanans? A class which contains only functional properties, one of which is the property of being exothermic, will result in that property selecting the explanandum. A class which contains only biochemical or physical properties of the lizard will result in these sorts of properties doing the selecting.

Neither of these suggestions is quite satisfactory. If it is a matter of complete indifference as to which description we should go for, then at least in principle we should seek the most causally informative. Yet this goes against the scientific grain; there seem to be appropriate levels of explanation of phenomena. Some phenomena call for explanation in terms of physics or biochemistry. Others seem to demand explanation in terms of functional states. The next section shows how this could be so.

4 Program and Process Explanations

A useful way of seeing how it could be appropriate to provide explanation in terms of higher level properties is to approach the question via Frank Jackson and Philip Pettit's notion of *program explanation* [Jackson and Pettit 1988]. Their strategy is to show that there is a kind of explanation which provides something *extra* over and above what other kinds of explanation provide, and that this something extra is often what is called for in explanation requests. I think that the core of this idea is crucial for an account of explanation at higher levels. I have reservations, however, about

the program/process *distinction*. I will first give an account of the difficulties in the idea, and then try to extract from it what will be useful for my present purposes.

Jacson and Pettit call this kind of explanation *program* explanation, and contrast it with what they call *process* explanation. A process explanation is a kind of causal explanation where the actual cause is cited for the case in question. A program explanation cites a range of relevantly similar states of affairs, any of which, if actual, would have caused the explanandum event, and one of which in fact did cause it.

They give various examples of this distinction, such as trying to explain a conductor's annoyance. Two explanations are offered: someone coughed, and Fred coughed. Suppose that Fred in fact did cough, and that his coughing caused the conductor to be annoyed. The process explanation is supposed to be that Fred coughed. The explanation that *someone* coughed, however, is a program explanation. In virtue of the fact that someone coughed it follows that some causal story is instantiated which will in fact lead to the conductor's being annoyed. In addition, however, it provides *extra* information of explanatory significance. It tells us that it is not in fact explanatorily relevant (although it could be) that it was Fred who coughed. Any one else in the audience would have annoyed the conductor just as much. In so doing it draws our attention to an interesting generalization: anyone's coughing annoys the conductor. In virtue of this, we can put out leaflets pointing out that patrons must muffle their coughs without inquiring separately about each patron whether their cough is likely to annoy the conductor.

They provide numerous other examples of program explanations, but the essential features are here. The program explanation is not explanatorily

idle because, in virtue of the higher level property which is the program explanans (in this case the property of someone's coughing) there must be some lower level property which realizes it (in this case Fred's coughing, though we need not know what it is) and does the actual causing. Jackson and Pettit call this *programming* for the cause. The program explanation is not *redundant* because it tells us something extra that we may be interested in, namely that it does not matter for the conductor and hence for us who actually coughed; and thus it draws our attention to interesting generalizations.

4.1 Is There Really A Program/Process Distinction?

There are, then, two features which are supposed to be distinctive of program explanations. The first is that they are multiply instantiable; they advert to a description of the explanans which could be realized in some way(s) other than it is. The second is that the program explanans is not *causally efficacious* with respect to the explanandum, though one of its realizations must be. How does this distinguish a program explanation from a process explanation? Well, presumably a process explanation is *not* multiply instantiable, and the explanans itself is a property which is causally efficacious with respect to the explanandum.

Let us return again to the example of Fred coughing. The program explanation is that *someone* coughed, but in the case where, say, the conductor is especially annoyed by Fred's coughing, we may want to give a process explanation that *Fred* coughed. Fred's coughing caused the conductor's annoyance, so we have a straightforward causal process explanation.

4.2 The Weak Objection to the Program/Process Distinction

Or do we? Surely Fred's coughing is as multiply instantiable as someone's coughing. Would it have mattered whether the timbre of the cough was slightly different? Or the neural pathways which fired leading up to the cough fired in a slightly different way? If not—and we are bound to find *some* aspect of the cough any change in which would not alter the conductor's annoyance—then this, in so far as multiple realizability is supposed to be criterial, is as much a program explanation as the "someone" case.

There are obvious parallels here with the difficulties for the traditional version of the function/structure distinction. In that case, we saw that even if a description at one level is realized by anything of which a description at a lower level is true, that description at a lower level will also be multiply realizable. In this case, the realization of someone's coughing is by Fred's coughing; Fred's coughing is realized by his narrow behaviour which in turn is realized by some more fine grained specification of his physical state, to which a narrow behavioural description is indifferent.

Consider the case of social explanations which are realizable as different (socially equivalent) combinations of psychological states of individuals and environmental factors. On some accounts of individual psychological factors (see below) these will not turn out to be causally efficacious, and on most accounts they are certainly multiply realizable. Even simple mechanical explanations which advert to mechanical properties like being a camshaft or a piston, are realizable by different geometric properties; and even *these* properties are not causally efficacious. Presumably it is the molecules of the different parts interacting that actually

does the work. Iron atom A, and B and C and so on striking atoms A', B' and C' may be the level at which the actual causal interactions happen.

This is the *weak* objection to the program/process distinction. It is the claim that most of the explanations that we give whose explanantia are multiply realized, have as realizer states multiply realizable and possibly causally inefficacious properties. So, at very least, almost all explanations are in fact program explanations, and I will argue below that whether our interest is in rôle states or realizer states (which may be in turn rôle states with respect to their own realizer states) is accounted for by the pragmatic apparatus of ch. 2.

4.3 Program Explanation and Choice Classes

The apparent distinction between program and process explanation when cashed out in terms of multiple realizability above, can be handled by the characterization of explanation given so far, if we add a principle that the explanans is individuated at the highest level of description required to distinguish it from the other members of the choice class of explanans. Remember that the characterization proceeds as follows:

For some choice class of explanans $A = \{a_1, \dots, a_n\}$ and some choice class of explanandum $B = \{b_1, \dots, b_k, \dots, b_n\}$, some member(s) a_j of the choice class of explanans rather than any other member will be said to *select* the explanandum b_k iff

(i) a_j is *causally relevant* to b_k ,

(ii) for all b_i in B : $i \neq k$, b_i depends causally on the non-occurrence a_j , and

- (iii) the selection of b_k by a_j depends, in some way,
on real causes in the microstructure.

Let us consider again the case of Fred's coughing. The explanandum is the conductor's annoyance. If the choice class of explanans includes various people's coughings as well as Fred's, and the choice class of explanandum includes the conductor's simply proceeding with the performance unperturbed as well as his being annoyed, then the "more process-like" explanation of the two cited by Jackson and Pettit will be indicated, since it is the cough *qua* particular cough which distinguishes it from the other members of the choice class of explanans.²¹

Fred's particular coughing will be chosen from the choice class of explanans rather than other people's coughings, and B, i.e. 'the conductor was annoyed because Fred coughed' will be true iff

- (i) The conductor was indeed annoyed.
- (ii) The conductor did not proceed unperturbed
(nor is any other member of the choice class of
explanandum true)
- (iii) Fred's coughing was causally relevant to the
conductor's annoyance

²¹It is worth noting that there are two ways that a member of the choice class of explanans could fail to select the explanandum; either by failing to have the right causal relevance, or else failing to occur. So the choice class of explanans in this case could contain both coughs that did not annoy the conductor and coughs that would have done so had they occurred, but which did not.

- (iv) The other members of the choice class of explanandum are causally dependent on the non-occurrence of Fred's coughing.
- (v) There is some real causal process in the microstructure which connects Fred's coughing and the conductor's annoyance.
- (vi) No other member of the choice class of explanans satisfies (i) to (iv).

If the choice class of explanans includes only factors *other* than coughings—such as failures of the air-conditioning, what the conductor had for breakfast or his being forced to conduct Tchaikowsky—then the more 'programmatic' account is indicated, since it *was* someone's coughing rather than the rest of the elements of the choice class which explains the annoyance. If, on the other hand, the choice class includes various kinds of cough (and they do make a difference—say the conductor hates high-pitched half muffled coughs), then the explanation will have to be at a lower level, using Fred's coughing in a certain way to explain the annoyance.

The choice class of explanandum can also indicate what level of generality we are interested in. If we are interested in why the conductor was annoyed to such and such a degree at time *t* (rather than to some other degree at time *t'*) then a lower level explanation is likely to be called for, since more of the details of the actual realizer of what annoyed the conductor are going to be relevant to his annoyance so specified. If, on the other hand, we are interested more generally in why he was annoyed during the performance (rather than not annoyed), then the higher level explanation can be appropriate, since any cough would do the trick.

In cases of overdetermination this is particularly compelling. Alan Garfinkel in [Garfinkel 1981] gives an example about rabbits and foxes. If we are concerned with why a particular rabbit was eaten at a certain time, then the details of the rabbit's particular causal history are appropriate. If, however, we are interested in why the rabbit was eaten *simpliciter* (or, better, rather than lived to a ripe old age) then—in the cases where had the rabbit not been eaten by the particular fox at the particular time it would have been eaten by another fox—we are more likely to be interested in more general structural facts about the environment. The rabbit to fox ratio is an example of the kind of structural fact that might be of interest. Of course in the actual case it follows from having been eaten that the rabbit does not live to a ripe old age. Garfinkel thinks that the principle to apply is the preservation of the explanans under perturbation of initial conditions. A better principle is preservation of the explanans under perturbations which preserve the explanandum. So the right explanation in the case where we are concerned with why the rabbit was eaten (rather than lived a ripe old age) is one which will survive a range of perturbations of initial conditions which nevertheless preserve the explanandum—namely the rabbit's being eaten (rather than living to old age). Some structural fact about the environment will do this. However, in the case of the explanation of why the rabbit is eaten at time t (rather than at time t'), there will be some realizations of the structural fact about populations that will *not* preserve the explanandum in contrast to the rest of choice class; for it may be compatible with the rabbit being eaten at t' . But the principle of allowing maximal perturbation which, together with other pragmatic constraints, preserves the explanandum, may also rule out micro-physical specifications as explanata, since there are plenty of perturbations of the micro-physical initial conditions (ie microphysical realizations of the rabbit) which will preserve it's being eaten at t .

Of course the constraint of preserving the explanandum in contrast to the rest of its choice class is not a *sufficient* constraint; there are plenty of perturbations of the initial conditions which will preserve the explanandum, but which do not give rise to intended realizations. Obvious cases are gerrymandered disjunctive explanantia. Consider the general explanation of why the rabbit is eaten at some time; one proposed answer is some structural fact about the population. This is realizable by a whole range of physical states of the environment. By saying that we are interested in the structural fact, we are signalling indifference to this range of variation of realization. But what if the proposed explanans was the structural fact *or* a nuclear holocaust? This is also multiply realizable, and presumably on each of the realizations the rabbit fails to live to old age. Does this disjunction program-explain? Perhaps, perhaps not. It depends on whether the pragmatic criteria which determine what we regard as having features relevantly in common count as criteria for program explaining at all, or whether they are measures of how good the explanation is. The gerrymandered explanation specifies a state²² such that, if it holds, there *is* a cause of the explanandum; and it gives us some of the extra information that a program explanation is supposed to give. It is just that it is wholly useless information.

At least some of the examples in Pettit and Jackson of the supposed Program/Process distinction, then, are special cases of context-dependent

²²One strategy, if one leant to the side of claiming that there is no such program explanation, would be to require that not just *any* weak functional state which programs for causes is required as an explanans, but a strongly functional state is required. Presumably there is no scientific account of functions on which having a certain fox-rabbit ratio or being subject to a nuclear holocaust counts as a strong functional state.

specification of the level of description at which an explanation is required. It remains to be seen if this is true for the account as a whole.

5 Causally Efficacious Properties

It could be objected, however, that I have so far ignored the other crucial feature of process explanations. This is that process explanations advert to *causally efficacious properties*, and program explanations do not. Perhaps, (although Jackson and Pettit do not explicitly formulate it this way) multiple realizability is a necessary but not sufficient condition for program explanation, but together with failing to advert directly to a causally efficacious property, and one of its realizations (or one of its realizations' realizations!) being a causally efficacious property, we might have a jointly sufficient condition. For process explanation, adverting directly to a causally efficacious property will be sufficient, and it can even be allowed to be multiply realizable if the preferred account of causal efficacy allows causally efficacious properties to be multiply realizable.

On this version the theoretical burden of process explanation is placed firmly on the notion of a causally efficacious property. I do not have an account of what a causally efficacious property might be, but presumably the idea is that causally efficacious properties (or at least the things picked out by them) *do the actual causal work*. The example in Jackson and Pettit that seems to have the most chance of adverting to genuinely causally efficacious properties is the example about boiling water and glass. In this example we are asked to consider the explanation of why a glass beaker breaks in terms of the fact that it contains water which is boiling. They point out that it is the particular history of certain water molecules hitting silicon dioxide molecules with certain velocities at certain angles which actually causes the glass to break. The story about boiling water is a program explanation in as

much as it guarantees that there will be *some* molecules which have the right velocity etc, and it does the required extra work by telling us that certain differences between states of affairs where different actual molecules with slightly different velocities are involved, are irrelevant for many purposes.

The intuitions at work in this example are much easier to understand than in the Fred coughing case. There does seem to be a sense in which the level of molecular interactions between glass and water is where the causation is going on; and the higher level account, while it may be explanatory for the reasons outlined above, does not add anything to the story about the *actual* causes involved.²³

Let us suppose, then, that there is some basic physical level at which real causal processes take place, and that causally efficacious properties are (at least) properties of things at that level. My strategy here will be to argue that while it is possible to advert to processes at this level, such adversions will typically not in themselves be *explanations*.

Take the boiling water and glass case. The process explanation on offer is, say, that water molecules $M_1, M_2, \dots, M_n$ have velocity $V_1, V_2, \dots, V_n$ at time T , and that silicon dioxide molecules $S_1, S_2, \dots, S_n$ have masses $W_1, W_2, \dots, W_n$. Such an explanation, fleshed out with enough detail, might satisfy the constraints of the Deductive Nomological account of explanation; or if it turns out there are hitherto unnoticed irreducibly probabilistic factors, some other suitably modified covering law account. But we have

²³It would be possible to argue that the molecular level is not the causally efficacious level at all; perhaps it is the sub-atomic level, or perhaps there is a progression down levels. I shall not pursue this here, since the comments to follow apply to *any* account of the bottom level at which causation actually takes place.

seen in ch. 2 that such covering law accounts are not enough. For there is a potential infinity of such accounts. Why pick the disposition of those molecules at those times? Why not pick the disposition of them at some other time—a moment before or after? Why not pick the (admittedly vastly more complicated) disposition of causally efficacious facts much earlier on from which (in virtue of the supervenience of mind on brain) it could be predicted that I wanted to boil the water to make tea?

The answers to these questions is that we have certain interests in asking the why-question in the first place, and an adequate answer to the question will have to be responsive to those interests. Perhaps we are interested in facts about molecular motion 30 seconds before the break because for some reason we have control over these facts, or perhaps we are interested in the more distant facts at that level because we have influence over them. Or, perhaps we are interested in some part of the causal history and geography of the breakage at that level because that part is a part that we *don't* have control over, and is therefore the source of the mysterious break. In all of these cases, though, we can pick the level of description as being the one at which the real causal processes occur.

How are these interests cashed out in terms of choice classes? It seems that our old friend the choice class of explanans does the work of describing these interests. If we care about facts of a certain kind at this level, at a certain temporal and geographical remove from the explanandum event, then it is the class of such facts which constitutes the contrast class of explanans. And, from above, it is membership of that class which is necessary for the fact at this level to be an explanation. So, although it is possible to *stipulate* that one is concerned with every detail of the actual process as it occurs, in fact the interests that usually inform such an explanation suggest that it is just in virtue of having a certain higher-level

property that we are interested in it. It is rather like the reverse of rigid designation by a higher level property. We may refer to the higher level property by using a property which is an actual property of the higher-level property's realizer.

This does not completely undermine the status of process explanations. For it shows only that process explanations do not explain solely in virtue of the causal efficacy of their explanantia. Something else is required; membership of a contrast class of explanans the membership criteria of which are determined by explanatory interests. On any account of causal efficacy, the property of being a member of *that* class will surely not count as causally efficacious. However there may still be an important distinction to be had between explanations for whom a *necessary* condition is that their explanans be causally efficacious, and program explanations which require only that some realization of the explanans be causally efficacious.

I do not think, however, that there is any principled reason to suppose that causal efficacy is a necessary condition of any explanation's explanans, unless by needless stipulation. For if program explanations are held to explain in virtue of the contrast classes and the requirement that some realization of the explanans be causally efficacious, then a process explanation will explain *for the very same reason*; but it just so happens that the realization which is causally efficacious is the explanans itself. I take it that there is nothing to stop something being a (or even the!) realization of itself. So it is not the causal efficacy which is *required* in these process-like cases to make the purported explanation explanatory; though adverting to the causally efficacious property is one way of having what is required.

This is not just a terminological question; it has to do with whether you think that explanations whose explanantia are themselves causally

efficacious, and explanations whose explanantia have a realization which is causally efficacious, explain for significantly different reasons. Since the criteria which make program explanations explain will do for process explanation, considerations of theoretical unity and economy might make us want to dispense with the program/process distinction, unless some other special argument is produced which makes the distinction look more important.

5.1 Why Does The Program Process Distinction Look so Plausible?

The reason that the distinction seems such an important one can be found in the problem that Jackson and Pettit begin with. It is the problem of wondering how something explains anything which does not advert directly to properties causally relevant to the explanandum. To wonder about this at all, you would need to think that a *basic* feature of explanation is to point out the actual causal processes which led to the explanandum. It is only *then* a puzzle why an explanation which adverts only to a higher level property the actual realization of which is causally efficacious is anything other than incomplete; or why an explanation which tells us *extra* things, like some higher level property of the explanans, is anything other than redundant.

Having answered this puzzle, though, and produced an analysis of explanation which allows these 'odd' explanations to explain, and discovered that this account works just as well for the 'basic' cases, it is time to reassess how basic the basic cases really are.

One thing is certain: if there is a philosophical or scientific theory to be had on which the level of basic causation is buried deep in the microstructure of the universe, and according to which macroscopic apparent causation supervenes on this, then the process explanations in

terms of this basic level of causation are certainly not basic in the sense of being the normal or standard kind of explanation. Indeed they are extremely rare; most explanatory talk is at the macroscopic level; and if the 'basic' level of causation is well down at the subatomic level this will include most of physics as well. The intuition which leads people to distinguish between the process explanations and the program ones is surely not going to be met by the mass of (irrelevant for most purposes, *especially* purposes involving controlling things) detail at this basic level. If we have no theoretical reason to distinguish explanations at this level from others, it seems we may have no practical ones either.

5.2 Causal Efficacy at Higher Levels

This would be true if causally efficacious properties were only to be found at this bottom level of reality. If not, then perhaps many of our explanations do indeed invoke causally efficacious properties, and if there is a distinction to be made between causally efficacious property and causally inefficacious property explanations *at the same level*, then this will be a distinction which will not be as easily captured by treating process explanations as the limiting case of program explanation at the bottom of a hierarchy of levels of description.

What if one were to *specify* that the explanans was not intended as a multiply realizable higher level description, and in a Kripkean mood insisted that what explains the conductor's annoyance was *that very cough*; one which has all the features at every level which it does have? There are various ways one could go about this. We could have a Davidsonian story in which we invoke a well individuated event, Fred's cough, which is how it is in virtue of its individuating causes and effects, regardless of how it is *described* [Davidson 1969, 1970, 1971]. This particular event would cause the

annoyance, and partly in virtue of this explain it. The particular event is not multiply realizable since it is determined by its causes and effects, and on the Davidsonian account those causes and effects are what they are *sui generis*. Any difference in realization would require some difference in cause or effect, and thus mean that we were not dealing with the same event.²⁴ Or, we could go for a more ontologically modest account, and say just that, whatever else is true, there is a network of causation which leads into the region of the world in which (or supervenient on which) we find the token annoyance. Amongst these causal networks, an important tributary is the one which is picked out in some way by Fred's Coughing, and an explanation of the annoyance is *whatever is (actually) really going on causally* in the region which contains the coughing .

Both of these stories seem to avoid any hint of multiple realization. In the Davidson account this is achieved by stipulating that we are dealing with a particular, even in cases where we do not have epistemic access to the complete nature of the particular. In the second case this is done by stipulating that we are concerned with *every* level of description, right down to the ones at which real causal processes are at work.²⁵

²⁴This seems as though it reduces the doctrine of the non multiple instantiation of events to the doctrine of non multiple instantiation of causes and effects. Since causes and effects are themselves events, there is a suspicion of circularity here. However I am not concerned to argue for or against the doctrine here. If it stands, then what follows above is argument enough for my purposes; if it falls then so much the better.

²⁵If there is no such level, and there is no bottom level of nature, this would still pose no problem for this account since we could simply stipulate that we are concerned with *every* level of nature. If causality were, for example, to consist in ineliminably probabilistic relations which depended on ineliminably probabilistic relations at lower levels *ad infinitum*, then stipulating that we intend *all* of these would do.

These different possibilities have three things in common. They all permit of causally efficacious properties at the high levels, and they do not invoke multiple instantiability and they advert to the actual causal processes at work in each case. Now if some story about real causation at the bottom level of nature is true, then these higher level causally efficacious descriptions will be causally efficacious *because of* the causally efficacious properties at the bottom level of which they are higher level descriptions. In this case my first response will be to say that they are redundant; the causal work is done by the story at the bottom level—what is added by the higher level description? Of course in the *program* explanation case a lot is added; information about what counts as relevantly the same for explanatory purposes. But in the process case, the sense of ‘that cough’ in which being ‘that cough’ is viewed as causally efficacious, is just the sense in which it is viewed as being a particular network of causal chains at the bottom level.

Another way to put this is to note that it is non-trivial supervenientism (i.e. supervenience with multiple realizability—see ch. 5) which gives ontological respectability to higher level properties; they are needed because all that we know about them is not exhausted by the actual realization of the higher level property. Things could be otherwise and the property still realized. But if we stipulate that we mean by some higher level description the thing as it *actually is* in detail, then this guarantee of ontological respectability disappears and the property, not being multiply instantiable, becomes reducible. So the claim that there could be, say, a causally efficacious and a causally inefficacious version of ‘that cough’ both at the same level, becomes less plausible. There will be the sense in which ‘that cough’ is multiply realizable by a range of physical processes, the actual one being causally efficacious—and this sense will be causally inefficacious. There will also be the sense in which we mean by ‘that cough’ the actual

process; but by reducibility it will not be at the same level. In fact it will be at the bottom level. And, while it will be causally efficacious, by the argument in the preceding section, it will explain for exactly the same reason as the properties in program explanations, and indeed will be a limiting case of them. Chapter 5 will provide further arguments to show that considerations of multiple realizability eliminate the problem for views of causation that place causal processes somewhere deep in the microstructure, that there might be a *need* for higher level causally efficacious properties to make sense of our ordinary explanatory practice.

6 The Constraints Reformulated

It should now be possible to reformulate the constraints of chapter 3 on the selection relation in order to show how explanation by higher level properties is possible. The difficulty was with the third constraint: that the explanation depended in some way on whatever science tells us is the real level of causal processes. If this turns out to be some kind of microstructural process (see ch.5), then it would be hard to see how the extra information in a 'program' or higher level explanation adds anything to the information about causes. The answer is that it adds *counterfactual* information about real causal processes. Exactly what kind of counterfactual information must wait until we have considered the first two constraints.

What of the first and second constraints? There is no need for a reformulation in these cases. These constraints were:

- (i) a_j is *causally relevant* to b_k
- (ii) for all b_i in B : $i \neq k$, b_i depends causally on the non-occurrence of a_j .

The definition of causal dependence, and hence of relevance, for events was that 'Some event e *causally depends* on some event c iff $(O(c) \Box \rightarrow O(e)) \& (\sim O(c) \Box \rightarrow \sim O(e))$ '. Now, in the case of a multiply realizable higher level explanans and explanandum, how are we to interpret these requirements? Both rely on the notion of causal dependency, of course. Let us first consider the case of causal dependency between two multiply realizable higher level events, each of which actually occurs. Let us consider them as a disjunction of realizations. If they both occur, this means that both are realized. In this case, the first half of the requirement is fulfilled, so the truth of the conjunct depends on $\sim O(c) \Box \rightarrow \sim O(e)$. The criterion for this counterfactual holding is that there is some world at which $\sim O(c)$ is true and $\sim O(e)$ is true which is more similar to w than any world where $\sim O(c)$ is true and $\sim O(e)$ is not. This means that there is some world which contains *no* realization of c (and thus $\sim O(c)$ is true), which is more similar to the world w of the explanation than any world which contains *no* realization of c and in which there is some realization of e .

In the case where neither of them occur, the truth of the conjunction depends on $O(c) \Box \rightarrow O(e)$, so there must be some world where there *is* some realization of c and in which there is also a realization of e which is more similar to w than any world in which there is a realization of c and no realization of e . Note that in these multiple realization cases the possibility that there is not necessarily any unique world which is more similar to wetc, is plausible. In, for example the second case, it is plausible that there are various worlds in which there is some realization of c and in which there is also a realization of e which is more similar to w than any world in which there is a realization of c and no realization of e , corresponding to different realizations of c and e . This is because there might be different realizations of c and e , which are equally different from the actual realizations of c and e ,

different from each other, and nevertheless realize whatever properties are required to make them *bona fide* realizations of *c* and *e*.

We are now in a position to reformulate the third constraint using the classes of worlds above to pick out what kind of counterfactual information about causal processes is required. The constraint can be reformulated in two parts as:

(a) the actual realization of the explanans is related by whatever the real causal processes are to the explanandum.

and

(b) that in every world where the explanans and the explanandum are realized, *and* which is more similar to the world (*w*) of the explanation than any world where the explanans is realized and the explanandum is not, there is, in that world, a real causal process which links the realization of the explanans to the realization of the explanans.

What this tells us is that not only is there a particular cause linking the realizations of the explanans and the explanandum in this world, but it gives us counterfactual information about the kinds of causes operating in a range of relevantly similar worlds. It does not, of course, necessarily tell us what these real causal processes actually are; the requirement on the explanation is just that there *are* such processes.

The third causal constraint has been reformulated to allow the possibility of explanation at higher levels. The fact that the first two do not need reformulation forms part of my motivation for distinguishing between

real causal processes, and causal relevance. What Lewis calls the causal relation turns out to be able to hold between disjunctions, or multiply realizable events. Yet an intuition which is worth preserving is that causes are discoverable in the actual world.²⁶ Calling it the causal relevance relation allows us to enhance its considerable power in explanations, with the requirement that there be real (and no doubt different) singular causes in the various worlds which realize the higher level events.²⁷ The next chapter explores more closely whether such a notion of actual microstructural causes could be sufficient for our understanding of explanation.

²⁶For someone who wants to hang on to the Lewis account of *causation*, yet without the possibility of causal relations between multiply realizable events, one way to go would be to take the special case of intrinsically non-multiply realizable events (say, exhaustive microphysical accounts or Davidsonian events) and apply the Lewis analysis to those. They could then avail themselves of my three constraints, with this as an analysis of the 'real causal relations' in the third constraint.

²⁷A possible problem for a scientific analysis of what the causal relation really *is* in terms of a physical reduction, is that on this account, we would have to restrict ourselves to worlds which share this physical feature with our world. To release oneself from this shackle, it would be necessary to abandon the intuition that causes are whatever they actually are in the structure of this world, and take some analysis like the one suggested in the footnote above as *criterial* of causation, and allow that there might be different underlying regularities in other worlds which seem to behave this way.

Summary of Part One

and some last remarks

The account of explanation which has been developed in Part One has two basic components, the pragmatic component and the causal component. In addition there is the recognition that some explanatory information is counterfactual information about causes. To modify slightly the slogan of ch. 3, explanation is the pouring of sometimes counterfactual information about causal structures through a pragmatic net. The basic notion is of *selection* of the explanandum from the choice class of explanandum, by a member of the choice class of explanans.

Selection was defined as follows: for some choice class of explanans $A = \{a_1, \dots, a_n\}$ and some choice class of explanandum $B = \{b_1, \dots, b_k, \dots, b_n\}$, some member a_j of the choice class of explanans will be said to *select* the explanandum b_k iff

- (i) a_j is *causally relevant* to b_k .
- (ii) for all b_i in B : $i \neq k$, b_i depends causally on $\sim a_j$.
- (iii) the selection of b_k by a_j depends, in some way, on real causes in the microstructure.

The third constraint was reformulated in ch. 4 to read:

- (a) the actual realization of the explanans is related by whatever the real causal processes are to the explanandum.

and

(b) that in every world where the explanans and the explanandum are realized, *and* which is more similar to the world (*w*) of the explanation than any world where the explanans is realized and the explanandum is not, there is, in that world, a real causal process which links the realization of the explanans to the realization of the explanans.

The last constraint is in some sense a gloss on the first constraints, available to those who think that the basis for a notion of causality can be searched for fruitfully in physics. If you had a *purely* counterfactual *analysis* of causation, then the first two might suffice.

The three components of the account of explanation are the pragmatics, actual causation and counterfactual causation. The pragmatics enter into the account in at least two ways. (1) Through the selection of the choice class of explanans—the sorts of factors in terms of which one is seeking explanations. (2) Through selection of the choice class of explanandum—the range of alternatives one of which is selected rather than the rest by the explanans, and which consequently constrains what will count as causally relevant under the second constraint.

The second component is the causal component, and this is introduced in the form of the constraints above on the selection relation. Selection of the explanandum over the rest of the choice class is selection *in virtue of causal relevance*. Finally, the third component of the account is the recognition that we are not only concerned with information about actual causal processes, but also about counterfactual causal processes. As is shown in ch. 3, the causal relevance constraints can be extended to cover counterfactual cases

given the analysis offered there of what it is for a multiply realizable supervenient entity to satisfy the causal constraints.

A last comment about the relationship between selection and explanation: it is tempting to have a model on which, relative to a why-question, if there is a unique member of the choice class of explanans which selects the explanandum from the choice class of explanandum, then the explanans explains the explanandum. Given the possibility of counterfactual information which tells us that the explanans selects the explanandum counting as explanatory, however, this lets in unmotivatedly disjunctive explanations. If *As* cause *Bs*, *Cs* cause *Ds* and *Es* cause *Fs*, and we arbitrarily stipulate that there is some higher-level event, *X*, which can be realized by *As*, *Cs* or *Es*, and another higher level event, *Y*, which can be realized by *Bs*, *Ds* or *Fs*, then whenever we can explain, say, *B* by *A*, we are in a position to explain *Y* by *X*. At every point in a theory of explanation there comes a point where we must distinguish between no explanation and bad explanation. My preference is to say that the explanation of *Y* by *X* is an explanation, because in some contexts, with gerrymandered pragmatics to match the gerrymandered events, it will meet the constraints above. It is just that it is *bad* explanation; the required gerrymandered pragmatics just won't answer to any genuinely occurring interests, only explanations in terms of events picked out by natural or non-gerrymandered properties will do that. On the other hand, I have no real quarrel with someone who wanted to say that if the multiply realizable properties are completely unmotivated, then there is *no* explanation in terms of the higher level property. To get this view off the ground, all that is required is an extra constraint that the explanans and the explanandum must not be gerrymandered. I leave the choice between these two possibilities to the reader.

Part Two

Chapter Five

Explanation and the Microstructural Cause Hypothesis

Table of Contents

1	Introduction	158
2	Why We May Want To Restrict Causation to the Bottom Level.....	159
3	The Competing Accounts.....	163
4	The Case Against Causal Reductionism.....	164
5	Kim's Event-Supervenience.....	164
6	Menzies' Objections.....	166
	6.1 Symmetry.....	167
	6.2 Events and Physical Objects.....	168
7	Varieties of Supervenience and Menzies.....	171
	7.1 Circumstantial Supervenience.....	171
	7.1.1 The Counterexample	174
	7.2 Logical Supervenience	176
	7.2.1 The Counterexample	177
	7.3 Typical Causal Rôle Supervenience.....	178
	7.3.1 The Counterexample	178
	7.4 Mereological Supervenience.....	180
	7.4.1 The Counterexample	182
8	Explanation and Supervenient Causation.....	183
9	Against Supervenient Causation	184
	9.1 Explanation and Causation	185
10	Explanation and Event Identity.....	187
11	Summation	188

1 Introduction

The desire to accommodate ordinary talk about causation with some of the spirit of scientific reductionism leads to some alarming theoretical tensions. On the one hand, there is the intuition that the actual physical transactions which make the universe tick take place somewhere in the universe's microstructure. Adrian .i.Heathcote; [Heathcote forth.]¹ even has a candidate place for the real causal transactions—interactions between Quantum fields. On the other hand, our everyday talk of radiators causing the room to be warm, politicians causing parliament to be dissolved and Magpies causing fear in the breasts of cyclists needs to be accommodated. It will be the thrust of this section that if one's account of explanation does not require that the explanation itself give information directly about the causal processes, but rather gives more abstract information in terms of properties, any of the realizations of which would in fact also be realizations of the causal properties, then some of these tensions can be relaxed. I will argue that the right way to look at the relationship between macroscopic entities is that

¹His proposal is that causation be identified with certain basic interactions in quantum field theory. The theory is exclusive in the sense I will outline below, in that it requires that should anything else that does not contain such interactions meet our pre-theoretic intuitions about causality, we should deny that such things are caused at all. Thus he (tentatively) suggests that one way to deal with the counter-intuitive implications of the EPR experiment is to deny that the regularities which emerge are causal ones. I do not think, however, that a proponent of a physical theory of causation need go this far in flouting our intuitions about the causal basis of regularities. It is one thing to admit the possibility of accidental regularities in the universe; this is just to deny that regularity is *critical* of causality. It is another to blithely allow massively systematic regularity to be uncaused. We may be suspicious of this not because we think that regularity is critical of causation, but rather because the massive regularity leads us by induction to suspect that we have not yet found the most general theory of causation.

they *explain* each other rather than cause each other, and that this is a better strategy than either abandoning causal reductionism entirely, or attributing a kind of Clayton's causality to macroscopic entities through supervenience. Supervenience, it will be argued, is properly a doctrine about events and objects, not about causes. My aim in this chapter is not to argue in general for the view for the view that causation might be reduced to certain features that physics might discover in the microstructure, although I do provide some motivation for it. My concern is rather, on the one hand, to defend this view from particular objections using the account of explanation in Part One, and on the other hand to elaborate on the discussion of explanation and causation in the last two chapters.

2 Why We May Want To Restrict Causation to the Bottom Level

There are at least two motivations for restricting the proper use of causal language to some basic level of nature. The first is an Occamite observation; in the case where two event types are constantly conjoined because they have a common cause, we are not tempted to claim that they cause each other, or that the earlier causes the later or whatever. In the case of microstructural processes and high level causation there seems to be a similar structure. If the microstructural realizer of a higher level event causes it, and also causes another microstructural realizer of a higher level event, then there will, if the state of affairs is lawlike, be a more or less constant conjunction between the higher level event types. But the situation is analogous to the common cause story: there seems to be a common cause of the higher level events in the microstructure; the only complication being that there is a two-place process rather than a unitary cause in each instance.

The second motivation is that it allows us to take seriously the possibility of a scientific analysis of causation. This does not mean that there will necessarily be no stipulative element in such an analysis. When some candidate for the basic unit of causal transaction comes up, it may well be an act of philosophical stipulation to say that *that* is a *causal* transaction, even though the physical nature of the event is known through physics. This just says that physics may not exhaust our grasp of what cause *means*, because perhaps we could imagine other possible worlds, albeit with different natural laws, in which cause was something else. But this does not mean that our linguistic intuitions about cause exhaust causation.

This is, I suggest, where Humeans about causation—or neo-Humeans who introduce counterfactual sufficient conditions—go wrong. For the intuition that cause just is cosmic constant conjunction rules out the possibility that there could be such a cosmic constant conjunction that was not an instance of a causal regularity. So much the worse for non-causal cosmic coincidences, I hear the friends of Hume say. After all, any theory has its consequences, and we are prepared to pay that price.

There is another price that has to be paid, though, and that is that a regularity account puts strong *a priori* constraints on the possibility of scientific discovery about causation. If there is an empirical hypothesis that causal transactions are a certain kind of micro-interaction, and there are constant conjunctions which do not feature these interactions, then the theory about the microstructural relations is falsified. Yet surely we would not want such an easy victory for the folk notion. It is consistent with taking science seriously that *some* weight be given to pre-scientific intuitions. I take it that (c.f. appendix) we now think that water is H₂O because *most* of what answered to our pre-scientific intuitions about water turned out to be H₂O. Similarly, a scientific hypothesis about causation which guaranteed no

constant conjunctions nor supported any counterfactuals could properly be objected to on *a priori* grounds—that just isn't what anyone means by cause. But that is not to say that the fit must be perfect, just as we now exclude from counting as water some of the wet stuff which might once have been counted as such. Just so with causality, if we want to treat it scientifically. Pre-scientific notions have *some* weight, inasmuch as they roughly delineate what we are talking about. But there is no reason for them to continue to have overriding weight in the light of any good empirical hypotheses we may come up with about causality.

Purely counterfactual accounts—ones where the counterfactual analysis is taken not as an exposition of a feature of causation but as exhaustively constitutive of it—are not much better, since they are just extensional across possible worlds. The same sort of problem is at least conceivable. A good scientific theory about the nature of causation is forthcoming, but it turns out that there is some constant conjunction across the (physically) possible worlds such that, in every one where event, E, happens, so does another event, E*. If this seems odd, then it is because we think that the range of physically possible worlds is constrained by whatever the true account of causation is anyway, so such a coincidence ought not occur. If it did, it must be because there is buried in our physical laws some *other* candidate for the proper site of causation. So perhaps we should allow this to count against our candidate, but not for *a priori* reasons. This means that a counterfactual account of causation is plausible just in so far as it merely organizes our physical knowledge. Whether the counterfactual story is really constitutive of causation would depend on the further test of whether you think that in some set of worlds which do *not* share our causal laws, the constant conjunctions which held across them would be constitutive of causation. At least the simple constant conjunction

account has the merit that it is possible to identify the relevant world—the actual one—independently of an account of what the causal law(s) are. In the counterfactual case, since we will not require constant conjunction in all *logically* possible worlds, a way of distinguishing which of those worlds should count as the physically possible ones is required. Such a way could only be knowing the physical laws. So whereas it is (in principle) possible to read off what causation is from the actual constant conjunction account, no such benefit attaches to a counterfactual account.

Constant Conjunction stories and their counterfactual kin would license causal claims between all sorts of entities, including the macroscopic ones. But if we want to avoid allowing such accounts to be sufficient for causation, so as to allow a physical story to have a slice of the action, then *this* reason for supposing that causation is a relation which can hold between macroscopic entities is not available.

So are there other reasons? One obvious candidate is the possibility of an account of the physical identity of causality which is located at some macroscopic level. If, for example, Newtonian gravitational theory were true and irreducible, then there would be both action at a distance and, plausibly, holistic causal interaction between one macroscopic entity and another, since for the purposes of Newtonian theory it is the total mass of, say (for simplicity's sake) each of two objects in a gravitational interaction, that is relevant. But Newtonian gravitational theory is not true. I do not propose to argue here that there could be no physical theory of causation which locates causal interactions at the macroscopic level, although I do not think it likely that there is, since my purpose is to decide what to say *if* the right physical theory locates it at a microscopic level. However, even if some physical account of causality is forthcoming at macroscopic levels, it should be remembered that not all such accounts will make *all* of what seem pre-

theoretically to be macroscopic causal interactions into legitimate causal interactions. If (physical, not socio-political!) masses were to turn out to have irreducible causal interactions, would that mean that social classes and economic conditions would do so too? Probably not, with the exact outcome depending on the physical theory of causation on offer. In any case, even were there a possibility that the physical basis of causation was sometimes locatable macroscopically, we would still need an account of the relationship between microcausation and a whole range of macroscopic causation not covered by the theory. If we need such an account anyway, and if it could deal in general with the relationship between macroscopic causation and a putative micro level physical base, then in the absence of overwhelming *empirical* reasons to suppose that there is unmediated direct physical causation somewhere at the macroscopic level, then theoretical parsimony will advise us to leave it out of our philosophy.

3 The Competing Accounts

What, then, is the competition? I take it that, with a background assumption about the truth of physicalism, there are at least three views about causation which remain obstacles to the thesis that causation occurs only at the bottom level of nature, and that explanation is the relationship between things at higher levels. Two of these views differ less markedly from mine than the others. These are the doctrines of supervenient causation of Jaegwon Kim [Kim 1973 and 1984] and Donald Davidson [Davidson 1980a and 1980b]. Common to these views and mine is that whatever it is that is *called* causation between macroscopic entities, it at least *depends* on microscopic causation. The third view is rather stronger, though. This is the view of Peter Menzies in his 'Against Causal Reductionism' [Menzies 1988]. He argues that macroscopic causation not only happens, but does not even depend on microscopic causation. He calls views which require this dependence *causal*

reductionism. I shall deal with this third and more radical claim first, before returning to argue that a notion of macroscopic explanation can take the high level causation out of the supervenientists' world view.

4 The Case Against Causal Reductionism

The structure of Menzies' argument is as follows. He argues that Kim's fine-grained model of events and event supervenience does not mesh with some central intuitions about supervenience and dependence. He therefore formulates a notion of event-supervenience which he thinks will cope. Loosely following Goldman, [Goldman 1970 ch. 2], he thinks that supervenience is not a univocal notion. It turns out, on his view, that there are *four* proper notions of supervenience, and for each of these he finds a counter-example to the thesis that the causal relations between an event and other events depends on the causal properties of the event(s) on which it supervenes. He then moves on to Davidson's coarse-grained conception of events, worries about it, then claims that it can be reconstructed as a special case of the fine-grained view and is therefore subject to his (Menzies') counter-examples.

I will first discuss Menzies views on Kim's supervenience, and defend certain features of the Kim view. I will then go on to look at Menzies' kinds of supervenience in this light, and argue that the counterexamples do not hold. Finally I will reply to an answer that Menzies gives to an objection to his position.

5 Kim's Event-Supervenience

Kim's account of events is a *fine-grained* one. It is called fine-grained because on his account events are numerous. Indeed there are as many events as there

are instantiations of properties of objects.² Thus they carve the world up into events finely. While on a coarse-grained conception of events micro-level and macro-level *descriptions* may be true of the the same event, on a fine-grained account there are separate micro-level and macro-level events. Thus, unlike Lewis [Lewis 1986a p.241], for Kim any nominalization is an event. So some physical object A's having the property F at t counts as an event. Some other event—say B's having G at t'—counts as the same event iff B is the same object as A, G is the same property as F and t' is the same time (instant or period) as t. For Kim, event supervenience depends on property supervenience. Property supervenience is defined as follows:³ a family of properties, A, supervenes on another family of properties, B, exactly if for any property F in A, if any object x has F, there is a property G in B such that x has G, and necessarily anything having G has F. In this case, G is a supervenience base for F. Supervenience between events which share the same physical object is thus easily defined. The event of a's having F at t supervenes on a's having G at t just in case a has G at t and G is a supervenience base for F.

On Kim's moderate supervenientist plan, a's having F and b's having G are related by supervenient causation just in case there are more basic events a's having F* and b's having G* such that a's having F supervenes on a's having F*, b's having G supervenes on b's having G*, and a's having F* causes b's having G*. Kim's analysis of macroscopic causation is that it always consists of supervenient causal relations thus defined, and there are

²In the language of Sydney metaphysics there is a Kim event for every trope.

³Roughly following [Menzies 1988]: this is little clearer by way of omitting the time specification than the definition in [Kim 1984].

at some microlevel basic causal relations which are not supervenient in nature. I shall return to my criticism of the Kim position after a defence of it from Menzies' objections.

6 Menzies' Objections

The first of Menzies' objections that I want to deal with is that Kim's definitions make event supervenience a reflexive and non-symmetric relation. The reflexivity is supposed to be objectionable because it is supposed to 'run counter to the intuitive conception of event-supervenience as a relation of *dependence* between events' [Menzies 1988 p. 4].⁴ He supposes simply that no event depends on itself, and that this is intuitively obvious. I do not share such an intuition. Indeed, in the case of basic events at the bottom of a supervenience tree my intuitions go the other way. But in any case, intuitions about what we think is true of dependence are not very robust; surely not robust enough to be an obstacle to a well worked out *theory* of supervenience, since dependence is just the kind of hazy notion that is supposed to be *explicated* by the theory. More importantly, however, reflexivity does seem to be allowed by a more precise metaphysical basis for supervenience than talk of dependence, namely, realizability. One source of the push for a notion of the supervenience of the macroscopic on the microscopic arose because of reflection that, on a physicalist world view, there can be microscopic change without macroscopic change, but not vice-versa. So there must be macroscopically irrelevant differences between microscopic objects or events. Thus the event, say, of Kim's laughing, while supervening on his exact physical state, is not reducible to that physical state

⁴Page numbers refer to a pre-publication copy. These may not match those in the October 1988 *Mind* which was not available by the time of submission.

*because other physical states could just as well have realized that state.*⁵ So Kim's laughing could be said to be *multiply realizable* by a number of physical states. For Kim's laughing to supervene on some physical state, then, would just be for that physical state to be the actual realization of his laughing. If this forms the intuitive ground for our notion of supervenience (and it is not so far incompatible with Kim's definition) then it is more obvious that supervenience should be reflexive, since there is nothing odd about something being a realization of itself, it is merely trivial in most cases. And to see that it is not always trivial consider the possibility that there are objects or events at the very bottom level of nature. If they were not realizations of themselves *they would have no realizations*, which would be very unfortunate for the intuition that what makes something existent is that it is realized.

6.1 Symmetry

Menzies also objects that Kim's definition makes event-supervenience non-symmetric when it should be asymmetric. He says '...if one event depends on another, the second event cannot depend on the first.' [Menzies 1988 p. 4]. While this may be mostly true, there is one case where, from what I have said above, it is not: when those events are identical. It simply follows from reflexivity that the relation is not asymmetric; no special argument is required if reflexivity is motivated. That it is non-symmetric, rather than symmetric, is motivated by the intuitions which I take it guide Menzies at this point. It would be odd if symmetry were true just because it would be odd if any two things supervened on each other unless they were identical.

⁵Not possible, of course on Davidson's coarse grained view. I address myself to the fine-grained view here since in any case Menzies thinks it is the more intelligible.

6.2 Events and Physical Objects

Another worry which Menzies has about the Kim definition is that it rules out the possibility of an event's supervening on another event, if the events involve different physical objects. Menzies claims that it is in fact *typical* of interesting event-supervenience claims that the events involve different physical objects. He cites as an example the supervenience of temperature of a body of gas on the conjunction of states, each of which gives the kinetic energy of a constituent molecule of the gas. Surely in this case one event is the property of the *body of gas* at *t*, and the other event is a conjunction of properties of *individual molecules*.

Certainly we want a notion of supervenience which will allow the temperature of a body of gas to supervene on the kinetic energy of individual molecules; but we also want to rule out certain possibilities. It would not do if it were possible for the temperature of a body of gas to supervene on the kinetic energy of the molecules in *another* body of gas. Kim's requirement that event supervenience be a relation between properties of the same physical object guarantees that, even if it guarantees too much. Whether it does guarantee too much depends on the conception of physical object at work. It is too strong a requirement if the conception of physical object involved is one involving common-sense objects such that molecules of hydrogen and clouds of gas are different objects or, more pertinently, the conjunction of molecules is a different object from the cloud of gas. But is there any reason Kim is obliged to work with such a notion? Just as there are fine grained and coarse grained theories of events, there could be fine-grained and coarse-grained theories of objects. On a coarse-grained theory of physical objects, the cloud of gas and the conjunction of the molecules could count as the same object under different descriptions, and thus Kim's definition would not prevent cloud events from

supervening on molecule-conjunction events. It could be objected that it is on the conjunction of individual properties, not some property of the conjunction, that cloud events should supervene. But it is surely a property of the conjunction of molecules that each conjunct has a certain kinetic energy, and I do not see that there is any objectionable holism here. I do not favour such a view, for reasons which I hope will become clear (and are expressed in ch. 4), but nothing Menzies says excludes the possibility.

My preferred option would be to incorporate a notion of object-supervenience (see ch. 4), defined by multiple instantiability, and to reformulate Kim's version as follows: A family of properties A supervenes on a family of properties B iff for any property F in A, if any object x has F, there is at least one property G in B such that an object y has G, x object-supervenes on y, and necessarily any object which has G subvenes on something which has F. A could be said to *exclusively* supervene on B iff, in addition, for any object z, if there is a property H of z such that necessarily anything which has H subvenes on something which has F, then H is in B. The *event* of a's having F at t. then supervenes on the event of b's having G at t. iff F supervenes on G (and from the above definition it follows that a supervenes on b).

Further, in the light of what I have said about symmetry above, if any event E supervenes on some other event F, and F supervenes on E, then E is the same event as F. By analogy with the notion of proper subclass, an event G could be said to *properly supervene* on H iff (i) G supervenes on H and (ii) G is not identical to H.

This definition makes it possible for events involving different physical objects to supervene on each other, just in case the physical object which the supervening event involves itself supervenes on the physical object which

the subvenient event involves. This satisfies the intuitions underlying both the Kim's and Menzies' desiderata; on the one hand it allows events involving different physical objects to supervene on each other, and on the other hand it does not allow events to supervene on some part of the world which is not itself part of the physical location of the event. It would be very odd if the event of, say, a conference agreeing to a resolution, were able to supervene on some event which was not itself the property of something which was subvenient to the conference at that time. It may supervene on the physical state of the participants of the conference, but surely not of, for example, the road system in Hanoi. At least this is true of *intrinsic* properties; do relational properties create a problem? Suppose the constitution of the some town planning conference where such that a resolution could only be passed if the road system of Hanoi had been reconfigured in a certain way. In these cases we could say either that a relational property of the conference supervenes on a relational property of the physical state of the participants,⁶ the conference supervening on the conjunction of the physical states of the participants, or else that a straightforward property of the conference supervenes an a straightforward property of the physical states of the participants and various environmental factors including the configuration of the road system, and the conference supervenes on the physical states of the participants and

⁶This is acceptable as a doctrine about ontology of events; not, however, about explanation—at least not good explanation. The relational properties of the physical states of the participants are bona-fide states in ontology, but they are hopelessly gerrymandered and so do not make for good explanantia—which is just to say that the kinds of taxonomies that are devised at(say) the biological or physiological level, are not designed to neatly account for properties of conferences.

various environmental factors including the configuration of the road system.

7 Varieties of Supervenience and Menzies' Counterexamples

Because of his dissatisfaction with the features of Kim's version of supervenience which are outlined above, Menzies goes on to outline four new varieties of supervenience. For each of these he provides a purported counterexample to the causal reductionist thesis. In the light of my defence of aspects of Kim's account, and with my modification of it in mind, I hope to show that reconstrued with a notion of supervenience which depends on multiple realization, each of the purported counterexamples to causal reductionism in each of Menzies' varieties of supervenience is in fact no counterexample.

7.1 Circumstantial Supervenience

The first of Menzies' kinds of event-supervenience is *circumstantial* supervenience. On Menzies' account, what distinguishes this kind of supervenience is that the relationship between the supervenient event and the subvenient event depends on some extraneous circumstance. He cites as an example the supervenience of Xantippe's being widowed on the death of Socrates. This supervenience would only seem to hold if Xantippe is Socrates' wife.

He dismisses the intuition that this kind of supervenience is capturable by supposing that Xantippe's being widowed supervenes on Socrates' death just in case some extraneous circumstance—in this case Xantippe's being Socrates' wife—holds. This is because it seems that it will not capture the right kind of *dependence* between the supervenient and subvenient event. For

consider; while this allows Xantippe's being widowed to supervene on Socrates' death in virtue of Socrates' death, it *also* allows Socrates' death to supervene on Xantippe's being widowed. This is because there is some extraneous event—namely Xantippe's being Socrates' wife—which together with Xantippe's being widowed, entails Socrates' death. Thus the two events would, on this model, be *interdependent* in an objectionable way.

He remedies this by some considerations of intrinsicity. I shall not examine this proposal; rather I will diagnose how the problem arises, show that it does not arise on my account of supervenience, and that the counterexamples to causal reductionism he provides for this kind of supervenience do not go through on my version.

The root of the problem is that on Menzies' account, events with a circumstantial component should supervene only on the narrow or non-circumstantial parts of their base. Menzies' motivation for this is in the philosophy of psychology—he wants psychological terms to supervene on narrow facts so as to make them, on his lights, causally efficacious. I argue in Ch. 6 that supervenience can and should be on the circumstantial factors as well as narrow factors, at least in the psychological case. Here I am content to show what advantages of simplicity accrue if supervenience is on both the narrow and the circumstantial component. One of these advantages is that this particular problem does not arise.

Suppose that, rather than Xantippe's being widowed supervening on Socrates' dying *given* that Xantippe is Socrates' wife, Xantippe's being widowed supervenes on Socrates dying *and* Xantippe's being Socrates' wife. Now Socrates' dying *and* Xantippe's being Socrates' wife obviously does not supervene on Xantippe's being widowed, since for that to be the case Xantippe's being widowed would have to entail Socrates' dying and

Xantippe's being Socrates wife, and it does not do so. In worlds where Xantippe was married to Plato, Xantippe could be widowed without the conjunction of Socrates dying and Xantippe's being Socrates' wife holding. So the objectionable interdependence would be defeated. On the other hand, Xantippe's being widowed and Xantippe's being Socrates' wife would supervene on Socrates dying and Xantippe's being widowed *and vice versa*. Is this an example of the objectionable interdependence? No: I showed above that non-symmetry is an acceptable property for the supervenience relation to hold in virtue of reflexivity's holding. This reveals an interesting possibility; what if all cases of symmetry are cases of reflexivity? In particular what if the event of Socrates dying and Xantippe's being his wife *is the very same event* as Xantippe's being widowed and being married to Socrates, under different descriptions. This is still in the letter of the law of Kim's nominalization account (although it sits more comfortably with Lewis's sparser events of [Lewis 1986a]), since there will still be as many events as there are realized properties, just perhaps not as many *distinct events*. If we make event inter-supervenience a sufficient condition⁷ for event identity in this way, then this prevents us from counting too many events. It also emphasizes the rôle of multiple realizability in non-trivial supervenience. If one event is multiply realizable at a lower level, then it non-trivially supervenes on its realizer. If there is only one possible realizer at the lower level, then we may as well allow this to be a case of trivial supervenience, i.e. identity.

⁷I shall have more to say about identity conditions for events later. It must by now be obvious that on my strong causal reductionism, the traditional criterion of event identity—identity of causes and effects—cannot be right for macroscopic events for there are, properly speaking, *no* causal relations between them.

So no special kind of supervenience is required to accommodate these cases where circumstantial facts are involved. I shall now examine the alleged counterexample in this category.

7.1.1 The Counterexample

Menzies' general strategy is to attempt to show that two events, each of which supervenes on the same base event (while being neutral about whether there are *ultimately* basic events) can have different causal rôles. If they do, then they could not have their causal rôle in virtue of the causal rôle of the base event. Obviously an account, like mine, which denies that macroscopic events *have* causal rôles has an easy answer to the claimed differences in their causal rôles, but that would be *too* easy. I will couch my reply in the more usual language of macroscopic causation.

Menzies' first example is one of a footrace:

Fred runs 100 metres in a footrace in 10 seconds and in doing so outruns the competitor on his left, but is outrun by the competitor on his right. Here his outrunning his competitor on his left and his being outrun by his competitor on his right circumstantially supervene on his running the race in 10 seconds in virtue of different circumstantial facts about the running times of the other two competitors. Yet we might say that the cause of his losing the race was his being outrun by his competitor on the right, not his outrunning the competitor on his left. [Menzies 1988 p. 15]

The claim here is that two higher level events—being outrun by the competitor on the right and outrunning the competitor on the left—each supervene on the same base event, but differ in what they cause. I have two responses to this.

First, this alleged difference in causal rôle from the base event only occurs on the assumption that the supervenience bases for the two events are indeed the same. If, as is argued above, it is preferable to allow the

supervenience of higher level events to be on a base which includes the circumstantial factors, then this counterexample is defeated, since the two higher-level events can derive their causal rôles from their different supervenience bases.

Second, even supposing that the bases are the same, Menzies' general strategy in these counterexamples is too concerned with the *actual* realization of the macroscopic events. Another way to put this is that perhaps the thesis that higher level events do not get what looks like their causal rôle from their actual supervenience base in each case is not a strong enough claim to count against the view that actual causation proceeds only at the microscopic level. For consider: on my account what justifies talk of higher level objects or events is that they are multiply realizable. What prevents a higher level event from being identical with a lower-level event is that they do not co-supervene, so there must be *other* lower level events which, if they were realized, would entail the realization of the supervenient event. I shall call such a class of events the *subvenience class*. So the difference between being outrun by the competitor on the right and the (narrow) supervenience base are other situations, where the narrow supervenience base was different, which if they were realized, would still entitle us to say that the same high level event was realized. But this is compatible with the claim that in each of these cases the only causation which proceeds is at the low level. The talk of high level causation groups classes of possible causation together in ways which have relevant similarities for particular purposes. Because of the pragmatic function of such grouping, it counts as *explanation*, but the *grouping* need not count as causation. There is a subvenience class of events which would be realizations of outrunning the competitor on the left, and there is a subvenience class of events which would be realizations of being outrun by the competitor on the right. There

is at least one member of one of these classes which is causally different from any member of the other class; this is what the explanatory difference amounts to, but that is hardly any reason for denying that the causal work is done at the micro level.⁸

7.2 Logical Supervenience

The next of Menzies' kinds of supervenience is *logical* supervenience. Logical supervenience is supposed to hold between events when one event depends on another because the base event logically entails the occurrence of the supervenient event, but not vice-versa. Examples are supposed to include Jones' saying 'hello' supervening on his saying 'hello' loudly; a light ray's having some wavelength supervenes on its having a determinate wavelength of 1080A and so on. In general 'a determinable state logically supervenes on a determinate state; and a disjunctive state logically supervenes on its occurrent disjunct.' [Menzies 1988 p. 9].

It is only with Menzies' requirement that in some cases there can be supervenience on a base state *due* to some circumstantial factor (see previous section) that logical supervenience is distinguishable from other kinds. For consider what classical entailment of the occurrences requires (and I take it this is the kind of entailment Menzies has in mind): only that in every world where the subvenient state occurs, the supervenient one occurs as well. This certainly holds for Jones' saying 'hello' loudly and his saying hello; but on my version of supervenience it holds for *all* the earlier cases of supervenience. For if the circumstantial facts are included in the subvenient event, *whenever* the subvenient event occurs so does the supervenient. To

⁸This is much the same notion as the program explanation of [Jackson and Pettit 1988]. In their terminology the two high level events program for different causes.

take an example from the last section, if Xantippe's being widowed supervenes on Socrates' death *and* her being his wife,⁹ then in *every* world where Socrates dies and Xantippe is his wife, Xantippe is widowed (though there may indeed be worlds where Socrates dies and Xantippe outlives her husband, Thrasy machus).

So on my version of supervenience there is no special class of logical supervenience. All supervenience would have this feature. Nor does the purported counterexample to causal reductionism have any sting

7.2.1 The Counterexample

The counterexample in this case, due initially to Goldman [Goldman 1970 p 36?] is again that of saying 'hello':

Smith says 'hello' loudly and abruptly. Both actions logically supervene on the same event, namely his saying 'hello' loudly and abruptly. But it might be that his being tense caused him to say 'hello' abruptly, but not to say it loudly. Conversely, his saying 'hello' loudly caused Fred, who is hard of hearing, to pay attention, whereas his saying 'hello' abruptly did not. [Menzies 1988 p. 15]

The response to this, unsurprisingly, is much the same as to the previous case. Saying 'hello' loudly can be realized in many ways, as can saying 'hello' abruptly. Not all of the realizations of the one event have the same causal rôles as the realizations of the other. In particular, all of the realizations of Smith's saying 'hello' loudly would cause Fred to pay attention, whereas not all of the realizations of his saying 'hello' abruptly would do so. This difference is the content of the intuition that the supervenient events have different causal rôles; but in any such case the

⁹There are of course a number of simplifying assumptions; that she is his *only* wife, for example.

actual causal work would be done by whatever the actual realization is. Which makes it compatible with causal microreductionism, which is just the claim that it is whatever is the further realization of *that* event at the right microstructural level, which does the causal work.

7.3 Typical Causal Rôle Supervenience

This is the relationship which holds between functional rôle states and their occupier states; a typical example is the supervenience of the fragility of glass on some molecular structure. Menzies requires that subvenient events in this class have more fundamental properties than supervenient events.

I do not think that this kind of supervenience—while it may be a proper subclass of supervenience on my definition—has any special features which make a counterexample to causal reductionism go through.

7.3.1 The Counterexample

The first of Menzies' counterexamples is due to David Lewis:¹⁰

The high conductivity of a metal, its ductility, and its opacity and distinctive lustre are explained in terms of the cloud of free electrons which permeates the metal and holds the atoms of the metal in a solid state. In other words, the cloud of free electrons occupies the typical causal rôle associated with a metal's conductivity and its opacity. Nonetheless, on a particular occasion when a current has passed through a piece of copper and produced some effect, we would say that the effect occurred because the copper is a good conductor, not because it is opaque.[Menzies 1988]

A reply to this putative counterexample along the lines of the responses given so far would be that conductivity and opacity are multiply realizable;

¹⁰This was a suggestion made when the paper was given at the Brisbane conference of the AAP.

and not all realizations of opacity would give rise to conductivity. There is some difference in this case, though: perhaps all realizations of conductivity *are* realizations of opacity and *vice versa*. If it a consequence of the laws of nature being as they are that all conductors are opaque *vice versa*, then I would be committed to the view that conductivity and opacity are different names for the same property. As it happens, this is not true: saline solution is a conductor, but is not opaque, and rubber can be opaque, but not a conductor.

The example could, however, be strengthened against these objections. Suppose one were to say that we were not interested in the too general properties of opacity and conductivity, but interested in the special properties of *metal-conductivity* and *metal-opacity*. These properties are properties which metals have in virtue of the electron cloud; here I would agree with anyone who pushed this extension to the counterexample. Both the event of having metal-conductivity at *t* and the event of having metal-opacity at *t* supervene on the event of having a certain electron configuration at *t*. But with these properties, it really is the case that anything which is metalopaque is a metalconductor and *vice-versa*. In which case, if we want an explanation of the intuition that it is metal-conductivity which is causally responsible for the current's being able to pass through copper, it cannot lie in what the different causal rôles of other realizations are.

So in every physically possible world, anything which is metal-opaque is a metal-conductor. I will bite the bullet and say that metal conductivity is the very same physical property as metal-opacity, and so the event of *X*'s being metal opaque at *t*. is the same event as the event of *X*'s being a metal-conductor at *t*. So what of the intuition about the differences? One response is to say that with such specialized properties as metal-conductivity and

metal-opacity, we *have* no reliable intuitions; and the intuition which one may have that there is a difference comes from the case of conductivity and opacity, where microstructurally different causal rôles may be exhibited by the realizations of the different properties. But there is another response; note that only in every *physically* possible world are all metalconductors metalopaque; this is enough to make them the same physical property in one sense. But surely not all metalconductors are metalopaque in every *logically* possible world; and perhaps intuitions about asymmetry of the terms comes from this. Such an intuition seems, however, to be one that would come from an incomplete conceptual shift: if the properties of metalopacity and metalconductivity actually *did* come into linguistic currency because they served some useful rôle, and it was discovered that they were always realized by the same basal property, then I do not think that this concern about logical possibility would persist, as it is likely that we would take it that it is an empirical discovery that they are the same property, and by taking this to be the case, come to *mean* the same thing by them, thus excluding the logical possibilities in question [see appendix on kind terms].

7.4 Mereological Supervenience

The last kind of supervenience which Menzies distinguishes is *mereological* supervenience. This is simply where an event supervenes on a complex event, the components of which are *parts* of the whole event. The first of his examples should make this clear; the event of a batallion's advancing in a battle, mereologically supervenes on the complex event consisting of each soldier of the batallion advancing on the enemy. He gives further examples: that of the supervenience of a body of gas's having a certain temperature on the complex state of each of the molecules having a certain kinetic energy,

and the supervenience of a macroeconomic state on a complex microeconomic state involving individuals.

Menzies then goes on to give a definition of this kind of supervenience:

a's having *F* *mereologically supervenes* on the complex event of b_1 's having G_1 &...& b_n 's having G_n if and only if (i) b_1 ..., b_n are constituents of a ; and (ii) the proposition that b_1 has G_1 &...& b_n G_n , together with the appropriate bridge laws, entails that a has F . [Menzies 1988 p.11]

A revealing thing about this definition is that, like Menzies' other definitions but perhaps more explicitly, while it *allows* of multiple realizability as a special case, it in no way makes a *feature* of it. The definition is also a definition of a necessary condition for identity of wholes with the sum of their parts, with no explicit mention of what can be true of supervenient wholes with are *not* identical with the sums of their parts—namely that there are *other* mereological sums which would equally, in Menzies' language, entail the relevant supervenient property together with the appropriate bridge laws. This is perhaps what disguises the work played by multiple realizability in accounting for the appearance of differences in causal rôle between different events with the same supervenience base.

Once again it is unclear what is supposed to be distinctive about this kind of supervenience which makes the putative counterexample go through—or fail to do so—in a distinctive way. Indeed it is not clear that mereological supervenience is especially distinctive or even coherent. While there are some events which it seems more natural to suppose supervene on their parts than others—for example because they have more than one part, or because there is a natural taxonomy of events at the subvenient level which makes it more usual to think of them as a number of distinct events rather than as one composite subvenient event—no motivation is provided for supposing that this gives rise to particular problems for causal

reductionism. On the coherence front, while one can see why supervenience *might* be a notion useful for considering the relationship between a whole and its parts, when we talk of the *complex event* of having the mereological sum of the constituent events, I am inclined to the view that the relationship between this *complex event* and the apparently supervenient event is one of identity—unless of course the supervenient event is multiply realizable. But in that case the ‘constituent’ events are not *parts* in any ordinary sense of the supervenient event; the relationship they have to it is of being parts of the subvening event. The relationship between the complex event and its parts severally in the non-multiply realizable cases, of course, is just mereological; there is no need to import supervenience into *that*.

7.4.1 The Counterexample

The purported counterexample follows much the same pattern as the previous ones; on Kim’s theory of events, two events can be distinguished which, while supervening on the complex basal event, seem to differ in causal rôle.

Suppose that a body of gas undergoes a sudden and steep rise in temperature. Within Kim’s theory of events we can distinguish the event of the rise in temperature from the event of the sudden, steep rise in temperature. Both of these events, however, supervene mereologically on the same actual increases in kinetic energy of the constituent molecules. Yet we might say that the glass containing the gas shattered not because of the rise in temperature, but because of the sudden, steep rise in temperature. [Menzies 1988 p.16]

Once again, the justification for claiming that there are supervenient events which are not identical with the base event is that they could be multiply realized. The intuition about different causal powers is explained by the consideration that some realizations of the steep rise in temperature would not be sudden; and thus the realizations at the micro level of the steep but

not sudden rises in temperature would not cause realizations at the microlevel of the glass shattering. Thus it is straightforwardly no counterexample to the causation at the bottom level of nature thesis.

8 Explanation and Supervenient Causation

The common theme in most of what I have had to say about Menzies' putative counterexamples is that the intuition that there could be different events which supervene on the same basal event, but which differ in causal rôle, can be accounted for by reflecting that the relations that hold between supervenient events do give us more information than (or different information to) information about what the actual single case cause is. They tell what kinds of causal relations might or might not hold if the supervenient events were realized in different ways; but this is consistent with any *causal* story being, strictly speaking, at some as yet unspecifiable low subvenient level of reality. In Jackson and Pettit's [Jackson and Pettit 1988] terminology, supervenient events could *program-explain* each other, rather than actually causing each other. Menzies' attempt to refute causal reductionism founders in part because of its concentration on the *actual supervenience base* of events. When one realizes that the extra information given by relations between supervenient events might not be about *actual* causal differences, but rather be about *other* possible instantiations, then it becomes clear that Menzies was looking in the wrong place for the microstructural foundations of the differences between events supervening on the same base. The differences may not be actual; but they may be microstructural for all that.

My claim is that having this notion of explanation removes the *need* for making causal claims about the relationship between macroscopic objects or events. It is certain that they can explain each other; is there any reason to

suppose that they cause each other? At this point I have only addressed half of the opposition; the claim that macroscopic causation is *independent* of causation. So far I have been in agreement with Kim; whatever the relationship is between macroscopic events, it is dependent on microscopic events. It remains to argue that the relationship between macroscopic events is not causal, and that it is not a mere terminological nicety to call the relation explanatory rather than causal.

9 Against Supervenient Causation

Jaegwon Kim [Kim 1974] and [Kim 1979] argues for causal reductionism of a sort. By this he means the view that all causation is *dependent* on microcausation. The relationships between macroscopic entities are causal; but they are examples of *supervenient* causation which depend on subvening causal relations.

I reject this claim; roughly because so-called supervenient causation turns out to be so very different from basal causation that it requires a separate justification from basal causation. Since the notion of explanation is already to be had, and since it captures the epistemic and pragmatic elements which supervenient causation turns out to have, and since supervenient 'causes' are at least candidate explanantia, I will argue that supervenient causation is best taxonomized as explanation, and that causation is best reserved for whatever basal transactions actually perform the work in the universe. That notion may have to be a primitive.

Menzies' arguments construed as arguments against causal reductionism, when they are defused by considerations above, become powerful weapons *for* causal reductionism. This is because combined with the view that all causal relations between supervenient events depend on subvenient events, Menzies' arguments show that whatever Kim's

supervenient causation is, it is not *causation in the very same sense as subvenient causation*, if it is meant to depend on the actual realization of the supervenient event. There are two possibilities; either the so-called causal rôles of supervenient events are exactly the same as the subvenient causal rôles, in which case the supervenience involved is the special case of identity; and talk of supervenient causation is idle. Or else there is a relevant difference in rôle; in which case either this kind of causation is not dependent on the causal rôles of the actual realization¹¹ of the supervenient event; or else it is no kind of causation at all. This last possibility is what is required to keep the intuitions that causation depends on actual microstructure, and that there are important pragmatic differences between the kinds of apparently causal relations that exist between events which supervene on the same base.

9.1 Explanation and Causation

In the last section I argued that what Menzies' counterexamples *do* show is that it is not possible to hold *both* that all or most of the relationships

¹¹Two distinct though related issues can get entangled here: microreductionism and supervenience. If you thought that causal roles of supervenient events did not depend on their supervenience bases, *and* you thought that supervenient events do have causal rôles, you might still think that how the world is causally is determined by the micro-level *but not just the microlevel of the actual world*, because it is micro-level differences in *other* possible worlds make a difference to the causal description of the actual world. It is this feature which tempts complete reductionists such as John Bacon to argue that supervenience is just reduction to disjunction. This is, even if it is formally true, surely not in the spirit of reductionism, since it should not matter what is included in a reductive base other than the true disjunct—the actual realizer. In addition, there is no reason to suppose that our scientific theories about the microstructural levels will motivate the kinds of disjunctions required at the microstructural level in terms that advert only to microstructural features (c.f. ch 4). Another way of saying this is that even if supervenience is *reductio* to disjunction, reduction to disjunction is not reductionism.

between supervenient events which are usually called causal *are* causal, and that all causality at supervening levels depends on subvenient causation. There remains one possibility; accept everything I have said about the relationship between supervenient events, and characterized as explanation, but insist that it is *precisely this* that is meant by causation at higher levels, and thus that higher level causation does not depend on lower level causation.

Menzies has three arguments against the view that it is explanation which best characterizes the relationship between these high level entities. They are (i) that no good ground is provided for the distinction between explanation and causation, (ii) that the violation of usage which would result if most of what we commonly call causation turns out not to be so is objectionable, and (iii) that these higher level causal claims are supported by counterfactuals.

(i) It would be tempting to say that the distinction between explanation and causation does not *need* independent motivation; this motivation is motivation enough. But I shall resist. One motivation which is very powerful is the need to account for the rejection of a lot of obviously causal information as non-explanatory. As is argued in Ch 2, a substantial pragmatic apparatus is required to make sense of the contextual explanatory relevance or irrelevance of causal information. On the other hand, just the contextual pragmatics plus the right kind of probabilistic relations do not do a good job of identifying what causation is, as is argued in Ch. 3.

(ii) Violation of usage is not objectionable in itself; if it is discovered that what used to be called macro-level and micro level causation have very different properties (one is a physical transaction, say, and the other *merely* counterfactual supporting) then we can expect usage to be disrupted for a

while—popular usage may, harmlessly, never change. In any case, it is not clear that usage is so clear on the point. Causal and explanatory language is intermingled, sometimes clearly one, sometimes clearly the other and sometimes ambiguous. Claiming that ‘because’ is unambiguously causal seems almost disingenuous; it is the paradigm beginning of an answer to a why-question.

(iii) The counterfactual supporting nature of explanations need not detain us too much. I sketch reasons above why this should not be taken as *critical* of causation. And in the case of these explanations, we do not have to assent to astonishing global coincidences to deny that counterfactual supporting explanations are not themselves causal, because the counterfactual supporting work is done by the *organization* of information about causal processes in various circumstances.

10 Explanation and Event Identity

I indicated above that my view has problems on the traditional account of event identity. If events are individuated by having the same causes and effects, and if there are in general no causal relations between macroscopic events, then all macroscopic events are identical since they both have the null set as their cause and their effect.

The response will come as no surprise; higher-level events are individuated by information about the causal rôles involved in all their possible realizations which have causal rôles.¹² If all an event’s possible

¹²A *direct* realizer of an event may have no causal rôles because it is also of too high a level; indeed this will be usual. In which case we must turn to the realizations of the realization. I take these to be realizations of the higher level event as well because of the transitivity of supervenience.

realizations at the level at which causation occurs have the same causes and effects, then the two events supervene on each other and are hence identical. It is differences between events in this regard which allow them to have different explanatory rôles; indeed you could say that events are individuated by what they may explain or could be explained by.¹³ Certainly *interestingly* different events will be exactly those which actually *do* explain different things or are explained by different things.

11 Summation

A Kim-like theory of events has been defended, except that events are more sparse than in Kim, while less sparse than in Davidson [1980a]. They are more sparse than in Kim because supervenient events which are not multiply realizable are identical with their subvenience base, and events all of whose realizations are identical, are themselves identical. Thus multiple realization is crucial to the notion of supervenience; while non-multiply realizable events may share the supervenience relation, this is the limiting case of event identity. What is *distinctive* about supervenience is the possibility of multiple realization. Also event supervenience is found to depend on object supervenience [discussed in ch 4]. Thus properties of events can supervene on properties of different objects provided the first object supervenes on the second. Events also have no causal relations with each other, unless they are at whatever level of nature is found to be the physical base of causality. They do, however have *explanatory* relations. These are relations which give us either information about, or information

¹³I express it in this way, rather than saying by their explanations and what they explain, because it is possible that, given the pragmatic account I have, that while they might be *candidates* for explanation, they might not actually be involved in explanations because no-one has the purposes which generate the right choice classes and so on.

which is true *in virtue of* causal relations amongst the realizations of the events. A theory of explanation which relies on the notion that explanations are pragmatically based organizations of possibly counterfactual information about causes, allows intuitions about causal reductionism and the physical basis of causality to be accommodated in ordinary talk about macroscopic causation, by analysing it as explanatory information.

Chapter Six

The Explanatory Rôle of Folk Psychology

Table of Contents

1	Introduction	192
1.1	Folk Psychology	194
1.2	Cognitive Science.....	196
1.3	The Potential Clash.....	196
1.4	A Taxonomy of Relationships.....	202
1.4.1	High Level / Low Level Axis	202
1.4.2	Broad/Narrow Axis.....	204
1.4.3	The Revisable/Unrevisable Axis.....	206
1.4.4	The Manifest/Non-Manifest Axis.....	207
2	Positions on the Relationship.....	207
2.1	Proto-Science	208
2.2	Voodoo.....	209
2.3	Compatibilism.....	211
2.3.1	McCulloch.....	213
2.3.2	Normalizing Explanation.....	215
3	Explanation at Higher Levels and Broad Psychology.....	218
3.1	Program Explanation and the Doppelgänger Challenge	219
3.2	An Objection and a Response.....	222
3.2.1	The Structural Realist Version	223
3.2.2	The Rôle Attributive Version.....	224
3.3	A Moral To be Drawn By the Broad Stater.....	225
3.4	The Challenge.....	229
3.5	Synchronic and Diachronic Broad States.....	230
3.5.1	Synchronic Broad States	231
3.5.2	Diachronic Broad States.....	232
3.6	An Objection.....	234
3.7	The Rôle of the Synchronic Environment.....	235
3.8	A Twin-Earth Example.....	235
4	External Functionalism	238
4.1	Some Candidates for the Rôle.....	240
4.1.1	Problems With the Simple Solution.....	241

4.1.2	An Evolutionary Constraint.....	242
4.1.3	A Dispositional Account.....	243
4.2	Why External Functionalism is Not Behaviourism.....	245
4.2.1	The Case if Fodor is Right	247
4.2.2	Some objections by Steve Stich.....	249
4.3	The Return of the Axes.....	255
4.3.1	High Level/Low Level	255
4.3.2	Broad/ Narrow Axis.....	256
4.3.3	The Revisable/Unrevisable Axis.....	256
4.3.4	The Manifest/Unmanifest Axis.....	257
4.3.5	The Explanatory Motivation.....	257

1 Introduction

People, it seems, are always explaining one another's behaviour by means of beliefs and desires. It is a habit that the folk have had as far back as recorded history and beyond. Even legends in the oral tradition feature explanations of the behaviour of deities and the hapless mortals that were supposed to suffer under them in terms of their beliefs and their desires, malevolent or otherwise. Yet this practice—folk psychology—is under attack; a new approach to explaining behaviour—scientific psychology or cognitive science—has emerged. Cognitive science is one of the few growth industries in the academic world apart from weapons research. It is an interdisciplinary field which includes work in psychology, philosophy, logic and computer science. The programme is very exciting: to find out how the mind really works; to lay bare the nature of thought through a thoroughgoing scientific, theoretical, and empirical investigation of human psychology. The controversy is about what the relationships could be between the kinds of results which cognitive scientists may eventually produce, and the everyday manifest image of the mental as it presents itself to us: that assortment of theories, ways of talking and ways of understanding commonly called folk psychology. At least some of its practitioners believe that it is in competition with folk psychology. The San Diego eliminativists¹—Steve Stich and Patricia and Paul Churchland [Stich 1983] [Ramsey, Stich and Garon 1988] [Churchland 1981] [Churchland 1986]—go so far as to claim that there is *no*

¹An early discussion, in effect about the difference between reductionism and eliminativism, (though for the case of sensations rather than propositional attitudes) is to be found in [Rorty 1965].

explanatory rôle for folk psychology, and that its usual objects of discourse, beliefs and desires, simply do not exist.

Of course exactly what the difference is between folk psychology and cognitive science is a little unclear, as there does not seem to be anything univocal about what the folk believe about psychology; nor for that matter about what cognitive scientists believe about it.

In this chapter I examine the various axes of difference between the various formulations of folk psychology and cognitive science, and conclude that the basis for the difference is different views on the nature of psychological explanation. The difference between the Eliminativists and Fodorians is explicated in terms of a difference in their location of the causally efficacious psychological properties, while sharing an assumption that psychological explanation endeavours to explain by discovering those properties. The ground is laid for a rejection of Fodor's proto-scientific view of beliefs and desires in the next chapter, and I go on to discuss the range of positions I call *compatibilism* about folk psychology, including both an *instrumental* kind (Dan Dennett's [Dennett 1978a], and Gregory McCulloch's [McCulloch 1986]) and a more robustly realist kind (Philip Pettit's [Pettit 1986], and Frank Jackson and Philip Pettit's [Jackson and Pettit 1988]). While I am in sympathy with much of the spirit of these defences, none of them seems wholly satisfactory, and I argue for a version of compatibilism about the explanatory aims of a particular disambiguation of folk psychology and cognitive science which conforms to the constraints on explanation laid down in the earlier chapters.

1.1 Folk Psychology

Folk psychology is a term which has been used both sneeringly and as a term of praise, but what is it? It is the explanatory and predictive strategy

which, it seems, people have always used in their dealings with each other. It trades in propositional attitudes—beliefs, desires, hopes and so forth that are pre-analytically intentional—that is, they are *about* things. They are called propositional attitudes because they express attitudes to propositions: one can believe, desire or hope any proposition. Propositional attitudes can vary along the axis of the proposition—one could believe that P, believe that Q and so on. They can also vary along the axis of the attitude to the proposition—one can believe that P, hope that P fear that P etc.. They can, naturally, vary across both axes. Of course like any folk theory there is no uniformly applicable account of it, and the formulations of folk psychology in the literature vary. But for a long time it reigned supreme in a quiet kind of way in the philosophy of psychology, particularly during the heyday of conceptual analysis. Even so radical a theory as Ryle's behaviourism [Ryle 1949] did not, on my reading, really deny the importance of folk psychological entities²: it just identified them with behaviour complexes. Not surprisingly, because no matter how much the analyses may have differed, the task was seen as analysing the psychological practice of the folk and getting out a theory which was a maximally consistent subset of them, or at least one which respected those elements of usage whose retention was most demanded by the analysing philosopher.

²Unlike that of Watson [Watson 1925]. The eliminativist/compatibilist debate was alive and well amongst early behaviourists. Watson seemed to believe that that compatibilism was an exoteric doctrine designed not to scare people away from behaviourism. The esoteric doctrine was that there was, properly speaking, 'no such thing as the mental'.

There are, though, plenty of elements which can be used uncontroversially to describe the theory. Action is typically explained by reference to beliefs, desires and intentions, thus:

(Q) Why did Merlin put himself to sleep?

(A) Because he believed that he was not living in an age suited to his talents, he believed that a time would come when his thaumaturgical techniques would be more appreciated and understood, and he desired to reap the benefits of living in such an age. Believing also that the long snooze would be an effective strategy, he formed the intention to implement the strategy and did so.

The details of how the intentions are supposed to manifest themselves in behaviour and of all the other complexities of belief-desire psychology belong to a vast literature which has little bearing on the main point, which is that for folk psychology much of the mental life of ourselves and of others is explained by these propositional attitudes, and they are taken to have *content*.

In dealing with each other as people we regularly ascribe beliefs, desires, wishes, hopes and so on to each other and believe (how hard it is not to use that word!) that they are hopes, wishes and beliefs about things outside ourselves. How does this square with cognitive science?

1.2 Cognitive Science

Cognitive science aims to give us what Clerk Maxwell³ called the 'particular go' of the mental. It is naturalistic; it is not concerned with merely analysing usage but wants to find out how the mind actually works.

It is not just concerned, however, with any old description of how it works. On a case by case basis, explanation of behaviour could be given by an account of the causal processes in the brain described at the molecular or even atomic level; and no-one (not even the Churchlands!) would think that such an explanation of a piece of token behaviour by physical states was a piece of cognitive science. Rather, cognitive science seeks to show how the capacities of the mind might be explained by hypotheses about *psychological* properties which, if it possessed them would account for those capacities.⁴ The further step is to show that these properties are in fact realized by minds. A completely successful cognitive science will do both of these. Of course exactly what makes something a psychological property is in this sense is up for grabs; the consensus amongst cognitive scientists seems to be that these properties are found at some level (see ch.4) higher than the organic chemistry of the brain, but that they should nevertheless be intrinsic structural properties of the brain.

1.3 The Potential Clash

There are, however, potential clashes between the research program of cognitive science and aspects both of the folk story about beliefs and desires,

³William James was fond of this phrase, he mentions it on p. 95 of [James 1975].

⁴This is, I think, roughly similar to what Cummins [Cummins 1983 Ch 1] means by property-theoretic explanation.

and some doctrines which, if not actually believed by the folk, are at least common amongst philosophers. The first is the most important clash, but the approach that will be defended here will also save the second.

To approach the second question first, much of the classical theory of the truth evaluability of belief is at least in jeopardy if folk psychology or something like it does not turn out to be a respectable enterprise.

It seems to be some part of the folk psychological view that our propositional attitudes are about the world. Indeed what they are about is what differentiates different propositional attitudes of the same attitude. For example this belief of Mary:

(1) Canberra is clean

is a different belief from

(2) Perth is clean.

This folk intuition is where a classical theory of truth of beliefs gets a foot in. On the Classical view, their content—what they are about—is given by their truth conditions. The truth conditions of the propositions believed are different. (1) is true just if Canberra is clean, and (2) just if Perth is clean. So they are different beliefs. There can even be cases where there does not seem to be anything relevant about the internal states of believers which distinguishes beliefs, but they are distinguished by the truth-conditions of the propositions believed. Consider the case of Mary and Jane. Mary lives in Perth, Jane in Canberra. Both believe 'the city in which I live is clean'. There seems to be at least *some* sense in which these beliefs have different truth

conditions; Mary's belief being true just if the city in which she lives is clean, Jane's belief being true just if the city in which *she* lives is clean.⁵

Since Putnam's 1975 paper "The Meaning of Meaning" [Putnam 1975] it has become something of a commonplace that facts about the intrinsic structure of brains may not determine meaning. We can, in principle, imagine a being in exactly the same physical state—a *doppelgänger* to use his picturesquely Gothic term—where every neuron is organized in the same way as in our Mary of the preceding example, and yet her environment is completely different, so much so that it is not Canberra at all that she has beliefs about, but rather some other city whether on a Putnam-like twin earth or some place rather more mundane. For the folk their beliefs are different in at least one sense (perhaps the folk have a narrow sense in which they are the same as well) and the Classical meaning of the propositional objects of the beliefs is different as well. One is talking about Canberra, the other about some other place. Truth is also implicated, because the truth values of the beliefs could vary in virtue of their differences in Classical meaning of their propositional objects: maybe Canberra is in fact filthy, while the twin city is impeccably clean.

The cognitive scientist is looking to explain behaviour. And it may seem that all that is relevant to psychological explanation of behaviour is

⁵It is no part of this claim that beliefs are *fully* individuated by truth conditional content. The claim is merely that the truth conditional content of belief seems to have a powerful rôle to play in the prediction of behaviour, in particular in the prediction of the consequences of belief-holding. This is compatible with the distinction in [Perry 1979] between belief-states and the propositional objects of beliefs, *both* of which are required for much of the explanatory work to which beliefs and desires are put. So, in this case one should, of course, not be committed to the view that there is nothing significantly *in common* between Mary's and Jane's beliefs.

whatever the causal processes are which go on in the head, and the states in virtue of which we can suppose that such processes will go on. Of course things outside the head *cause things to happen in the head*; and a full causal history of a piece of behaviour would mention them but explanation in terms of these things is not psychological. In my terms, psychological explanation, as cognitive science construes it, limits its choice class of explanans to facts about the structure of the brain. Jerry Fodor calls this doctrine *methodological solipsism* (see [Fodor 1981a])⁶. It is a doctrine which is shared not only by those who, like Fodor, want to keep much of the folk taxonomy while insisting that it can be captured by a story solely about syntactic tokens in the brain, but also by those who, with Stich and the Churchlands, think that the kind of story about the operations of the brain which will best explain behaviour will most likely be no *realization* of the folk notions, but will be an alternative to it which will render it redundant. Methodological solipsism gets its plausibility from the consideration that it is the internal workings of the mind under their best scientific description which cause and hence explain behaviour. Consider the twin-earth examples above which are supposed to show that narrow states do not determine meaning. The methodological solipsist can accept this, while saying that there are important similarities between Mary and her twin as well as differences, and that it is the similarities which explain behaviour. For one thing, if they have parallel histories, then they have displayed the same behaviour; and if these uniformities of behaviour are going to be uniformly explained—as they had better be if we want to capture as many

⁶In [Fodor 1987] ch. 2 there is a revised doctrine, *methodological individualism*, which plays much the same rôle. The difference is that methodological individualism rules out relational states of things as irrelevant to their causal powers only if they are not causally efficacious.

generalizations as possible—then the candidate for explanans is uniformity of narrow state, since there is no uniformity of broad state. If both Mary and her doppelgänger set out on a journey as a result of believing 'Canberra is Clean' we want to be able to offer the same explanation and say that it is because they have the same beliefs that they do this. A terminological trick, like saying that they have different beliefs but that there is uniformity of explanation since their behaviour is in each case explained by the fact that their behaviour is caused by the belief tokened by the sentence 'Canberra is clean' in their idelect, doesn't help much. So, according to this line of thought, it is the actual intrinsic state itself that does the causal work. If that state isn't the belief, then beliefs don't explain action.

A research programme which is committed to *only* these sorts of contents is going to have to take a long, hard look at the relationship between meaning and truth that the Classical theory prescribes. Fodor sees room for a theory of the relationships of subjects to the world which is independent of a theory of content, and combining such a theory with an account of narrow content would perhaps do some of the work of the traditional theory, but unless we can find some rôle for folk psychology, at very least the traditional unity of the theory of meaning and truth will have to be modified. A theory of truth of beliefs on such an account becomes rather hard to formulate. Something like the doctrine that a belief would be true iff the state of affairs which standardly held in the given context during veridical belief-holding in fact held, becomes a front runner; and it would be nice if it weren't even a starter given the smack of circularity that comes with the use of 'veridical' in this context.

Put in a nutshell, the problem is this: it is in virtue of their meaning that beliefs can be evaluated as true or false. But Putnam has shown that,⁷ to a greater extent even than was supposed before, meaning is not given solely by narrow factors; factors in the head. Now if propositional attitudes are individuated purely narrowly, then being a certain propositional attitude won't determine the meaning of the proposition. So being a certain propositional attitude won't give the truth conditions of the belief, so beliefs as such won't be truth evaluable without some further story.

But there is still more at stake, even if one has no brief for the truth evaluability of beliefs. The folk theory of beliefs and desires has been extraordinarily successful. Its predictive and apparently explanatory success has been much greater than that of any theory which has turned out to be an almost accidentally useful heuristic in the light of the true theory. All of the arguments in, for example, ch. 1 of *Psychosemantics* [Fodor 1987] for the indispensability of belief-desire psychology do not depend for their plausibility on the truth of Fodor's account of beliefs and desires. If Fodor's attempt to rescue them by a cognitive science which shows how they are systematically structurally realized by intrinsic architectural states fails (and I provide some arguments to suggest that this is a possibility in ch 7), there will be no comfort for beliefs and desires in the arms of the San Diego Eliminativists. So an account of beliefs and desires which reconciles their explanatory power with whatever the true account of the architecture of the brain is will be sorely needed.

⁷I argue in the appendix that although Putnam has shown that there is more determination of meaning outside the head than previously thought, there is a little less than *he* thought in [Putnam 1975e], and less than his acolytes still think.

1.4 A Taxonomy of Relationships

At this point it must be clear that there is a whole range of issues buried in the very general descriptions of cognitive science and folk psychology as a whole that we have given so far, and that characterization of the relationship between folk psychology and scientific psychology or cognitive science given at this stage is hardly adequate. A reason for its inadequacy is that the defenders of folk psychology, and those who urge its elimination in favour of a cognitive science, have many different goals and are defending or urging the elimination of a wide range of positions. The distinctions between them are often blurred in the literature. There are at least four axes along which views about propositional attitudes vary: I shall give an account of these, and the various positions should fall out as combinations of them. They are the *high level /low level* axis, the *broad/narrow* axis, the *revisable/unrevisable* axis and the *manifest/unmanifest* axis. Folk psychology as such does not get mentioned in this list since it is not clear to me how the folk psychologize. Certainly some of the combinations of positions on these axes which claim to be folk psychology are positions more subtle than was ever dreamed of in the folk's psychology. This is not meant as a criticism; perhaps such positions play useful rôles. Indeed, the rôle of this chapter will eventually be revealed as defending the explanatory indispensability of just such a position. It does, however, chasten the zealot who thinks she is defending the views of ordinary people against the incursions of philosophical obscurantism.

1.4.1 High Level / Low Level Axis

Most accounts of folk psychology place it at a more abstract level than whatever a scientific psychology will tell us about propositional attitudes. It is a difference between the *descriptive level* at which minds or brains must be

the same for them to have the same propositional attitude. In the sense of ch. 4, the terms of folk psychology are at a higher level than those of cognitive science.

Indeed most accounts of folk psychology tend to place it at a high level on this axis. In Fodor, Dennett, Stich, Jackson, Pettit and McCulloch there is agreement that that Folk beliefs and desires, if there are any, supervene at least in part on, but are not supervened on by, intrinsic states of the brain.⁸

There are differences in the question of at just how high a level folk psychology is located. Some accounts make the folk psychological a very high level of description, which will supervene on the narrow scientific but whose justification is on more narrowly instrumental grounds than functional. Thus, on this kind of account (for instance Gregory McCulloch's) although folk psychological descriptions are high level descriptions, and *important* ones at that, they are neither (in my terms) structural nor strongly functional (see ch.4). Other views [Fodor 1968 onwards] [Devitt and Sterelny 1988] maintain that folk psychology is at a slightly lower level, since what we are referring to in our folk psychological discourse are probably the propositional attitudes of the true scientific psychology. Most views on cognitive science, however, will put the level of description of the propositional attitudes of scientific psychology at a fairly low level by comparison. The High Church Computationalist story of Fodor is at the *highest* descriptive level of the various competitors for the prize of Best

⁸Except perhaps in the case of Fodor where the second clause may not hold, since plausibly he holds that the folk psychological states are identical with the syntactic structural realizers posited by the Representational Theory of the Mind. But in any case, folk psychological states are not at a *lower* level of description than the cognitive scientific ones.

Scientific Description of the Brain. Presumably the computational level supervenes on, and is not supervened on by the neurological level and possibly levels between. The Churchland's account of the mind/brain [Churchland 1981] [Churchland 1986] which eschews states which bear much resemblance to the propositional attitudes of folk psychology at all, and insists on the neural level as the important one for psychological explanation, will be at an even lower level, supervening more or less directly on the neurological or neuro-chemical.

1.4.2 Broad/Narrow Axis

This is the axis of difference between views of psychology which invoke *broad* states and the views which invoke *narrow* states. A broad psychological state is one which supervenes both on properties of the environment⁹ the subject is in, as well as intrinsic properties of the subject. A narrow psychological state, on the other hand, supervenes only on intrinsic states of the subject; on how things are "inside the skin", so to speak. Finding an *exact* border might be difficult (are we concerned with the brain alone, the rest of the body or do we include tools and extensions of the body?), but in practice this does not seem to be a real issue in the debate.

An example may help. Suppose we are trying to establish identity conditions for Fred's belief that he is holding that cup. If we require that the cup be *that actual cup*, and that Fred be causally related to it in the right way for him to have the belief, then we are talking about a *broad* state. If, on the other hand, we require only that Fred have a relevantly similar intrinsic

⁹The environment is construed generously here, so as to include causal relations, histories and so forth. I mean simply how things are outside the subject.

state to the one which causes him to behave as he does in virtue of his belief, then we are talking about a *narrow* state.

Some of the more interesting defences of folk psychological propositional attitudes to have appeared recently have been couched in terms of a defence of the indispensability of broad states. Philip Pettit in [Pettit 1986] argues for the explanatory indispensability of broad states, and I think the *motivation* for this is, at least in part, that he takes broad states to be closer to the folk notions of psychology that, if abandoned, would require a complete change in our view of ourselves as rational agents. Jackson and Pettit in [Jackson and Pettit 1988] have a further defence of the explanatory rôle of broad states, in terms of what they call *program* explanation (see ch. 4).

The four axes of difference that I am discussing cross cut each other, however, and not all views on folk psychology occupy positions in the resulting matrix that make them broad. I shall argue that Gregory McCulloch's views on folk psychology are at least *compatible* with its being narrow. Also, much of actual folk discourse about psychology is couched in terms that seem at least *prima facie* narrow.¹⁰ If, while I was dreaming, I believed that my bed was on fire, it had better have been a *narrow* belief, since it presumably was not, in any direct causal sense, about my bed. In addition, folk intuitions do not seem to rebel against the possibility of brains in vats having beliefs about the ordinary objects of folk discourse. So in actual folk discourse there might be a sense in which beliefs are *considered* to be narrow. But it is open to the proponents of the broad view to argue that the folk view *also* allows the broad view and it is this component of the cluster of folk notions which cannot be reduced to or replaced by a narrow

¹⁰I am indebted to a conversation with David Lewis on this point.

cognitive science. So a distinction between the psychology of the folk, and folk psychology seems to be in order. It may turn out that the most fruitful line of inquiry will be the defence of some position which may be *called* folk psychology, and which may preserve some important folk intuitions, but which will not preserve the psychology of the folk (if there is such a univocal thing, which I doubt), in its entirety. The other side of the same coin is that I do not see that there cannot be resolutely scientific psychological taxonomies which invoke broad states; what is in common between quite different species that display the flight response, for example, may well be nothing which can be captured any other way than as a broad state involving the creature, the danger, and a teleological story about biological function to cement it together [see e.g. Sterelny and Godfrey-Smith forth.]. So not all broad states need be *merely* folk states. Perhaps this is where the folk and scientific quests will meet; but that tantalizing suggestion must wait for its elucidation until the end of the chapter!

1.4.3 The Revisable/Unrevisable Axis

One of the cluster of intuitions which sometimes go together to comprise folk psychology is that our folk taxonomy of mental states somehow just *couldn't* be wrong. This is not, however, a necessary partner of the other views. A view like that of Gregory McCulloch's [McCulloch 1986] which relies on the alleged impossibility of our getting on in the world with substantially revised notions about propositional attitudes may require that folk psychology be pretty thoroughly unrevisable; but there is no reason why broad psychological states, if they are explanatorily indispensable, should provide their content in virtue of retaining the folk taxonomy of propositional attitudes. Perhaps they will; indeed I think that, roughly, that this is likely. It is, however an open question to be decided by the very criterion of explanatory indispensability which justifies the broad states. A

defence of folk psychology which relies on the apparent irrefutability of the presence of the Manifest Mind will have a high reliance on the irrevisability of the folk taxonomy. To the extent that such accounts rely on how things *seem*, they have to. Which brings us conveniently to:

1.4.4 The Manifest/Non-Manifest Axis

One last axis of variation is the extent to which a folk psychological rather than a cognitive scientific account relies on an introspective notion of *what it is like* to have a propositional attitude as criterial of having it. It is at least conceptually possible that a very high level, unrevised in taxonomy and broad account of propositional attitudes could be true and yet not yield the Manifest Mind: perhaps some doppelgänger could have such folkish propositional attitudes and not have the experience of what it is like to have them; if indeed there is anything which it is like to have them. This will be, trivially, impossible on the Manifest account. The relationship of propositional attitudes whose keep is earned by explaining behaviour on the one hand, and subjective experience will be kept, for the most part, in the too hard basket; but on occasion I will take it out, uncrumple it, and have a look. For present purposes it is enough to point out that this is another axis of variation which makes it impossible to pretend that there is a univocal cognitive science/folk psychology distinction.

2 Positions on the Relationship

Preliminaries out of the way, let us return to the first question about the relationship of folk psychology and cognitive science. There are broadly four positions which seem to have been adopted:

- (1) Folk psychology is a proto-science.
- (2) Folk psychology is voodoo.

- (3) Folk psychology is indispensable and cognitive science contributes *nothing* to our understanding of how psychological states, properly understood, are realized.
- (4) Folk psychology is indispensable in principle, and is doing something quite different from, and is therefore not in competition with, cognitive science. I call this range of positions *compatibilism* about folk psychology.

2.1 Proto-Science

The *proto-science* view is held most famously by Jerry Fodor [Fodor 1968, 1978, 1981, 1987]. The view is that folk psychology is like many other folk wisdoms: a remarkably good first step in systematizing our understanding of its domain. But it is only a first step: it will be replaced by a scientific psychology which will give both an analysis of how folk psychological states are structurally realized, and may revise the notions of propositional attitudes somewhat in so doing. There will be a two-way process: folk psychology will give an idea of the kind of things to look for in our cognitive science, but when we find them they may—or indeed probably will—turn out to behave a little differently from how we thought they would, but be similar enough to count as a scientific reduction of the folk notion rather than a complete replacement. This is supposed to be similar to other pre-scientific generalizations which, while crude, were able to yield reasonable results and started us off on the right track. Pre-scientific botany, for example, often conflated species and made numerous other errors, but nevertheless on the whole tracked real natural differences sufficiently well

that it was the investigation of the objects of the pre-scientific discourse which was able to lead to the development of the science.

The proto-science view of Fodor, or for that matter of Devitt and Sterelny [Devitt and Sterelny 1986], assumes that the true science that folk psychology is the prototype for will be a science of narrow states. What fuels this is, I think, a combination of views about explanation. Firstly (1) that an explanation must cite causally efficacious properties (cite the actual cause in the schema of ch 4), secondly (2) that high level states of systems can themselves be actual causes of the system's behaviour, and finally (3) that only intrinsic (or narrow) states of a system could be such causes. The argument against Fodor in the next chapter, and the argument for a certain kind of folk psychology that is not Fodorian, will depend on rejecting the first two claims, and showing that extrinsic properties satisfy the criteria of causal *relevance* outlined in part I.

2.2 Voodoo

The voodoo position, held *inter alia*, by the Churchlands, has an eloquent champion in Stephen Stich, whose book *From Folk Psychology to Cognitive Science* argues that perhaps there just aren't any beliefs: cognitive science has no room for the propositional attitudes of folk psychology at all. Just as Voodoo wasn't a proto-science and was just plain wrong, so too is folk psychology. There is just no such animal as a propositional attitude in a scientific psychology, for nothing could play the right kind of causal rôle. Just as the entities that Voodoo theorists were committed to turned out not to exist, so too will beliefs and desires turn out not to exist. It is a thoroughgoing eliminativism about folk psychology. Whereas the Fodorian plan is to *reduce* folk psychology to cognitive science, the San Diego Plan is to *replace* folk psychology with cognitive science, and argue that the

ontology of cognitive science contains nothing which is a realization of the things posited in the psychological ontology of the folk. I shall not deal with the state of the evidence for or against the truth of this view when stripped of the rhetoric about voodoo, since the rejection of a folk-psychology-like structure for cognitive science is not, I hope to show, incompatible with (4) above. Indeed I am roughly sympathetic with one construal of the line in [Ramsey, Stich and Garon 1988 forth.] that folk psychology as a theory about *structural states in the operations of the brain* is incompatible with various kinds of reasonable connectionist hypotheses about the functioning of the brain. In other words, that these kinds of connectionism might be incompatible with folk psychology as Fodor sees it. I shall, however, consider later Stich's objections to there being an autonomous rôle for folk psychology.

Eliminativism gets its explanatory motivation from the acceptance of (1) and (3) above, but with a more stringent attitude to (2). High level states do not count as causes for the Churchlands or Stich just if they represent functional descriptions that are physically realized; what their talk of a 'bottom up' approach to doing cognitive science amounts to is that high level states could only be causally efficacious if they are in my terms *structurally* realized. Which means that for some taxonomy of neural states to count as a realizer of, rather than a replacement for, the folk taxonomy, it would have to have independent scientific motivation in narrow terms. But the best account of structural states which are causally efficacious (somewhat misleadingly called syntactic states, in [Stich 1983], see ch. 7 footnote 6), according to this kind of cognitive science, does not share features which are supposed to be essential to the folk account. So the states which do the causing are not the folk states, and since they take it as essential to the folk notion that folk states are causes, there are no folk states.

I have no plans to offer comment on view (3). This might be held by idealists, dualists and wholesale epiphenomenalists.

My plan is to argue that the last position, Compatibilism, can be defended by taking a different view of the requirements on explanation. Both (1) and (2) above are rejected. Instead the constraints of part I are applied, and I show how folk psychological states which were functional states supervening on both the environment and the organism could be causally relevant to behaviour. It will be further argued that folk psychological states may not be *mere* functional states, but could be *strongly* functional states, thus sidestepping the worry that while it may be possible to construct a folk taxonomy of states which are physically realized, it might be a scientifically uninteresting, gerrymandered functional description. Before doing this, however, I shall examine some existing versions of compatibilism.

2.3 Compatibilism

The modern version of this view seems to begin with Dan Dennett's talk of *stances* in his paper 'Intentional Systems' [Dennett 1978b],¹¹ The idea starts with the observation that there are different ways of predicting, and therefore more generally of understanding and explaining complex systems. Three such ways are introduced: the *design stance*, the *physical stance* and the *intentional stance*. He takes a chess playing computer as an example. In principle, if one knew everything there was to know about the physical state of the make-up of the computer, one could, with a knowledge of physics,

¹¹Dennett now considers this paper superseded by "True believers" in [Dennett 1987]. Here he takes a line which he regards as closer to realism about propositional attitudes, roughly by being a realist about calculational *abstracta*. In his recent review of [Fodor 1987], however, [Dennett 1988] he seems to have reverted to instrumentalist type.

predict what it would do. It would be easier, though, to view it from a design stance, if one knew how it was designed, or even knew its program. One could then simulate its processes more simply than by the purely physical description—since you would have an idea of at what level one has to include detail to be sure of preserving the right features to be predicted. Crucially, it is only by making predictions based on this stance that it is possible to make sense of the idea of a malfunction—a malfunction will have occurred if a thorough design stance prediction is not borne out.

Either of these stances, though, will make prediction fairly cumbersome and in fact the only really effective way of predicting what the machine will do when it is your chess playing opponent, is to view it from the intentional stance. This is to regard it as a chess player that has certain information at its command. Rather than working through the *actual program* that is realized in it, you use your knowledge of *what the program is designed to achieve* to make your predictions. You predict what the chess-playing computer will do by asking what the best move would be, and, crucially, you can explain its previous moves as manifestations of chess-playing principles even if they are nowhere explicitly represented in the machine's program. Of course, one will often be wrong because of either deficiencies in the machine's implementation of the ideal player, the possibility that the ideal player is a non-unique notion or even a fuzzy one, or deficiencies in one's own grasp of the ideal player! Nevertheless, it may still be the only way to interact with such a machine in the situations for which it was designed, and as an explanatory strategy it will capture generalizations that might not be capturable at the physical or design level because of enormous differences in implementation.

The idea of the intentional stance, though, is one which is applied more full-bloodedly to humans who are taken to exhibit rationality. We predict

what they will do on the basis of taking an intentional stance to them. Thus folk psychology is instrumental—there aren't necessarily any entities which correspond to the folk psychological categories, but it is nevertheless going to turn out useful to attribute it to them as a kind of instrumental tool: calculational fictions, like lines of force, but necessary fictions for many purposes none the less.

This "fourth position" is developed rather differently (both from Dennett and from each other) in papers by Philip Pettit and Gregory McCulloch, both of which appear in [Pettit and McDowell 1986].

2.3.1 McCulloch's Patterns of Engagement

McCulloch takes over Dennett's notion of stances, calling them the Physical, Design and Intentional descriptive stances. He adds to these a set of what he calls *intervention strategies*, which differ according to which level one actually intervenes in a system: direct physical manipulation, changing the design or by reasoning with it. The supervenience relation is taken to hold between these relations. One cannot intervene at a higher level without doing so at a lower one, but not *vice versa*.

He combines these ideas of stances and strategies into what he calls a *pattern of engagement* (PE), which is a mixture of descriptive stances and intervention strategies with respect to some system. Folk psychology is taken to be one of these patterns; the pattern formed when the intentional descriptive stance is combined with the intentional intervention strategy. This means that we 'treat others as people' and 'take them seriously as thinkers'. It is compounded both of ways of regarding or understanding another or oneself, and ways of intervening. When using the folk psychological pattern of engagement one reasons with someone, for example, rather than trying to change their views by chemical means.

This makes the folk psychological PE radically different from the design explanatory one. It is the complex of behaviours which demarcate a way of interacting, rather than a way of coming up with design-explanatory accounts of why behaviour takes place. McCulloch makes the bold claim that folk-psychological entities turn out to be properties which the mind has. Not, though, properties which the mind has in itself, but rather properties relativized to a particular point of view—the point of view of a particularly pervasive pattern of engagement. Folk psychological propositional attitudes turn out to be qualities visible from within engagement. He emphasizes the Nagelish [Nagel 1974] point that it *is like something* to engage in the folk psychological enterprise, however this supervenes on the physical, and folk psychological properties are what appear when the mind is like this. Folk psychological properties are secondary properties, but properties nonetheless. If it turned out that vastly different physical systems could realize redness, that wouldn't make redness any the less a secondary property of them.

The problem with this position is while it is all very well to say that the mind has these properties, properties are to be had at very little price. What we want to know is whether we *need* these properties for anything. In particular whether we can explain anything with them which we cannot explain with what we already have and need for other purposes anyway. McCulloch shows merely that things *seem a certain way to us*; but we knew that already. What he doesn't provide a substantive argument for, is that they actually *are* that way—or even that in general if they seemed that way, then things would be so as to explain the kinds of behaviour we expect our folk psychology to explain. Once again there is no argument forthcoming that these strategies are not *merely* instrumental, nor that the usefulness of them explains anything about the system that is not already explained.

2.3.2 Normalizing Explanation

Philip Pettit's approach is rather different. He aims to show that there is a distinctive explanatory rôle for folk psychology construed as a theory of broad psychological states, and that broad-minded states are not dispensable in psychological explanation.

He claims that on a standard account of explanation (which he calls a *regularizing* account) wide psychological states are indeed decomposable into narrow states and contextual information, but that a regularizing or causal-subsumptive account of action explanation is not the only one required for full explanation. Instead *normalizing* explanation is called for as well.

Regularizing explanation, according to Pettit, renders an explanandum intelligible by pairing it with a cause, and then subsuming the two under a causal principle. The explanandum is thus seen as a consequence of how the world actually is.

In contrast, *normalizing* explanation is concerned not only with how the world is, but also has a teleological component. It is concerned with the norms that a system *aims at*, and the explanandum is rendered intelligible by its being seen as the event which was 'required if the world is to continue to satisfy that norm'. He says:

The regularizing thought is that if the principle generally obtains, and the antecedent is satisfied in this case, then the consequent is fulfilled too. The normalizing counterpart is: if the principle is generally to obtain, and the antecedent is satisfied in this case, then the consequent has to be fulfilled. Either thought might be expressed in the original deductive formulation.[Pettit 1986 p. 38]

He tells a story about science-fictional robots whom the beneficent Martians have sent down to Earth. We are promised that they are useful and will work smoothly. We find them poking around under trees and try to explain what is going on. Finally we realize: they have discovered that the grass is matted and have set about aerating the soil. When we see their behaviour as *gardening* behaviour we can explain it by subsuming it under the rôle of gardening. We may also want a causal subsumptive account of the aetiology of this kind of behaviour, we may even also want a property-analysis of the architecture of the Martian robot's CPU, but these are different explanatory purposes.

Pettit goes on to argue that from a normalizing perspective broad folk-psychological propositional attitudes are less gerrymandered than composites of narrow states and contextual information.

Normalizing explanation seems to be irreducibly systematic: that it is concerned with the properties of systems as wholes. It is only at the level of token behaviour intentionally specified that normalizing explanatory deployment of propositional attitudes attributes anything non perspective-dependent to the objects of explanation.

It seems to be indifferent to the realization of the processes it describes; or at least to the causal processes underlying the candidate explananda. This is what guarantees its independence from the details of a true cognitive science. It could turn out—indeed it probably does—that there is nothing physically in common between states in which belief attributions are made. There is even the possibility that different individuals in relevantly similar circumstances, exhibiting relevantly similar behaviour, and the subject of similar folk psychological propositional attitude attributions could be instantiating greatly dissimilar internal states at all but the grossest

functional level—the output which enables the behaviour to be normalized. Indeed in this regard, the notion of normalizing explanations (and the explicitly instrumental position of Dennett) is, if anything, consonant with the admittedly hazy work being done in non-computational AI. The idea here is that there is no level of explicit computational representation in the mind: indeed possibly no levels of abstract high-level programme-like description between the very lowest levels of organization and the very highest level of behaviour. If this view is right, then the proto-science view of folk psychology is out, but the last three versions are unharmed. Of course non-cognitivism does no harm to the voodoo thesis.

So on the normalizing explanation view, we do not have to attribute some independently identifiable fact about the candidate explanandum's computational or physical state. But this seems at odds with the attempt to keep the causal relevance of normalizing explanations. For example, Pettit says

...on the normalizing conception of action, the intentional profile remains a cause...unless the attitudes involved were causally responsible for the action they could not be invoked to explain it. [Pettit 1986 p.47]

This suggests that Pettit sees the propositional attitudes of folk psychology as something more than just an heuristic aid or an instrumental strategy, but rather attributes these states to people as psychologically real entities. Something like this I take to be Davidson's [Davidson 1980c] anomalous monism: the token propositional attitude or mental state is identical with the physical state (and therefore explains the action in a causal subsumptive way) but because each instantiation of a propositional attitude is so different, there are no psycho-physical laws. But it is hard to see how token causal explanation is *required* for normalizing explanation in any strong way. Of course we can take it that if we are explaining some piece of behaviour

which we are certain has happened, then *something* has caused it. But what functional story licenses the *identity* of the states in a normalizing explanation with the states which actually cause? The justification for the strong attribution of a token propositional attitude which is identical with a token physical state is not really clear.

In any case, in a footnote Pettit says that 'the pattern of cause and effect is not enunciated by the normalizing principle; the principle merely provides evidence of its existence'.¹² This sounds different; a lot hangs on 'enunciated' and I am not sure how to take it. And I do not see how normalizing explanation gives evidence of a causal law much stronger than raw behaviour. Of course there is *some* causal explanation, but what light is shed on this by normalizing explanation?

3 Explanation at Higher Levels and Broad Psychology

The most recent defence of broad psychology is Frank Jackson and Philip Pettit's 'Functionalism and Broad Content' [Jackson and Pettit 1988]. Although it is not explicitly stated, the paper appears to be motivated by the hope that some of our folk intuitions about propositional attitudes will turn out to be true of broad states, and so that if broad states can be shown to be explanatorily indispensable, then some of those folk intuitions will be vindicated.

¹²I should point out that Pettit regards the views of [Pettit 1986] as superceded-or clarified-by [Jackson and Pettit 1988]. I retain this discussion because the 1986 paper provides a very good example of an attempt to clarify the often articulated but seldom worked out notion that rational psychology represents a *norm* which is both approximated to, and is explanatory of, human behaviour.

Whereas in [Pettit 1986] it was argued that, for normalizing explanation, it was more *natural* to provide action explanation in terms of broad states, and that the narrow state plus environment explanation was something 'gerrymandered out of it' [Pettit 1987 p52], the strategy in Jackson and Pettit is to show that there is a kind of explanation of which broad psychological explanation is a member, which provides something *extra* over what other kinds of explanation provide, and that this something extra is often what is called for in explanation requests.

They base their view of explanation by broad states on their notion of program explanation (see ch. 4). Despite some of my concerns about the program/process *distinction* in that section, none of these undermine what I take to be the *purpose* of Jackson and Pettit's program/process distinction: to show that an explanation need not have a causally efficacious property as its explanans, and that an objection to the explanatory efficacy of broad psychological states founded on their causal *inefficacy* is founded on sand. The next section examines this claim.

3.1 Program Explanation and the Doppelgänger Challenge

This notion of program explanation is used to defend the explanatory value of broad psychology from the doppelgänger challenge. The doppelgänger challenge is their name for the difficulty for broad content raised above. To recap briefly; suppose that there is a molecule for molecule duplicate of me typing at a twin-Macintosh keyboard on some Twin Earth. Call this my doppelgänger. His beliefs and desires about his twin-Macintosh are different from my beliefs and desires about my Macintosh if construed broadly because the objects of the propositional attitudes are different in each case.

Yet, as we sit here and type our theses, exhibiting the same behaviour,¹³ if we want an explanation which is the same for both of us and thus captures the obvious generality, it had better be one in terms of *narrow* propositional attitudes, since that is all we have in common, and since they are what does the actual causal work in each case.

Jackson and Pettit accept that along this axis of change—sameness of narrow state, change in environment—that an explanation in terms of broad states will indeed be redundant (though not idle, since the sameness of broad state would still guarantee *some* narrow state which does the work), but posit a different axis of change along which a broad state explanation would be a program explanation with all the attendant benefits. This would be one where there is difference in narrow state, but sameness of broad state. Suppose that on one occasion I see a system error sign on the Macintosh screen and quickly reboot by selecting restart from the displayed menu. On another occasion I hear the characteristic beep of the machine's crashing and, without looking at the screen, reboot by turning the power off and then back on with the startup disc left in the machine. Both of these cases seem to be cases of knowing that the machine has crashed, desiring that it be restart, knowing that rebooting will have the desired consequence and therefore doing so. And any account of how I came to have these capacities may well involve these descriptions. Yet quite possibly there is no independently identifiable narrow state in common in me on these two occasions. My narrow behaviour is in both cases uncontroversially different, and so the causal paths in me that led to them may also be different. The

¹³For the sake of brevity in these cases I shall not discuss the objection from the broadside that it is only the *narrow* behaviour which is the same in these cases. This results in detour around the argument for the dispensability of broad behaviour.

environment, however, remains the same. An explanation in terms of my broad belief that the machine has crashed would cash in on the multiple realizability of my broad belief by narrow states and environments; it will explain because it 'programs' for some narrow state that does the causal work; and in addition it will draw our attention to the fact that it doesn't matter for computing purposes how my belief that the computer has crashed is caused, or how it is realized as a narrow state.

Jackson and Pettit do not even require that the environment remain the same for the same broad state to be instantiated. They write:

For a given broad content B there will be a number of ways of realizing that content by the appropriate combination of narrow state content N and environment E, say: $N_1 \& E_1$, $N_2 \& E_2$, One of these ways, say $N_a \& E_a$ will be the actual way B is realized. Now each of $N_i \& E_i$ will explain and predict different behaviour in the subject, but it may be that there is a common thread T running through these different pieces of behaviour. In this case ascribing B explains and predicts just as well as ascribing $N_a \& E_a$, and does something distinctive besides—it tells us that it did not matter as far as getting T goes that it was $N_a \& E_a$ that was actual instead of, say, $N_2 \& E_2$

From this we can see that in the broad state case the environment does not need to remain fixed to realize the same broad state. $N_1 \& E_1$ could instantiate a certain broad state B, and so could N_2 and E_2 . So if we replaced my second example above with one in which both my narrow state was different (I heard the beep and used the power switch—presumably caused by different narrow states) *and* my environment was different—someone had switched my computer for Philip Pettit's, this could *still* realize the same broad state. This point will be important in the light of the objection and the response to it in the next section.

3.2 An Objection and a Response

One objection to the explanatory indispensability of broad states roughly is this:¹⁴ in each of these cases, it is agreed, there is *some* narrow state that does the work. So take the following set of combinations of narrow state and environment:

$N_1 \text{ \& } E_1$

$N_2 \text{ \& } E_2$

$N_3 \text{ \& } E_3$

$N_4 \text{ \& } E_4$

Let each of these combinations of environment and narrow state make up the same broad state B on the Jackson and Pettit story. Now there *is* something in common between the narrow states in each of these combinations: each of them does, however differently, the causal work required to produce the same broad state when it is in the right environment. Perhaps there won't be any narrow *realizer* states in common, but who said that broad psychologists have a monopoly on rôle states? In each of these cases the narrow state involved could be the same narrow rôle state. In other words, there must be something they have narrowly in common to the extent that their causal commerce with the environment is similar. In which case the same explanation in terms of narrow state is available in every instance in which the combination of environment and allegedly different (perhaps lower-level in the sense of ch. 4) narrow states. In which case the

¹⁴Something like this objection was raised in a paper read by Michael Devitt to the Department of Philosophy in the RSSS at ANU, early in 1988.

doppelgänger objection goes through, since the broad story is explanatorily redundant.

I distinguish two versions of this argument, the *structural realist* version and the *rôle attributive* version.

3.2.1 The Structural Realist Version

This version of the argument supposes that there will, at some level of description, turn out to be real structural similarities between the putatively different N states. It hopes that the best description of the mind, from the bottom up, will have at one of the higher levels functional similarities which can be attributed in virtue of how the mind *really works*, not just in virtue a state attribution made on the basis of a perceived commonality of rôle within broad states. This is just the kind of possibility that Stich [Stich 1983] amasses evidence against, and which, if connectionism were true, would certainly be ruled out. On a state-realist reading of Fodor, the LOT hypothesis, if true, could perhaps underpin this version of the argument. What would be narrowly in common between the various N states would be sentences in the language of thought, however different their actual causal consequences or predecessors.

Two points need to be made about this suggestion. The first is that whether these supposed intrinsically motivated structural similarities will in fact turn out to exist, is an entirely empirical matter. The next chapter will go some way towards arguing that there is no good reason to suppose that they will.

The second point is that even if such similarities are forthcoming in human psychology, this may turn out to be species chauvinistic. If, say, connectionism is false for us but true for the Venusians, then on this kind of

account, explanations of Venusian behaviour which are open to the broad stater simply by positing the same broad state realized by some bizarre Venusian narrow state and its environment, are blocked for the pure narrow stater. The broad psychologist has *prima facie* claims to a more general psychological theory due to possible multiple realization across species, even if it is in fact realized in one way in humans. This last point is really by way of a promissory note; I will pay at least some of the balance outstanding on it in section 4.

3.2.2 The Rôle Attributive Version

This version of the argument simply attributes a narrow similarity to the various N states in virtue of the rôle they play in the broad states. There is an example made famous in [Pylyshyn 1980]. Mary knows that there is a fire, and runs away. The problem is that Mary can detect fire and run from it in a number of different ways. She can smell it, see it, feel it, hear the fire alarm and so on. Even if it turns out that the states that result from these different inputs do not turn out to have anything in common looked at from a narrow perspective, we should (in some way in principle!) be able to list all the different narrow states which, in an appropriate environment, would play the rôle of mediating between the fact of fire and the running away. Let us characterize these combinations as $F_1 \& N_1, F_2 \& N_2, \dots$, where the N states are token neurophysiological states which do the job, and the Fs are the appropriate fiery environments. Then the N states have the functional similarity that they all play the rôle of recognizing fire, and initiating fire avoiding behaviour. So, we rewrite our list of states with the narrow functional rôle state in the N position and we get:

$F_1 \& \text{NRS}$

$F_2 \& \text{NRS}$

F₃ & NRS

and so on. What do all these broad states have in common? The same narrow rôle state. So what provides the non-redundant explanation? The narrow rôle state. The problem is that it is all done with mirrors; the narrow rôle state explanations are reflections of the broad state explanations in that there is no way of independently identifying them. They are the high level state of having a state which is part of a broad state which explains the behaviour.

3.3 A Moral To be Drawn By the Broad Stater

All is not entirely well for the broad state account, though. Consider the form of the reply to the Rôle Attributive narrow account. The narrow rôle states were granted to have explanatory force in each case, but there was nothing intrinsic about the narrow rôle states which accounted for their explanatory force. The states were intrinsic by courtesy alone; they were disjunctive states, the disjuncts of which could apparently only be determined by reference to the broad states. It is much the same move as that made by the kind of reductionist who argues that supervenience can be reduced to reduction, just so long as we countenance reduction to potentially infinitely disjunctive states, the identity of which may not be intrinsically determinable. There is nothing to *stop* these moves, but they are notational variants on the broad state story in the one case, and the supervenience story in the other. And what is more, by introducing special kinds of narrow states or reduction without using different terminology, they are confusing notational variants because they obscure important scientific differences.

Is the broad state account as Jackson and Pettit formulate it immune from a similar response? Consider again what makes a broad state broad. It

is broad because it supervenes on how things are, both within the skin and in the environment. Thus the characteristic $E_1 \& N_1, E_2 \& N_2, \dots$ form. Now, consider some such list of conjunctions of narrow states and environments $E_i \& N_i$ which are realizations of some broad state B:

$$\begin{array}{l} b_1 = E_1 \& N_1 \\ b_2 = E_2 \& N_2 \quad \left. \vphantom{b_2} \right\} B \\ b_3 = E_3 \& N_3 \dots \end{array}$$

Now each of the states b_i is a broad state; it supervenes on how things are in the skin as well as in the environment—even if it is the limiting case of supervenience where the relation goes the other way as well, i.e., reduction. In any case, the states b_i are *constituted* by the combination of the narrow state and the environment. The state B, which is what does the real work, could be realized by any of the states, b_i . But what is the ‘common thread’ which makes each b_i a realization of B? It is not broadness as such that does the work; B is no *more* a broad state than any state b_i . Nor is there any simple account to be had about how it is that the environmental part of the story binds the disparate narrow states into a set of similar broad states, b_i , each of which is a realization of the higher level broad state B. Remember from above that it is not required that the environment remain *constant* through variations in the narrow states that are part of the realization of the broad states b_i . It is not as though each E_i were the same, and it was this which somehow guaranteed sameness of broad state across relevantly acceptable changes in N state (although even there the work is being done by this talk of relevant acceptability). What this boils down to is that, formulated in the way Jackson and Pettit do, the broad state story is open to a similar objection to the one that the narrow state story was open to. The similarity of the b_i states does not depend on any obvious fact about them; rather it depends on

decomposing them from the identity criterion for the B state. But it is not at the B state level that broadness enters the picture; it enters the picture at the b_i state level. So there is nothing distinctive about *broadness* which explains the commonalities. If we can say that each b_i realizes the same B state because of some fact about B states, why is that any more acceptable than saying that each N_i is the same narrow rôle state because of *its* involvement in realizing B?

One way out is to abandon the conjunctive talk about the environment and the narrow state, and instead replace it with some kind of (yet to be explicated) *relation* between the narrow state and the environment. Of course just such a relation could be buried in the conjuncts, but we had better be explicit if we want the broad states—or at least those that are explicated—to bear the potential explanatory weight. Instead of the previous list, let us have instead:

$$b_1 = E_1 R N_1,$$

$$b_2 = E_2 R N_2,$$

$$b_3 = E_3 R N_3 \dots$$

In this formulation, there is something in common between each of the broad states b_i , which is not in common between the N or the E states, except by reading back from the b state. What is in common is the relation R, which would not appear on considering each b state as a mere conjunction. Broadness—or at least the sense of broadness which is explanatorily non-redundant—would then not just consist in the narrow and environmental component, but would consist in a relational component. The *higher level* broad state B would supervene on the environment and the narrow state,

just in virtue of each of its realizations having the same relation hold between the narrow state and the environment.

If something smells fishy here, it is because metaphysical problems ought not be able to be solved through varied notation. There remains a substantial question as to where this relation R comes from: if the relation R holds between some N_i and some E_i just in case the N_i and the E_i together constitute a realization of B , then the broad story will still have problems caused by an analogue of the reply to the narrow rôle state objection. In other words there will be nothing that makes any b_i more obviously a natural realization of B than some N_i .

This would not present any difficulty if it turned out that in each case the environmental part of the conjunction was required to make up the mental state, but this is just what the narrow stater denies. The broad stater does not aim for a cheap victory by saying that of course the environmental component is required, that is what makes it a broad state. This would be to do no better than, and probably a little worse than, the cheap narrow response of saying that the environmental component is *not* required for the narrow state, and the narrow state is all you need. After all, the narrow stater might continue, this gives the bones of a schema which accounts for things like misrepresentation and errors of fact; in these cases the same narrow state is there, but the environmental factors which determine *truth* of beliefs and so forth are absent. The broad stater must make the more subtle claim that while the narrow state may well remain the same narrow state in the absence of the environmental factor, this is *because* of the rôle it *would* play in the broad state, and that there is no independent way of identifying it. And the challenge of fleshing out the R relation is to find some relation which can only be identified by analysing broad states which provides the real similarity between the b_i and is thus the only way to ground the N

states. At this point then, the most that can be said for broad psychology is that it *could* be indispensable, it remains to show that it *is* indispensable.

3.4 The Challenge

One way of expressing the underlying worry that was at work in the previous section is to ask what it is about various broad realizations of a higher level broad state that make them all realizations of the same state. Consider another example; Fred enters the office on various mornings and sees the Professor. The lighting conditions on each morning are somewhat different, so the professor has rather different shadows on her face, and is angled in different ways. On each occasion, though, we want to say that Fred is in the same broad state B of believing that the Professor is in the department, and the actions he takes—various kinds of greetings—are explained in terms of that broad state, despite their different realizations. Now the challenge that we inherit from the section above is to show that what is in common between these situations is something which is more naturally seen as in common between the different broad states—the combinations of the various narrow states and the environmental features which vary around the common theme of containing the Professor.

If the relational analysis in the section above is correct, then what we are looking for is some *relation* between the environment and the narrow state and the environment which captures these similarities. More importantly, if we want our broad states to *explain* action, then this relation had better *explain why* there is such a similarity.

The narrow challenge will then be that it is hard to see how the environmental factors, even in this non-Doppelgänger kind of case, account for what is similar about each case. In each case there is, both sides agree, a narrow causal account of why the particular action has taken place.

Attributing the broad state does not add anything to the causal account, and it does not seem to add anything to the account of what is the same in each case, unless there is a *prior* stipulation as to the identity conditions of the B state. The narrow states in each B state could be different, the environments in each case could be different, the actual causal relations between them could be different. If the relation R is a causal one, there still has to be an account of what needs to be causally similar between the causal relations in each case to count as the same, higher level causal relation R. The sting of the response is this; just *adding* environmental components and causal relations between the environment and the narrow component does not clarify what is in common. There had better be some story about why they are scientifically important causal relata and environmental states. The original problem for the narrow stater is saying what the narrow states have in common; the additions only give us *more* things between which to find the common element.

3.5 Synchronic and Diachronic Broad States

The quest to find what is in common between various broad states which we take to be tokens of the same folk propositional attitude leads to the question of just what kinds of environmental facts these states supervene on; in particular to the question of their temporal location. Do broad states supervene on *past* states of the environment and narrow states (and possibly the causal history of the narrow state as well), or on *current states* of the environment and narrow states? If the former, it will be possible to individuate states where there is no relevant synchronic environment along lines similar to those in causal theories of belief. If the latter, then various time-sensitive indexical beliefs will be better handled. I will call the former kind of state a *diachronic broad state*, and the latter kind of state a *synchronic broad state*.

3.5.1 Synchronic Broad States

A *synchronic* broad state is a conjunction of a narrow state and some environmental factor which is, at least roughly, concurrent with it. So Mary believing the fire is burning could be a synchronic broad state comprising Mary's narrow state, the fire outside, and perhaps some environmental factors at that time which constitute the relation between them. If we taxonomize the narrow components according to the highest level of intrinsically motivated similarity, and do the same with the synchronic environmental states, then we will have the kinds of b states discussed above. But these, in virtue of the way they are taxonomized, will be *realizers* of the broad states in a folk psychology. But, as we saw above, this kind of broad state tells us nothing about what is *in common* between various realizations of the same, say, belief. So we need a *higher level* kind of (synchronic) broad state which is realized by the lower level broad states, and the identity conditions for which involve this elusive common element.

I take it that in the usual literature about broad states it is some kind of synchronic¹⁵ broad state that is intended. In [Jackson and Pettit 1988] the example of "my belief about *this cup* " is used, and the broad analysis involves a state involving me, the cup, and perhaps some other relevant environmental linkage. To the extent that I could be linked together with the cup in the right way on various occasions, a state of at least the lower level kind is needed; if the argument from program explanation for the indispensability of broad states is to get off the ground, then the higher level kind of state would be called for.

¹⁵I do not mean by synchronic *strictly* contemporaneous, of course—there must be enough time lag for the causal processes to work.

3.5.2 Diachronic Broad States

A *diachronic broad state* is a broad state composed of some narrow state and some environmental factors including causal relations *preceding* the narrow state. Just the same kind of sub-taxonomy as we used in the synchronic case can be used to distinguish the lower level kind from the higher level kind.

My motivation for introducing these diachronic broad states is whatever the *reason* for each of N states in the synchronic broad states being partial realizations of the same state, it surely has *something* to do with its causal history and its past grounding in objects of belief (or other propositional attitudes) which has brought it about that the particular narrow state has particular behavioural consequences.

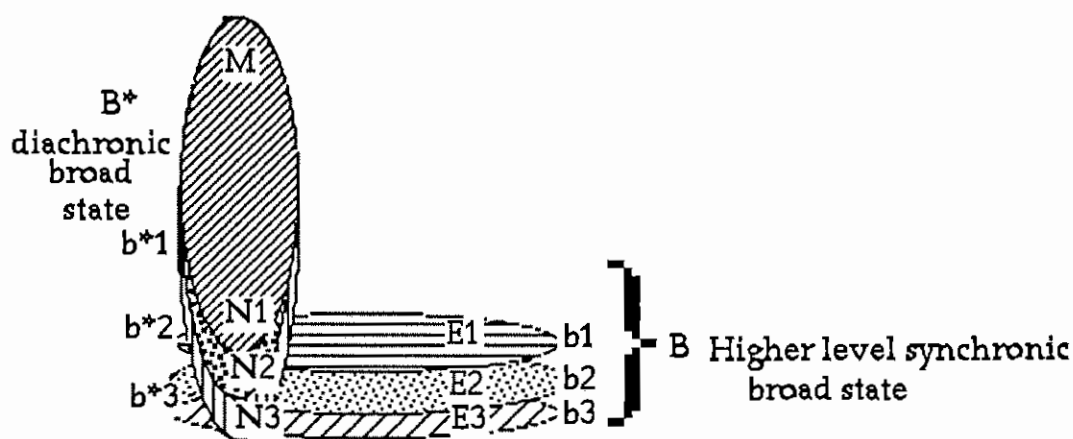
Consider again the various cases of Mary confronted with fire, and her narrow states being heterogeneous and the environment varied also. What rôle does the environment play in each case? Suppose that Mary is hallucinating—or is deceived by appearances—and still believes there is fire. If what is to count as criterial of having the same propositional attitude is the synchronic broad story, then a peculiar difference between the true belief case and the misrepresentation case emerges. In the true belief case, the propositional attitude is realized just because it is one of the approved combinations of N states and E states; in the misrepresentation case the narrow state is realized just because, if the appropriate environmental state were realized as well, then the synchronic broad state *would* be realized. As a truth functional relationship this should come as no surprise to anyone, but I do not see how it *explains* what is in common between the narrow states. There may not be anything about the actual environment in the misrepresentation case which does the explanatory work, and to invoke the counterfactual above is not to give an account of the *mechanism*. If we were

to give an account of the mechanism by which this counterfactual gets its force, say call the account (involving perhaps past environmental states and causal processes) M , then we could ascribe in the misrepresentation case a diachronic state $N_i \& M$. It would be in some very strong sense *in virtue* of N_i 's being part of $N_i \& M$ that it would be the narrow state that it is. By the argument first used against the Attributive Narrow State objection above, the diachronic broad state $B^* = N_i \& B_i$ would be ineliminable, and any supposed commonality between the N_i would be parasitic on the M , since what narrow state is required for realizing B^* cannot be identified independently of the diachronic broad state.

If we take this road, we can eliminate the difference between the misrepresentation case and the true belief case; because although in the true belief case the full token synchronic broad state is there, if the explanation in terms of diachronic broad states explains the propositional attitude attribution in the misrepresentation case, then it should do so in the true belief case.

Consider again the example: in each of the true belief cases of Mary believing that there is a fire, what *explains* why the occurrent narrow state is part of a belief that there is fire is its connexion with past fires and environmental states; the occurrent state is indifferent to the environment: if it were a brain recently transferred into a vat that would be something causally isolated from the narrow state. The synchronic environmental facts need not cause in the right way the narrow state to be what it is; in some story past states may well do so. Considerations of multiple realization and program explanation would rescue the synchronic broad state only if they uniquely delivered the something extra; the account of what is in common between narrowly disparate states and environmental states so as to make them the same propositional attitude. But the diachronic broad states would

do this as well; and in a way transparent to the actual causal processes involved. Consider the following diagram



The horizontal axis represents the move from narrow, skin-bound state, to the external environmental components. As can be seen, a higher level synchronic broad state B can be realized by various b states which are in turn realized by the combinations N_1 & E_1 , N_2 & E_2 , The vertical direction moves back in time via causal relations with originator external states; I have called the cluster of originator states and relations M . So the N states are *also* members of b^* states realized by N_1 & M , N_2 & M , ... which in turn realize the propositional attitude B^* . The multiple realizability of propositional attitudes under different realization in terms of N states is just as well accounted for by their being realizations of B^* as by being realizations of B ; and in addition an analysis of B^* brings us closer to the true causal story.

3.6 An Objection

One worry about this is whether there is going to be any principled way of discovering a univocal M . Remember that M was supposed to pick out the cluster of originator states and causal relations such that if some N_i is appropriately related to it, it would be part of an instantiation of B^* . If M is a

disparate group of states for which there is no compelling scientific reason to join together, then the diachronic story will suffer from the same problem as the synchronic story; motivating the collection of different realizations as counting as realizations of the same state in other than an *a prioristic* way. All this is to say that at least *some* of the program of the causal theory of belief must succeed, and some gains be made in the direction of a solution for the misrepresentation problem, in order for the diachronic broad state story to have these kinds of advantages over the synchronic story.

3.7 The Rôle of the Synchronic Environment

Where does all this leave the synchronic broad state story? There is at least one rôle which the synchronic environment has: it *sustains* the representational rôle which a narrow state plays. My claim is that although there may be an account of propositional attitudes in which diachronic broad states determine, at any given time, what propositional attitude is instantiated, the environment at that time allows the narrow state instantiated at that time to *continue* with the representational rôle that it has. The explanatory rôle that synchronic broad states might have, then, is of explaining why the same narrow states *continue* to be involved in the same propositional attitudes.

3.8 A Twin-Earth Example

A twin-earth example may help to clarify the respective rôles of synchronic and diachronic states (c.f. the appendix). Suppose Hilary Putnamow is an astute observer of things, has most of the beliefs people in this speech community have about water, but (for simplicity's sake) she does not know much chemistry, having been more interested in mathematics when she was at school. All of a sudden she is transported unawares by a Zygorth transporter ray to Twin-Earth, where the stuff that looks like water is XYZ.

Fortunately things are pretty much like they were at home, and XYZ is potable thanks to certain safeguards built into the rematerialization program of the Zygorth transporter ray.

Hilary's beliefs about water include things like "water is the stuff that the experts think it is" and all the usual deferential apparatus, and she starts her new life.

On first encountering XYZ she slurps it thirstily, believing (perhaps falsely) that it is water. Presumably nothing about the environment determines that this is her belief. Her narrow psychological state is much the same as it would have been in similar circumstances back on Earth. It is the diachronic broad story which makes the narrow state she is in part of the propositional attitude of believing that this is water. Some other narrow state could have done the trick; different neural pathways could have been used than are; different percepts even could be involved in the belief; if qualia supervene on some more well understood level of description then water could have had quite a different qualitative aspect. But it doesn't; and it is a historical story which guarantees this. While Hilary was in transit, there was a fact of the matter about what narrow states could be involved in the propositional attitude of believing there to be water.

So when Hilary first arrives, her beliefs are about water—though they are false. This becomes less plausible, though, with time. After she has been living happily on Twin-Earth for some time, interacting successfully with its inhabitants and so forth, the pressure to describe her beliefs as about XYZ—called 'water' in the local lingo—seems pretty strong. Prediction of Hilary's behaviour by the locals, and both her's and their rendering intelligible this behaviour is best served by ascribing to her true beliefs about XYZ rather than false ones about water.

If we wish to preserve this intuition, what price do we have to pay? Well, whatever the details of the claim we want to make about meaning change in a case like this (see appendix)—for instance that Hilary's indexical beliefs in the new context shift her reference (the local experts have changed, for example, though she doesn't know this)—one factor remains constant. Whatever the reasons for such changes, they have nothing to do with changes in her narrow states, because there haven't been any.

A temptation that this might give rise to, is to suppose that since the (narrow) behaviour of both Hilary and the Twin Earthers is much the same as that of Hilary and the Earthers had she stayed there, then all of the apparently powerful intuitions that something has *changed* can be dispensed with. The fact that we need to shift the reference of "water" for her in order for the usual folk explanations to go through on a broad story about belief shows that what we should really do is realize that what we are doing is explaining commonalities in narrow behaviour, which are explained by commonalities in narrow state. But that is exactly what we don't want to do because of the problem of finding reasons for grouping potentially disparate narrow states together as single, identifiable propositional attitudes.

This example shows us that, especially if we want Hilary's beliefs to be about XYZ (called 'water') at her first gulp, some rôle must be found for the synchronic environment. Indeed, on the combined descriptive/causal story of kind terms suggested in the appendix, provided there are enough indexical beliefs in her belief cluster, she will pretty soon come to have true beliefs about XYZ when she slurps the stuff. If she believes that water is whatever the local experts believe it is, and she believes that *this* (XYZ) is water, she is surely right. It is because of the *synchronic* connexions she has to the local experts and local stuff that this is so. We cannot dispense with a diachronic story: only it can account for mental states being representational

in some way in the absence of an environment (Robinson Crusoe's thoughts about home?), or in misrepresentation cases (c.f. [Fodor forth.]); but if the sustaining rôle of the synchronic state breaks down, whether by transportation to Twin-Earth or more mundane environmental change, then the synchronic environment had better be involved.

4 External Functionalism

So far the argument for broad states has been that it seems that there may be heterogeneous narrow states whose only commonality is that they appear to be linked to an environment in relevantly similar ways; so a proper specification of the same state relation will involve mention of the environment. But it has only been a hope; as was argued above, broadness as such provides no guarantee that we will be able to specify what the commonality is. It is just a guess that the solution, whatever it is, will involve broad states of some kind. In this section I flesh this hope out just a little more. It seems that the higher level broad states are *functional states* of some kind, but the worry is that they might be *gerrymandered* functional states—or at least functional states which we can do without for scientific purposes. *We want some way to show that they are (in the terms of ch.4) strong functional states.* It is all very well to say that functional states which involve relations to the environment *could* be explanatorily important (which is I take it the claim made about 'broad functionalism' in [Jackson and Pettit 1988]); it remains to show that propositional attitudes are. The account of explanation in terms of high level states given in ch. 4 allows gerrymandered high level states to *select* the explanandum if they are chosen so as to satisfy, *inter alia* the causal relevance constraint; but for a decent explanation we want a strongly functional, non-gerrymandered state to do the selection. The proposal is that that narrow states N are possible part realizers of some broad state B because they all play *the same functional rôle in the environment.*

Hence the section heading 'External Functionalism'. Functionalism, as it is usually conceived in the philosophy of psychology, talks of the functional rôle the states play *within the organism*. The very notion of functionality is one which involves an environment. The functional theory of pain, for example, solves the apparent difficulty that different states which all count as pain states might be 'physically' (for which now read 'at a lower level of organization') different, by postulating that all these states play the same rôle *in the organism*. So there is a suppressed relational content to any functional claim. Some state plays some rôle in some system.

There is also nothing new about the idea of state attributions of bounded modules depending on the rôle they play in an environment. When one module of a computer program is replaced by another which "does the same job" we are concerned that, in the context of its environment, it has the same effects in important ways, regardless of how this is internally realized. When some chip in a computer is physically replaced with a different model with the IO modules adjusted to fit, and *it* does the same job, we tend to make the same state ascriptions of it as we made of the original model. And these states are relational; they are broad states. It is not as though the functional description of the chip is just an abstract higher level description of the narrow states of the chip. The state attributions we make about it, and in which its own different internal narrow states count as the same functional state depend on the relations that hold between it and the environment. Plug the same chip in a different environment—read 'machine'—and its intrinsic states will form part of different, or perhaps no, contentful states. How we taxonomize these states depends on what is the simplest, most predictively powerful and elegant theory of the system as a whole—however we balance those desiderata.

So much is uncontroversial. In the psychological case, though, bland talk of 'functional rôle' has always *assumed* in a rather unexamining way that the right environment to consider is the environment within the skin. The arguments above suggest that psychological state attributions based on the functional rôle they have in the broader environment are going to meet the criteria for causal relevance at a high level. In this case, no matter how varied the narrow states which make up the broad states are, no matter how close Stich and the Churchlands are to truth about how the brain works, then we will have criteria for including diverse narrow states as narrow components of broad states which may resemble the propositional attitudes of folk psychology—if there indeed are such functional rôles in the environment.

4.1 Some Candidates for the Rôle

The simplest approach to finding such a functional level of description¹⁶ is to say we have it already in folk psychology. We have a powerful and predictively successful tool which seems to predict and explain human behaviour. There must be *something* systematic about how our folk psychological sentences are linked to the actual causal mechanisms of behaviour, otherwise it would not be so predictively successful. In fact, whenever we have good reason to believe(!) that someone desires X, and believes that *ceteris paribus* doing Y will yield X and nothing else undesirable, and all things are equal, it more often than not turns out that they do Y.

So, even if there were nothing about brain states which looked like a belief or desire, and which was systematically taxonomizable by the

¹⁶I think that something like this view is now held by Frank Jackson and Philip Pettit.

cognitive scientist, there must nevertheless be (perhaps highly distributed) facts about the causal mechanisms in the brain and in the world in virtue of which these regularities hold. So let our talk of beliefs and desires be talk of states which supervene on these facts, whatever they are. There may be no regularity visible to neuroscience or cognitive psychology, but in *each case* there will be states which do the work. The elusive commonality between them is just given by the observed regularity and success of the folk psychological enterprise.

4.1.1 Problems With the Simple Solution

The problem with this simple solution is that it is not defended as well as one might hope against eliminativism. The danger is that *any* reasonably successful heuristic can be turned into realism about the theoretical objects of the heuristic by going functional. As Steve Stich complains, even phlogiston could be rescued if we allow that it is realized by oxygen in all of the interactions which it is supposedly engaged in. There is a response to this; that (i) there are parts of the phlogiston theory that are just unambiguously false and (ii) that these are part of the *conception* of the phlogiston theory. But so too there are no doubt many parts of the folk theory which are false as well, and it is not at all clear which of these are so central as to be included in the conception of the folk theory.

In any case, if a systematic science of human behaviour which involves intrinsically motivated states (as connectionism does) ever emerges, and which accounts for the success as well as the occasional failure of folk psychological terms, then even if by simply functionalizing folk psychology we can guarantee the truth of its state ascriptions, that is no great victory. We will not have shown that it has any distinctive theoretical merit; no generalities will be elegantly and scientifically captured by it that the

cognitive science might not capture. I think we can do even better than this by capturing what is in common between folk psychological states in a way which gives them a potentially broader domain of applicability than any science which is concerned with the details of mental operation at levels where they may be specific to our species.

4.1.2 An Evolutionary Constraint

One proposal might be that what is in common between these broad states is that they play some kind of rôle in our evolution as thinkers and language users.

The evolutionary case is interesting. Suppose that evolution does indeed select for rationality, or at least enough rationality that beings so selected for are capable of distinguishing rational from non-rational behaviour. Suppose that some mutation takes place amongst gorillas, and a gorilla emerges that is super rational, intelligent, and can learn English and interact with ordinary English speakers in a way that would pass the Turing test. Now on a view which requires that the entities whose behaviour can be explained by their propositional attitudes have an ætiological history which accounts for their rationality, the first few such gorillas could not have their behaviour explained in terms of folk psychological content attribution. However, if after a few generations this kind of gorilla is selected for and begins to predominate, then the ætiological constraint is satisfied, and it becomes possible to engage in normalizing explanation. Yet nothing has changed about the gorillas or the (human) environment they are in!

One intuition which seems to deserve preservation is that if folk psychological content attributions have an explanatory rôle in the latter case then they do so in the former; why should the history of the *species* have anything to do with the success of psychological explanations of

individuals? On the other hand, if we dispense with the ætiological requirement, then we seem to be left with nothing but a brute fact about dispositions to behave which is certainly not in itself going to explain why the dispositions themselves are there. We might be able to (and perhaps would) take an intentional *stance* towards the gorillas, but how would that *explain* anything? We may have to look further for our common element. No matter how far we look, however, there is one kind of explanation which the functional rôle of folk propositional attitudes will not do successfully unless there is an evolutionary (or other ætiological story) to be told about them. The thought is that perhaps the functional rôle of folk states explains *why we are in them*. In the case of the mutated ape that seems to think and talk, there is nothing about the functional rôle of its states which explains their presence, whereas in the case where organisms have been selected for operation which is roughly describable by folk psychology, perhaps the function of the states specified ætiologically explains the presence of the states. However we do not require that much explanatory power. What we require is not an explanation of the etiology of the states in terms of their function; rather in folk psychology we are looking to explain behaviour in terms of the states, and need some story about what is functionally in common between different tokens of a state. The strong ætiological requirement may be stronger than is required for this.

4.1.3 A Dispositional Account

The analysis of functions by John Bigelow and Robert Pargetter [Bigelow and Pargetter 1987], which I mentioned in Ch. 4 as a candidate for an account of what makes something at a strongly functional level of description rather than at a merely gerrymandered one, suggests another approach. That analysis was, roughly, that some property of an organism has a certain function just if it has dispositional powers to enhance selection. So rather

than (as an ætiological account of functions would require) insisting that the functional property have been selected for, it is enough if it is selectively advantageous, i.e. that in a normal environment it will be selected for. So in the example above, the apes who have mutated so as to behave in a way which might make one take an intentional stance towards them *will* have this property selected for, if the environment remains normal, and this accounts for what is important about that abstract property, and thus justifies our using it as a candidate explanans. This, coupled with the fact that that properties are ultimately realized physically by causally efficacious mechanisms, means that their behaviour can be explained by their beliefs and desires.

This concentration on dispositional powers leads to another thought about these functional rôles that is only trivially related to selection. Perhaps folk psychology describes how organisms or things must be connected to their environments in order for them to be information gatherers and users in the way that people are. It describes how, not individuals, but rather systems of individuals work at a high level of abstraction. The connection with selection is simple. If folk psychological generalizations (*however* they are implemented) *ceased* to be true of some system of individuals, it would collapse. Passing on information about food sources, cooperating to create things based on beliefs about natural science and so forth may all be possible only of a system which is connected up in this way. In other words, in some reasonable robust sense it may turn out that folk psychology tells us how a system *functions*; they are functional properties in virtue of the realization of which the system (our system in particular!) *works*. If this purely conjectural but suggestive idea is right, it would also give grounds for thinking they were important properties even in light of the One True Cognitive Science, should it appear. For folk psychology would have to be

true of *any* information using and transmitting species of partly autonomous organisms that was successful; regardless of whether connectionism or the language of thought (or, as Fodor now thinks possible¹⁷, both) were true of it.

4.2 Why External Functionalism is Not Behaviourism

One possible difficulty for External Functionalism is that it is simply a kind of behaviourism, and if so it suffers all the difficulties which have led to the demise of behaviourism in all its guises. The claim might proceed as follows. The only thing that an External Functionalist should take as causally relevant to the systematic environmental story that she will want to tell, is the narrow behaviour of individuals; it is these narrow behaviours which are grouped together in bunches to form the propositional attitudes of external functionalism, regardless of what it was that caused them. This is no different from Rylean behaviourism; after all, even he did not take token behaviours as constitutive of mental states, but rather grouped them together in behavioural types, in much the same way that the External Functionalist will want to do.

This spectre can, however, be easily laid to rest. The argument that it is only narrow behaviour which is relevant to the systemic analysis is arbitrary: one could equally say that all that is relevant is a state-description of the area just around the individual with all its causal properties; or else one could draw the line just inside the skin, as it were, in the nervous system. The point is that the narrow *internal* physical facts which are partial realizers of folk psychological states are required to ground the causal relevance

¹⁷See [Fodor and Pylyshyn 1988].

causally relevant to the outcomes, and are likely to function as important causal nodes in a systematic picture of what is going on. They alone will not do, as is shown above, but they are needed for their counterfactual supporting rôle which gives an account of content in the immediate absence of an environment, and content in the absence of the usual behaviour (attributions of which are certainly useful enough). So External Functionalism is committed to internal states in a way which classical Behaviourism is not.

Another important distinction between behaviourism and External Functionalism is that External Functionalism does not pretend to do *all* the work; indeed it is compatible in various ways with various other positions in the philosophy of psychology. Its most obvious compatibility is with the substantial scientific hunches of some of the most vociferous enemies of folk psychology. If the Churchlands are right about the best scientific theory of the brain, or if some kind of connectionism (see [Smolensky 1988] or ch. 7 of this work for a survey of connectionism) is true,¹⁸ then the external functional story would be necessary for its particular explanatory purposes, since nothing from those theories of the brain will look like having the kind of structure which will map on to the folk taxonomy—if indeed it is that taxonomy which will best describe the states which do the external functional explanatory work.

Most tellingly, however, behaviourism provides no account of what it is that these behaviours have in common. External Functionalism, however, at least promises some kind of account of what the similarities are in terms of their function.

¹⁸Or at least some radically non-Computationalist or non-Representational theory.

4.2.1 The Case if Fodor is Right

But what if some Fodorian scheme is right, and there is a theory of the mind in which propositional attitudes that roughly resemble those of folk psychology *do* supervene in a regular way on structural features of the brain, so that they can be identified narrowly, even if the environment provides their content? In this case, content might not be necessary for our behavioural explanations to go through, as the independently identified narrow states provide explanations of behaviour in each case. Since the (narrow) theory of the mind would sufficiently resemble the folk theory, there would be enough narrowly in common between states which we want to count as the same for our general explanatory purposes, to use the narrow states in each case. This is to ask whether the whole problem which led us to seek Broad states in the first case—the search for unifying features for propositional attitudes over an axis of difference orthogonal to the Doppelgänger one—might just not occur.

I do not think this is very likely. But even if it is the case, the broad external functional states may still have an explanatory rôle to play. For suppose that our states do taxonomize narrowly in a way which maps on to the propositional attitudes of folk psychology. Suppose that there is a language of thought which is productive, systematic and so on, and that a reading of our neural states could, through certain transformation laws, tell us what the sentences in that language are. This would provide us with narrow states we could use to explain actions of members of *our* species. But what is to say this would have to be the case for *any* species which nevertheless was able to engage in the kind of informational exchange which leads to the growth of control over the environment? Unless the arguments for the Language of Thought or computational theories in general are disreputably *a prioristic* (see ch 7) the possibility that

connectionism—or even some non-representational low-level theory—might be true for a species has to be allowed.

Suppose then that, having amused us with the gardening robots of [Pettit 1986], the Martian builders themselves arrive and talk to us about the philosophical test they were performing to see how we would interpret their behaviour. We get on splendidly with the Martians, and they give papers at the AAP conference which are greeted with acclaim. Then some of their neuroscientists and cognitive scientists talk to ours; horror! It turns out the Martians *don't* have a LOT. In fact they are a connectionist network that has evolved in a messy kind of way to interact in way which is recognizable by us, and seems *prima facie* to be explainable by us as exchanging beliefs, expressing desires and so on. Well then, the external functionalist story could still be the best, neatest most elegant and general account of Martian interaction, and thus provide a justification for the propositional attitude attributions we were making before anyway to explain their behaviour. And if this kind of analysis is needed for the Martians, there will be no loss of economy in applying it to us, even though in our case there is greater narrow regularity about what realizes the states within the skin. Having then *separated* accounts of the narrow states, however much they look like the folk states as the proto-science view hopes, we are left with the scientific freedom to allow the external functionalist and internal functionalist stories to go their own way. If there are differences, major or minor, between the taxonomies which these accounts give of propositional attitudes, then they will be settled on their own terms; not by trying to make the taxonomies of one kind of theory match those of another.

So the External Functionalist story provides a greater degree of generality than the narrow stories. It need not just be a theory of how human brains work, but rather is an account of what systematic interactions will

have to take place, at a fairly high level of abstraction, for a species to interact in a way which allows members of it to predict what other members will do and exchange information about the environment in which they find themselves.

There is no need to claim that External Functionalism is what the folk have always conceived themselves to be talking about. I am not at all sure what the folk have always been talking about, and I suspect there is no fact of the matter. Some folk intuitions are narrow, some broad. Some of the folk think that it doesn't matter how a mind is instantiated; others will never believe that a robot that passes the Turing test has mental states. One philosophically prominent member of the folk believes that the gooey grey stuff is logically necessary for the mental. An over obsession with purely causal theories of the reference of natural kind terms might lead one to suppose that, whatever the folk are talking about, it is grounded in the One True Theory of the Mind. If there is no OTTM, this makes life difficult. But then folk thinking about propositional attitudes is a mess; discovering that there could be a range of theories of mental states helps explain this mess. And if every such theory has some counterintuitive consequences, this is only to be expected, since the intuitions come from a set of folk conceptions in which the different theories (which have different explanatory purposes) are intertwined and confused.

4.2.2 Some objections by Steve Stich to Compatibilism

I want now to consider some objections to this kind of position by Steve Stich [Stich 1983] and [Ramsey, Stich and Garon 1988 forth.]. Stich's complaint against what he calls the instrumentalist view of folk psychology (but I think it is meant to have wider scope than just that position) is that it is a disreputable trick. It is as if a practitioner of alchemy said "Well, we were

never claiming that our theoretical entities represent intrinsic and primary properties, they are just instrumental *abstracta* useful aids in prediction and, like lines of force, perfectly respectable in this rôle."

The first response is that the plausibility of alchemy's disreputability is that the terms of chemistry are obviously microstructural; there is just no plausibility to the view that alchemists thought that they were positing properties which were not microstructural properties of the substances that they were dealing with. Also, our current use of chemical terms seems to be microstructural; and to the extent that we might want a high-level functional story about chemistry, alchemy is generally not going to be a useful one, even if in fact some orthodox chemistry *is* just such a high level story (solubility, for example). In the case of psychology, however, this is much less clear. It is not at all obvious that the folk are making microstructural attributions to people when they psychologize. It is plausible in a way that it is not in the case of chemistry, that when the causal-subsumptive strand is removed from folk psychology and replaced with cognitive science, there is something left. If it turns out that Reagan and some Twin-Reagan or Reagan android both act so as to inspire the belief in someone that Reagan believes the Sandanistas are all commies, there seems to be something left to explain even if it turns out that their internal structure at a functional and a physical level are quite different. Phlogiston, though, construed as a functional property at the highest level (the dispositions to behave of groups of chemicals) seems to be gerrymandered in a way that folk psychological content attributions are not. Further, there is an important difference between the psychological case and the chemical case. It is *much* more plausible in the actual world that systems which satisfy the constraints of a folk psychology are multiply realizable, than terms like phlogiston should be multiply realizable. phlogiston construed

functionally is just not very interesting; nothing but oxygen could (physically) possibly fill the phlogiston rôle,¹⁹ we have a good story about what oxygen does, so why bother with phlogiston? The generality argument above, though, suggests that there is at least the possibility of an interesting generalization over a wider domain to be had in the case of folk psychology that is not to be had in the phlogiston case. This is no knock-down argument, but the foe of folk psychology had better come up with an account of why there should be no difference.

Another objection comes from the proto-science camp; Stich thinks this objection is actually a concession to the friends of folk psychology. It goes like this: folk psychology is like cooking. It may well be that when you beat mayonnaise into mustard it will—mostly—reconstitute. There is some generalization here. But these generalizations are vague. When you replace them with scientific chemistry you get rid of the vagueness and totally supersede in principle the folk theory, even if in practice cook books are still useful.

The first thought is that mayonnaise is not multiply instantiable, so there is no need for a criterial theory of how it can be realized independently of its instantiation. This would be a crucial difference between folk psychology and cooking. Mayonnaise would be reducible to a chemical theory, where as FP supervenes on its base. Even if we do take mayonnaise to be a functional term though—and I gather that there are fake mayonnaises which would be saved on this account—and each instantiation

¹⁹This is meant to be independent of the [Jackson and Pettit 1988] point that oxygen could not fill the phlogiston rôle (and hence nothing does) since it is constitutive of the phlogiston theory that phlogiston is *emitted* during combustion.

was identifiable by passing the Taste Test, there would be no reason to suppose that there is a correlation between passing the Taste Test and obeying the Mustard Law. Many of the laws of cooking depend for their efficacy on the chemical properties of the usual instantiators of the culinary kind. It is no *culinary* fact about mayonnaise that guarantees its interaction with mustard. So there is no generalization to be had; there is no reason to suppose that some other substance *tasting* like mayonnaise would behave similarly. In the case of folk psychology, however, the performance requirements, social or evolutionary, which constrain a system which is to be explained, guarantee that there will be a maintenance of generalization over multiple instantiation. In any case, the generalization argument above comes to the rescue; the generalizations may be vague, but they have more scope. Moving from the One True Cognitive Science to folk psychology may indeed have some cost in vagueness, but equally it may pay dividends in generality.

Two other objections: Stich's intuitions [Stich 1983 p. 244]²⁰ balk at the idea that some entity which exhibited rational behaviour in virtue of an exhaustive list of rules governing behaviour could actually be rational and have psychological states. Suppose that the Block head²¹ is finally produced; an android that has an enormous look-up tree structure. What goes on in its

²⁰In [Stich 1983] the objection which follows is formulated as an objection to *instrumentalism*. I am indebted to a conversation with Stephen Stich for this objection formulated as an objection to *compatibilism*.

²¹Ned Block's original example is of two chess playing computers, one of which encodes the rules of chess and some strategies, the other of which just has a sufficiently vast number of pre-recorded games stored that it can just look through the list at each move, and and make a move that keeps it in one of its stored games.

CPU bears little resemblance to the kind of information processing that humans do. Instead it has in effect an *extremely* large but nevertheless finite list of taped 'lives'. It evaluates each input, looks along its list of 'lives' until it finds one which is recorded as receiving that input, and then provides the next output from that life. If the result is sufficiently complex it might be behaviourally equivalent to a normal adult human being over a sufficiently wide range of possible inputs to pass as human; in particular it might pass the Turing test. Nevertheless, Stich thinks it is obvious that such a being wouldn't have psychological states. Its internal structure is just too different from ours for it to be remotely plausible that we share psychological states with it. In particular, it isn't *intelligent*. It doesn't do any important thinking: that was done by the programmer (or the machine) which produced the taped lives. All this machine does is look things up. So you shouldn't want your offspring to marry anything like that, you shouldn't feel sorry for it when it seems to scream in agony and so on. The example is supposed to threaten an account of folk psychology which is indifferent to how the states are realized. If you do think that the Block head does not have the relevant kinds of states, then you think that how these connexions to the environment are realized internally matters. It is then a tempting (but not compelling) slide to suppose that it is having the same narrowly taxonomized structures that makes for identity of psychological state. In which case there is no extra generality for a folk story to depend on, because there would in fact be no way (or at least not many ways) the states could be realized, other than by the human cognitive apparatus or something which duplicates it at a computational level.

The easy response would be to dig my heels in and say that I believe the Turing test to be criterial for having psychological states. Perhaps the *computational* complexity which real people have is compensated for by the

tremendous richness of the data structure in the Block head. After all, it is impossible to underestimate the extraordinary complexity of a machine that really was a working Block head.

More consideredly, it is important to distinguish two quite distinct issues that are being conflated here. What the folk think about psychology is very complex, and there is not even any guarantee that it is consistent. Perhaps the folk think that it is constitutive of psychological states that they are, at least sometimes, accessible to introspective experience; that there can be something which it is like to have certain beliefs or desires. Some of Stich's use of the human version of the Block head example seems to trade on the possibility that subjective experience—cognitive qualia, if that is not a contradiction—depends on the actual architecture of the human brain. But the kind of argument for externally functional folk states I am running here does not depend on denying this (although I see no compelling argument *for* it either). Although beliefs and desires are sometimes available to introspection, sometimes they are not—and we are still prepared to attribute them in those cases. The predictive and explanatory rôles of beliefs and desires remain, even if their possessors were introspectively dead. To return to the alien creators of the robots: if it turned out (though it would be perhaps in principle impossible to find out) that they were introspectively dead, this would not diminish the usefulness of decision theory in predicting what they did; and it would not reduce the usefulness of a story about the systematic connexions of their narrow states to the environment as an account of how their species functioned as a whole. If it were true that awareness depends on being like us narrowly, although on some views their introspective deadness (due to being Block heads, or in some other way not being functionally equivalent to us in the right computational way) might be *morally* relevant, it might not be cognitively relevant. They believe things,

but since they aren't aware of this it might be OK to slaughter them. I doubt it—but even if this is so, it is no problem for an external functionalist account of beliefs and desires.

The last Stich objection [Stich 1983] I want to deal with is that folk psychology will be less general than a scientific psychology because it will not work for the very young, the brain damaged and so on. This does not strike me as a fault but rather as a virtue, since it captures how we do relate differently to different kinds of systems. There is surely something importantly different between the brain damaged and the psychologically normal. A functional account of the rôles that folk psychological propositional attitudes play in the environment might give us an account of what that difference is in terms of whether an individual satisfies a functional description, the satisfaction of which by most individuals gives an account of how the system works.

4.3 The Return of the Axes

This chapter began with an account of the four axes of variation along which views about psychological states differ, and all of which are enmeshed in the distinction between cognitive science and folk psychology. It ends, by way of summary, with an account of the position of External Functionalism along these axes.

4.3.1 High Level/Low Level

External Functional states are fairly high level states. Any states which neuroscientists discover, or which are posited by internal functionalism will be the part instantiators of External Functional states. What would make External Functional states *interesting* high level states, and not just gerrymandered or artificial ones, would be their strong functional rôle; so it

will not be the intrinsic nature of the states which will be the sole ground for their attribution.

4.3.2 Broad/ Narrow Axis

External functional states will be broad, but their broadness is not that of traditional broad psychological states where the narrow state and the object of the narrow state somehow together form a realization of a big B Broad state. Rather it is the rôle played by the state in the environment that is important. Since the narrow states that play this rôle cannot be identified independently of the environment in which they play it, the Devitt-Narrow strategy outlined in § 3.2 will pay no dividends. Also the broadness will not be only of the synchronic kind. Externally Functional states in which the rôle is played by the state in a past environment, will be needed in cases of misrepresentation²² and where the environment changes.

4.3.3 The Revisable/Unrevisable Axis

Externally Functional states are revisable. They are not, theoretically, an attempt to preserve folk intuitions or classical semantics, though they are *motivated* by the thought that the success of certain elements of the folk theory might have a grounding. But we cannot have our conceptual cake and eat it; if it is the rôles which narrow states play in environments which confer explanatory respectability on external functional states, then the exact taxonomy of them awaits a serious study of the rôles narrow states actually do play in those environments. Perhaps this is a little too extreme; perhaps

²²It should be stressed again that this is not a claim to a solution of the misrepresentation problem; rather it is a suggestion that whatever such a solution is, it will be found in some specification of the rôle that states and environments regularly have in virtue of how they were set up.

Proust amongst many others *has* given us a serious study. In the scientific spirit, though, we await falsification!

4.3.4 The Manifest/Unmanifest Axis

External Functionalism is neutral on this one. What is it which guarantees *awareness* along with everything else? One of the transformations in the philosophy of psychology over the past twenty years is that we now talk about mental states in a way which is neutral about their being subjective experience of them. Perhaps this is because it is too hard. Who knows what is required for subjective awareness or qualitative perception. Perhaps everything which shares external functional states has it, perhaps internal functional states the same as ours are required to preserve this quality, perhaps functional organization at the neural level is required—perhaps even the Grey Stuff (though I doubt it). Presumably there is a fact of the matter; but it is hard to see how one would ever find out.

4.3.5 The Explanatory Motivation

This kind of hypothesis about what constitutes folk psychology depends on two views about explanation; that it need not cite the particular real causes (perhaps ultimately microstructural if the views in Ch. 5 are right), but that it must be causally relevant. Abandoning the causal relevance requirement would allow even purely instrumental views about folk psychology to be explanatory; after all, presumably purely instrumental considerations would, for example, redistribute probabilities in the way called for by van Fraassen. Requiring that the actual microstructural causes (or causally efficacious properties) be non-redundantly cited, however, leads to an insistence on narrow states as the only possible realizers of folk states, and if there are no such narrow taxonomies, eliminativism will make some headway.

I think this is a most attractive consequence of the account of explanation offered in the first four chapters; by insisting that explanation is causally relevant we retain the feeling that a proper explanation involves citing facts in virtue of which the explanandum has come about, but also by allowing counterfactual information about what would cause things in relevantly similar situations to count as explanation, we are able to capture important abstract generalizations. And folk psychology, one of our most successful and long-lived complex explanatory theories, is nothing if not both important and abstract.

Chapter Seven

Explanation and the Language of Thought

Table of Contents

1	Introduction	259
2	The Language of Thought.....	260
2.1	What is the LOT Hypothesis?	261
2.2	Arguments for the LOT.....	263
2.2.1	Explaining Cognitive Capacities	263
2.2.2	The Methodological Argument	268
3	How Not To Get A LOT For Free.....	271
3.1	Syntax and Semantics.....	272
4	Explanation I: Synchronic and Diachronic Explanations	276
4.1	A Slightly Stronger Claim.....	280
5	Explanation II: Implementation and Levels of Explanation.....	282
5.1	Why High Level Causation Brings on the Language of Thought	285
5.2	The Fallacy of the Implementation Fallacy	287
6	Conclusion.....	290

1 Introduction

In the previous chapter I have argued that the account of explanation I have given makes room for the possibility of an external functionalist theory of propositional attitudes. This chapter has a more negative argument; that these same views of explanation make Fodor's account of propositional attitudes less compelling. It is not often that theories of explanation constrain what we think are good explanations, or what we think is in need of explanation. More usually the direction of fit is the other way, as in the cases of explanatory asymmetries. In the philosophy of psychology, though, intuitions about what counts as a good explanation are so hotly disputed that we would do well to see what light can be brought to bear on these areas from our preferred theory of explanation. Views in the philosophy of explanation can have considerable bearing on substantive explanations in the philosophy of psychology.

The purpose of this chapter is to examine Jerry Fodor's famous Language of Thought (hereafter sometimes LOT) hypothesis in the light of some of my views about explanation. Roughly, I will argue that if you have the right views about psychological explanation, then you don't *need* the Language of Thought to explain any of the available data. This is not to say that the Language of Thought hypothesis is wrong—I take that to be an empirical matter for sorting out by those psychologists who are not of the armchair persuasion—but, rather, it is to say that there is no *prima facie* case for it to be made out by philosophers or psychologists of a philosophical bent.

The plan is as follows: in section one I will outline what the substantial Language of Thought hypothesis is, and will run through the currently canonical list of Fodorian arguments for it. I will take passing swipes at

some of these, so as to leave the substantial arguments for the rest of the paper.

In section two I run through a weak version of the Language of Thought hypothesis, and explain why it might be the plausibility of this which has led to such acceptance as the substantial Language of Thought hypothesis has had.

Sections three and four deal with questions in the theory of explanation, and how they bear on the strong version of the hypothesis. Section three contains an argument that the explanation of the behaviour of complex systems may not require an elegant synchronic explanation in terms of its structure, if there is reason to believe that there is a diachronic explanation of why it has such a behaviour that does not necessarily support the elegant structure story. Section four is the most crucial: here I argue that if supervening state ascriptions (such as mental states) are not required to causally interact with one another, then while they may in some way *explain* behaviour they do not cause it. If they do not cause it, then there is no need to take them to be *intrinsic* states. Thus the argument for the strong LOT is blocked, though not the weak one. I conclude that it is Fodor's insistence that high level structural states must not only explain but also cause behaviour, which generates the strong Language of Thought from the weak one.

2 The Language of Thought

The Language of Thought hypothesis as it was first introduced made two claims: that we needed to postulate an internal representational system (probably innate) which was rich enough to support complex linguistic and cognitive skills, and that this system of representation had a particular structure much like that of a language. The gist of the former claim goes for

the most part unargued these days; just about all of us are representationalists of some sort. It's the latter claim that remains a point of in-house debate amongst representationalists. In this section I detail this latter claim and summarize the arguments for it.

2.1 What is the LOT Hypothesis?

The LOT is made up of three sub-claims. Firstly, the claim that mental representations:¹

... have a *combinatorial syntax and semantics*, in which (a) there is a distinction between structurally atomic and structurally molecular representations; (b) structurally molecular representations have syntactic constituents that are themselves either structurally molecular or atomic; and (c) the semantic content of a (molecular) representation is a function of the semantic contents of its parts together with its constituent structure.
[Fodor and Pylyshyn 1988 p.12]

¹ Fodor sometimes speaks as if he takes the LOT to be a doctrine about *mental states* rather than mental representations. He says:

LOT claims that *mental states*—and not just their propositional objects—*typically have constituent structure*. So far as I can see, this is the *only* real difference between LOT and the sorts of Intentional Realism that even Aunty admits to be respectable. So a defence of LOT has to be an argument that believing and desiring are typically structured states. [Fodor 1987 p.136]

The intuition that mental states have constituent structure (for example the view that the belief that Becker is playing at Wimbledon and will win, somehow has the belief that he will win as a component, and the belief that he is playing at Wimbledon as a component) could be preserved without buying in to the strong LOT hypothesis, if you have an account of mental states in which they are not *narrow* states of the brain. On the other hand, if with Stalnaker [Stalnaker 1984] you do not think that even *content* has structure, you may not want to preserve the intuition at all.

When Fodor says that mental representations have 'constituent structure' he is talking about (a) to (c). Because mental states are constituted in part by structured representations, cognitive processes may be defined in terms of those representations. A cognitive process is the transformation of 'any mental representation that satisfies a given structural description...into a mental representation that satisfies another structural description' [Fodor and Pylyshyn 1988 p.13]. An obvious example of this structure sensitivity of a mental process is that of inference. It is a process of inference, for example, that will transform a representation of the form 'P & Q' into a representation of the form 'P'.

The LOT also makes a substantive commitment to the physical instantiation of structured representations. Mental representations:

...are assumed to correspond to real physical structures in the brain and the *combinatorial structure* of a representation is supposed to have a counterpart in structural relations among physical properties of the brain. For example, the relation 'part of', which holds between a relatively simple symbol and a more complex one, is assumed to correspond to some physical relation among brain states. [Fodor and Pylyshyn 1988 p.13]

The requirement that the properties of mental representations proposed by the LOT are instantiated in the brain makes the LOT a considerably strong thesis. In order for a cognitive system to qualify as instantiating the LOT it must possess more than mere input-output properties. In fact the LOT is an even stronger thesis since it is also committed to the claim that:

... the physical properties onto which the structure of the symbols is mapped *are the very properties that cause the system to behave as it does*. In other words the physical counterparts of the symbols, and their structural properties, *cause* the system's behaviour.[Fodor and Pylyshyn 1988 p.16]

As we shall soon see, this final claim regarding the causally efficacious structure of mental representations is crucial for the current work.

2.2 Arguments for the LOT

We should accept the LOT if there are good arguments in its support. The arguments currently on offer, found in Fodor's 'Fodor's Guide to Mental Representation' and *Psychosemantics* [Fodor 1985 & 1987] and Fodor and Pylyshyn's 'Connectionism and Cognitive Architecture: a critical analysis' [Fodor and Pylyshyn 1988], come in two basic kinds: arguments from the explanation of cognitive capacities and a methodological argument. In the remainder of this section I review these arguments with the aim of assessing their support for the LOT in the next section.

2.2.1 Explaining Cognitive Capacities

There are four arguments from the explanation of cognitive capacities.² As Fodor himself admits, all these arguments are really very much the same [Fodor and Pylyshyn 1988 p.48]. So, a description of two of them will suffice in order to give a flavour of the style of argument. Cognitive capacities exhibit two properties—productivity and systematicity.³ Cognitive

²They are the argument from *productivity*, the argument from *systematicity*, the argument from *compositionality* and the argument from *influential coherence*.

³The argument from productivity gets played down by Fodor these days in favour of the argument from systematicity. The reason he gives is that because of our mortality only a finite proportion of our putative potentially infinite cognitive capacity in fact gets used. In order for a cognitive system to be truly productive we must *idealize* from the finite *performances* to infinite *capacities*. By refusing to idealize, one may claim that we are constantly thinking and believing new things while denying productivity; if we only lived long enough then we might well run out of novel things to think and believe. Fodor now favours the argument from systematicity because he claims we do not have to idealize:

you can make these points about the systematicity of
language without idealizing to astronomical

capacities are productive because we are constantly thinking new and novel thoughts and believing and desiring new and novel things. Cognitive capacities are systematic because our ability to think some thought or believe some proposition is intrinsically connected to the ability to think or believe certain *other* thoughts and propositions. It is in virtue of this property that you don't come across cognitive systems with the ability to think that Jill loves Mary *without* the ability to think that Mary loves Jill.

The strategy Fodor uses to explain these capacities derives from the work of Chomsky [Chomsky 1968]. Chomsky thought that *linguistic* capacities are also productive and systematic. To account for this he claimed that the structures underlying linguistic competence are *generative*. That is, one's (tacit) knowledge or *cognizing* of a language consists in the mastering of a combinatorial syntax and semantics, out of which the entities over which linguistic capacities range—sentences and utterances—are

computational capacities. *Productivity* is involved with our ability to understand sentences that are a billion trillion zillion words long. But *systematicity* involves facts that are much nearer to home: such facts as the one...that no native speaker comes to understand the form of words 'John loves Mary' except as he *also* comes to understand the form of words 'Mary loves John'. [Fodor 1987 p.150]

Contra Fodor, my claim is that systematicity is generated only when *all* cases of relevant word forms have the requisite properties. You might think that whenever we understand 'aRb' we also understand 'bRa' and yet deny systematicity by failing to idealize to all the other possible cases. So some idealization is also required in the case of systematicity. The facts associated with systematicity *appear* closer to home. That's because all the speakers we've come across can think both 'John loves Mary' and 'Mary loves John'. But at best this is systematicity on a *local* scale since we are dealing only with finite performances. However, in order for capacities to be systematic we want systematicity on a *global* scale, and to get that we are required to idealize. If idealization detracts from the argument from productivity then it's also going to detract from the argument from systematicity.

constructed. Fodor's argument for the constituency of the representations over which cognitive processes range immediately follows. Since we explain the productivity and systematicity of linguistic capacities by postulating the constituency of sentences and assuming the psycholinguistic premise that we use language to express our thoughts, then we make the same inference in the cognitive case as we do in the linguistic case: viz. that the productivity and systematicity of cognitive capacities is explained by the *constituency* of mental representations. Mental representations have constituent structure because there is a combinatorial syntax and semantics for cognition. In short, productivity and systematicity are explained by there being a Language of Thought.

This argument style rests heavily on the assumption that the Chomskian enterprise will be vindicated. By citing Chomsky in the premise of his arguments Fodor uses it as *evidence for* accepting the LOT. But why should we let Fodor use Chomsky to lend credibility to the LOT story? There are two reasons why we shouldn't. Firstly, it is surely still an *open* question as to whether or not grammars are psychologically real entities in the way Chomsky maintains. We shouldn't let the plausibility of one contentious empirical hypothesis depend upon the truth of another contentious empirical hypothesis.

The second reason why one shouldn't take the linguistic case as evidence for the cognitive case, is that both the linguistic and cognitive cases would seem to be two sides of the same coin. In both cases we are trying to explain a particular capacity of a subject by postulating some intrinsic psychological fact about that subject. The fact that we do seem to use language to express our thoughts and the analogous way we treat thoughts and sentences—both are representational, are semantically evaluable, etc.—would suggest that these hypotheses are closely related—probably closely

enough related so that they both either stand or fall together. *Of course*, by taking one as a datum and using it in an argument for the other the latter follows, and *vice-versa*. But that's because they are essentially the same style of answer to similar problems.

We can assume that in some sense the LOT can explain productivity and systematicity. But Fodor's claim in the argument from the explanation of cognitive capacities is stronger than this implies. He claims that *only* a LOT can explain these properties of cognitive capacities, since you have to have *structured* representations in order to get these two properties. To see why Fodor and Pylyshyn think this let's take a look at an alternative to the LOT which postulates *unstructured* representations to see how it tries to account for systematicity and productivity.

The alternative view is that of Connectionism or Parallel Distributed Processing (PDP).⁴ Connectionism is described as the 'new wave' of cognitive science. It proposes models of cognitive architecture which are highly parallel instead of serial, and are 'brain-styled' to the extent that they build models based in part upon the properties of neurons and neuronal organizations. With Connectionism one doesn't get structured representations which have a combinatorial syntax and semantics. Instead, one gets a network of atomic nodes with each connexion having its own excitatory and/or inhibitory thresholds, according to which the spread of activation within the network occurs. Connectionists want to interpret the nodes featuring in a network semantically. They might interpret nodes to be

⁴A useful introduction to the cluster of views that goes under the name of Connectionism is Smolensky's 'On the Proper Treatment of Connectionism' [Smolensky 1988]

representations such as 'A & B', 'A', 'B' etc.. Although the nodes are *labeled* in this way as being structured, this labeling is in fact irrelevant to the properties of the nodes; they're unitary. All they have are causal powers defined relationally with respect to other nodes via the inter-nodal connexions. They have no intrinsic structure relevant to their semantic interpretation. In order for 'A' to be represented in addition to 'A & B' the Connectionist cognitive architect must separately build 'A' into the architecture, unlike a LOT architecture where once one has 'A & B' represented one automatically has 'A' represented.

From this description Fodor and Pylyshyn draw some implications for productivity and systematicity. While the Connectionist can model a finite performance mental history, that very model is not going to generate an infinite capacity. In such a model, the architect also has the option of constructing a model in which you get, say, the thought that Mary loves Jill *without* the thought that the Jill loves Mary. Of course the Connectionist architect can build her network so as to be consistent with a finite and systematic mental life; you can build Connectionist and LOT architectures which are input-output equivalent. However, Fodor claims, it is just as likely that there are mental lives which do not satisfy systematicity, say, at this input-output level. If Connectionist models are accurate, then we should expect there to be *gaps* in cognitive competence because since the systems don't have representations with syntactic structure; the systematicity of the system doesn't *follow from* the architecture. Connectionist architecture treats mental representations as a *list* instead of a generated set. Where the list happens to differ, then cognitive gaps may appear. Cognitive gaps, however, *don't* seem to appear. For these reasons Fodor and Pylyshyn believe that only a LOT architecture can truly explain

the properties of our cognitive capacities. Sections three and four address these arguments directly.

2.2.2 The Methodological Argument

The second argument which Fodor cites in support of the LOT is the methodological argument. This argument provides a methodological basis for the inference to the types of structures required by a LOT architecture from the capacities of the system evident in the argument from cognitive capacities. The argument goes like this. Fodor comes up with what he takes to be a plausible (not surprisingly given his interests) principle of nondemonstrative inference:

Principle P: Suppose there is a kind of event c_1 of which the normal effect is a kind of event e_1 ; and a kind of event c_2 of which the normal effect is a kind of event e_2 ; and a kind of event c_3 of which the normal effect is a complex event $e_1 \& e_2$. Viz.:

$c_1 \rightarrow e_1$

$c_2 \rightarrow e_2$

$c_3 \rightarrow e_1 \& e_2$

Then, *ceteris paribus*, it is reasonable to infer that c_3 is a complex event whose constituents include c_1 and c_2 . [Fodor 1987 p.141]

For example, if e_1 is the raising of my hand and e_2 is the hopping on my right foot, then we infer that the cause of e_1 and e_2 is the conjunction of c_1 and c_2 and not some *other* cause c_4 . Fodor's claim is that unless we accept the LOT we are going to flout this principle. If mental representations are not structured (as in the case of Connectionism) then whenever we think the thought that 'A & B' that thought has a different etiology from the thought that 'A'.

Just when principle P ought be invoked is crucial. One is required to ascertain that the event being explained is in fact complex. If the event in

question is not complex then the principle should not be invoked. In the case of my raising my arm and hopping on my right foot, it seems unquestionable that this seemingly joint action is a conjunction of two other physical events. So, the adherence to principle P would be recommended. However, in the case of the outputs of our cognitive system, although it seems that our thoughts and beliefs have constituent structure, we had better be careful in adopting principle P, since automatically concluding that they have constituent structure might be to beg the questions at issue in favour of the LOT.

This can be seen in Fodor's own example of synergism [Fodor 1987 p.143]. Synergisms are behaviours which although appearing to be complex, are in fact behavioural wholes; the elements are in effect fused to one another. One way in which synergisms develop is through learning. Perhaps an organism's raising its arm and hopping on one foot is a synergism because it was learned as part of a rudimentary system of communication, the behavioural elements of the language having a different ætiology from that of the individual pieces of behaviour "fused" to form the linguistic behaviour. Invoking principle P in this case would lead us astray since we need some independent account of whether or not behaviour is to count as a synergism.

So too in the case of cognitive capacities. We need some story as to which behaviours *are* synergisms and which are *not*. Only then can we apply principle P in support of the LOT. What about the case of beliefs? Does the ætiology of the belief that P & Q have as a component that of the belief that P? Suppose we want to know whether an agent's uttering 'P & Q' is just a composite of the separate ætiologies of an agent's uttering 'P' and 'Q' separately. According to Fodor, principle P would suggest that the

proximal causes are the same viz. $\{P, Q\} \rightarrow 'P'$ and $'Q'$ and $\{P, Q\} \rightarrow 'P \& Q'$. There is at least one important sense, though, in which this may not be true.

Contrary to the Fodorian principle that systematic behaviour should just 'follow from' the architecture, I do not think that all consequences of an agent's belief set are automatically believed by the agent. Consider the case of closure under adjunction. Someone may believe that P, Q, R and S , but if asked in a quiz whether a sufficiently long conjunction is true, she may have to *form* a belief token that $P \& Q \& R \& S$. And she does this by considering the evidence in the same way as she would for any other belief, even if the *evidence* on which she bases her judgement is her own several epistemic states. This is a special case of the realization in AI that allowing beliefs to be closed under deduction in general, will lead to the inability to distinguish between the deductive consequences of a given belief set which have actually been generated, and are likely to be useful in future proofs, and those which have not been explicitly generated.⁵ This suggests the following alternative model of the proximate ætiology of our agent's uttering $'P \& Q'$, viz. $\{P, Q\} \rightarrow 'P'$ and $'Q'$ whereas $\{P \& Q\} \rightarrow 'P \& Q'$. In this case the proximate ætiology varies across the utterances, despite Fodor's principle.

Fodor might reply that the belief that $P \& Q$ has as *its* proximate cause P and Q in which case the model looks like this: $\{P, Q\} \rightarrow \{P \& Q\} \rightarrow 'P \& Q'$. In this way one's citing of the proximate causes will conform to principle P.

⁵Recent automated theorem provers which attempt to model natural deduction, such as those by John Pollock [Pollock forth.] and Jeff Pelletier [Pelletier 1982] represent as separate entities those formulae which, at the semantic level, seem to be *composed of* constituent formulae. For a discussion of the problem of all deductive consequences being generated in the context of the Frame problem in AI see [Dennett 1984].

There would, however, seem to be no necessity to go back that extra causal step in explaining the utterance of 'P & Q', since the reason why that utterance is made is because the agent believes that P & Q. The only reason to cite the extra step would be to ensure that principle is adhered to, and hence get a LOT. But opting for the extra causal step needs to be argued for independently, not from the assumption that we want to secure the LOT.

Whatever one thinks of this argument, though, it makes Fodor's Principle P less convincing as an *argument* for the LOT; since if the LOT is true and there is constituent structure then the methodological argument is applicable, but if it is not true, and the argument for distinct ætiologies of apparently constituent behaviours goes through, then the methodological argument is inapplicable. In sum, if perhaps a little too strongly, the methodological argument is good in the case of psychology if and only if the LOT is true—and there is no independent way to establish the validity of the methodological argument.

3 How Not To Get A LOT For Free

The LOT hypothesis as described in the previous section is essentially the claim that there is a combinatorial syntax and semantics for mental representations with the ensuing constituent structure being mapped onto the physical properties of the brain. This raises the following question: in virtue of what does such a mapping exist?

One way of getting such a mapping is to construe the LOT hypothesis as postulating an algorithm for generating our productive and systematic capacities. Such an algorithm might be the neatest and simplest way of describing those capacities. Of course, constituent representational structure might feature in that algorithm. If you think that no matter how the brain actually operates it is *that* algorithm which is realised, no matter how

irregularly it maps on to the actual structure of the brain, then you can have, trivially, a Language of Thought. On such a view it is an *input-output* specification which is constitutive of some algorithm's being realised.

There are, however, any number of algorithms which could account for our behaviour. There are as many algorithms as you like for performing the functions of a pocket calculator, let alone a human mind. If you cull these by saying that any algorithm that *does the same thing*—i.e. is an algorithm for a human brain or a pocket calculator—is the same algorithm, then you have returned to the mere top level of input and output.

Fodor wants more than this; he thinks that it is *internal functional rôle* which will identify internal states (see ch. 6) [Fodor 1987 Ch 2.]. This at least *sounds like* he does not want it to be a mere mapping of the top level. So we need some extra, independent, motivation for supposing that some algorithm or architectural description which is *compatible* with the description at the level of input and output is the real one. If a taxonomy of the system motivated in some other way reveals similar structures which could be said to *realize* the algorithm, then perhaps that would do.

Fodor makes of lot of the fact that *token* states are *syntactic* states. In the next section I consider whether a syntactic analysis could provide such a motivation.

3.1 Syntax and Semantics

Fodor takes constituent structure to be *syntactic* structure. But on this construal of structure a LOT can be had if not for free then very cheaply. We can see this by examining the relation between syntax and semantics. I have two related claims to make; first, that if you have a semantic interpretation and something to map it on to you can generate a trivial syntax and second,

that you *can't* have a syntax properly so described without a prior semantics of which it *is* the syntax.

The LOT requires that mental representations have syntactic structures realized in the brain. The problem here is what is going to *count* as syntactic structure. Syntax and semantics are intimately related. The practice of logicians to behave as though the syntax comes first and then an interpretation is applied puts the cart before the horse. A syntax is a simple, if not the simplest description of a supposedly meaning bearing system, given its intended meaning. A syntactic constituent of such a system is that which makes some uniform semantic contribution to that system. What this means is that a syntactic item is taken to be a syntactic item *because it stands in a signifying relation to some semantic interpretation*.

Now suppose that the One True Cognitive Science is completed, and we have state descriptions of the brain which we can pair off with attributions of content given the standard semantics. The important question, then, is what *kind* of similarity between these descriptions is required to get an account of what the syntactic tokens which represent the same content are. One possible syntax—perhaps the crudest and hence useless, but a syntax nevertheless—would be a disjunctive one. Simply disjoin all the state descriptions which are true whenever a given content attribution is made, and count the disjunction as a syntactic token. This disjunctive state would then be the syntactic token of the mental state. In such a case you can get a syntax just in virtue of applying a semantics to something which you stipulate is representational, just in the same way as you can, if you must, map *Principia Mathematica* onto the Canberra Telephone Directory. What is more, you can come up with a syntax in which various addresses and numbers represent thrilling theorems of meta-arithmetic, and, with a massively gerrymandered account of similarity

relations amongst syntactic tokens, get constituent structure off the ground. So it seems that the first claim—that whenever you have a semantic interpretation and something to map it on to you can get a trivial semantics—looks fairly plausible.

In fact I am neutral about whether such a syntax is a trivial syntax or no syntax at all; what is worth insisting on is that an account has to be given of what makes something a 'real' syntax or a non-trivial one. If trivial syntax is what you appeal to, then syntax will not do the job of getting the strong LOT hypothesis from the weak one. It will not provide the independent motivation that I mention in the last section of the chapter: the kind of motivation which will make the syntax a *structural* realizer in the sense of Ch 4. Some kind of independent taxonomy will be required on to which it could turn out as a matter of empirical fact that the ascribed syntax maps.

In their more *a prioristic* moods (especially toward the end of [Pylyshyn and Fodor 1988]) Pylyshyn and Fodor seem to think that they can provide an independent and intrinsic structural account of the mind by simply taking the trivial syntactic story and forgetting about the semantics from whence it came. The assumption seems to be that if you have mapped the semantics on to the brain, and you are left with a taxonomy which gives you syntactic tokens, that there is no problem in then determining whether the syntactic tokens have constituent structure.

Having got these tokens, however, how do we go about deciding whether the tokens whose *content* is constituent are a constituent part, *qua syntactic token*, of other syntactic tokens? What is crucial here is that there is one way which is too easy. If the taxonomy of syntactic tokens comes from their being the token states that are realized when certain content

attributions are made, then they can be described as having constitutive structure in virtue of their relationships to the ascribed content.

This allows you to stubbornly insist that it is syntactic tokens that you are talking about, while there is nothing *intrinsic* about the brain which determines which token is a constituent part of another token in any given situation. What has happened at this point is that the move decomposes the syntactic tokens from a story about the semantic content of the tokens, and then posits relations amongst the tokens which come only from the interpretation provided by the semantic content. The temptation then is to think that you have structure even if you jettison the semantic story which led to the taxonomy of those syntactic tokens. But in effect we don't have any syntactic tokens in the absence of the semantic content.⁶ For a substantial syntactic account to be given, two factors are required: a semantics, to ensure that it really is a *syntax* that is being given rather than any other kind of description, and an independent motivation for the taxonomy of syntactic tokens, so as to avoid the merely trivial kind of syntax described above.

The upshot of this is that we can get mental representations with some form of syntactic structure, which in some way gets realized in the brain,

⁶This obviously has repercussions for the Syntactic Theory of the Mind proposed by Stich in [Stich 1983], since on the syntactic account of cognitive science, syntax might be *derived* from a semantics in just this way. Interestingly, the issue of how strong a claim the LOT hypothesis is making, and whether there is a LOT, applies to Stich's *syntactic* programme just as much as Fodor's; the only difference between them would be whether or not the syntactic states over which the LOT quantifies, if there were one, are *representational*.

but which does not satisfy the demands of Fodor's strong version of the LOT.

If this is the version of the LOT that one finds convincing then it's easy to see how the arguments from cognitive capacities and methodology support the LOT. The cognitive inquirer chooses a syntax, or more accurately imposes a syntax, upon a cognitive system in order to most neatly account for the capacities of the system such as productivity and systematicity. The methodological argument's principle P provides a general strategy for imposing neatness onto our explanations in much the same way as the weak LOT does. Again, though, we have not generated the strong LOT hypothesis.

4 Explanation I: Synchronic and Diachronic Explanations

I do not think that the traditional arguments which have been outlined bear on the strong version of the Language of Thought hypothesis, partly because it is not at all clear what *kind* of explanation the LOT is supposed to be. In this section I attempt to determine what type of explanation the LOT is providing. Fodor thinks that we need the LOT because in order to explain capacities such as productivity and systematicity, a mechanism or a particular state of an organism *must* be postulated in order to guarantee the presence of those capacities. If there is a LOT then we get these capacities automatically. If there is not, claims Fodor, then it is *unlikely* that these capacities would be evident. While other architectures may *allow* systematicity and productivity, none *guarantee* it. My claim is that the presence of these capacities can be adequately explained without the postulation of some specific mechanism or state of the organism which neatly and elegantly captures features of the organism which are visible at

the behavioural level. Such an explanation is to be had from, roughly, the pressure of evolutionary forces. An evolutionary style of explanation raises the probability that a cognitive system generates systematicity and productivity without making the commitment to a specific mechanism such as the LOT. I then go on to claim that such evolutionary explanations can place constraints on what remains to be explained by other kinds of explanation.

I start by distinguishing *diachronic* explanations from *instantiation* theory or *synchronic* explanations.⁷ The diachronic explanations, including evolutionary explanations, are concerned to give a causal account of how a system came to be in the state it is now in, and includes as a special case. The most usual explananda are states of a system or states of affairs, and the usual explanantia are earlier states of the system or states of affairs together with transition laws which describe the generation of the later state from the earlier.

The second kind of explanation which might be asked for—which I contend the LOT hypothesis is providing—is of the *synchronic* or *instantiation theory* kind. In this kind of explanation we are concerned to give an account of what it (actually) is for a system to have a certain property in terms of the structural states of that system. An explanation of the ductility, colour and conductivity of Gold by appeal to its atomic structure and the interaction of its outer electron shells with other gold atoms would, for

⁷This is possibly a similar distinction to that between transition state and instantiation theory explanations in Chapter 1 of [Cummins 1984], except that I mean something a little more general than Cummins' notion of transition state explanation, since also included amongst diachronic explanations are long scale ætiological explanations such as evolutionary ones.

example, be an explanation of the properties of gold at the surface level by appeal to the structural properties of the system which *instantiates* gold. It is no part of such an explanation to claim that anything which has the phenomenal properties of gold *must* have the underlying properties that gold does have, but rather that in fact the phenomenal properties of gold are explained by the structure it actually has.

Are the facts which the friends of a LOT are trying to explain explicable by a diachronic explanation? A diachronic explanation requires only that some previous state of the system together with some transition laws entail the explanandum state. This will be trivially possible if physicalism is true; some complete neurological description coupled with a list of inputs and the right neurological laws will provide an account of why some new state is the way it is. Can we get an explanation of productivity and systematicity of thought out of all this?

We can certainly get something which looks like an explanation of productive and systematic *behaviour* in each instance—which is why the question of exactly *what* is being explained is so crucial. But the friends of the LOT want more; they want an account of why, in general, the behaviour is (almost) *always* systematic or productive. In short, they want an explanation of the systematicity and productivity of the system as a whole. This is why we should see what the friends of the LOT are engaged in as a kind of synchronic explanation. They want an account of what property of the system it is which realizes these capacities.

Must we provide an synchronic explanation of these capacities as the friends of the LOT seem to imply? I think not.

Firstly, it is far from the case that the bare story about neurophysiological states and laws exhausts what diachronic theories can

say about the mind. You can jump up a level, and ask of the system as a whole how *it* came to have the behavioural properties—or the functional properties at the highest level—that it does. And the best kind of candidate for *that*, it seems to me, is some kind of evolutionary account.

I do not have such an account on offer here; I do, however, think that there must be some such account of how the mind has been tailored to be what it is now. Nor do we *need* such an account in detail; that one is to be had is common ground between us and the LOTers (if the LOT hypothesis is true then a diachronic story will tell us *why* there is a LOT in explaining why cognitive capacities are the way they are). Our claim is rather that this sort of story removes the surprise with which both Fodor and Pylyshyn think we should greet the news that minds are, more or less, productive and systematic. So I think that the objection that diachronic stories, (or neurophysiological or connectionist ones supplemented with an ætiological account) do not explain because it is a mere *accident* that the system is productive and systematic, is misjudged. It is no accident: it was selected to be that way, and this selection plays an important part in diachronic explanation of minds if the time dimension is long enough.

So the Language of Thought hypothesis rests on an explanatory imperative to provide a simple, elegant synchronic explanation of the structural properties of the brain in virtue of which it displays its productivity and systematicity of output. If there is an empirically adequate low-level account of the operations of the brain is to be had, and if any surprise at high level regularities which it displays can be removed by some evolutionary account, then there is no requirement to produce the neat synchronic account.

The imperative to produce a neat synchronic architecture of the mind in which, in Fodor's and Pylyshyn's words, systematicity simply 'follows from' the architecture, rests on a confusion between the rôles of transition-state and instantiation explanations, caused by a neglect of other routes to eliminating the surprise which the Language of Thought hypothesis despatches all too quickly. Synchronic instantiation explanations are not required to be neat. With complex systems it is often the case that the details of their operation, even at a functional level,, are messy. When we explain the properties of a chemical by its physical structure, we do not look for analogues in this structure of the phenomenal properties that we seek to explain. It is enough that these structures account for, more or less regularly, the explanandum properties. In the thought case it may indeed be surprising that the complex physical system displays these properties, but having given this instantiation account *why are we obliged to try to remove the surprise at the instantiation theory level?* This seems to be the hidden requirement that lurks in some versions of the LOT argument. Instantiation theories explain simply by describing the actual mechanism; in some sense it is a fairly weak explanation, *but that is all they do*. Surprise at what they do is often best removed by a diachronic account. In the case of the mind and the productivity of thought we have a candidate in the form of evolutionary pressure; the need to remove this surprise by appeal to some structural feature of the mind mistakes the purpose of instantiation theories of the mind.

4.1 A Slightly Stronger Claim

Our slightly stronger claim is that an ætiological or evolutionary explanation can not only remove the *requirement* that a synchronic explanation of a certain kind be provided; it can also actually constrain the sort of account we should give.

I propose two constraints on an evolutionary explanation of how our cognitive capacities got to be the way they are:

- 1) That whatever the explanation, it must account for the continuity or discontinuity between the highly systematic and productive capacities of human minds and whatever capacities are exhibited by infraverbal mentation.

and

- 2) that the minds which have evolved must have done so in incremental stages; however the mind works it got that way by additions and changes of an *ad hoc* nature; much the way a tree which is pruned to look like a giraffe gets that way. Each change does not proceed according to a plan; it is only the overall direction of change which is determined.⁸

Something like these constraints has been used by, for example, Dennett [Dennett 1984] to argue against (or at least motivate the arguments of others against) computationalism as a doctrine of the mind. Regardless of whether they bear on computationalism (whatever *that* really is) I think they do put constraints of some kind on a synchronic theory of the mind. Consider the following passage from Fodor and Pylyshyn:

⁸If it turns out that some kind of punctuated equilibrium account in which evolutionary change often or mostly does *not* proceed by incremental changes is right (see [Eldredge and Gould 1972], [Gould 1980]), then this point will go by the wayside. This would still not, however, be evidence *for* the LOT architecture being selected.

It's possible to imagine a Connectionist being prepared to admit that while systematicity doesn't *follow from*, —and hence is not explained by— Connectionist architecture, it is nevertheless *compatible* with that architecture...The only mechanism that is known to be able to produce pervasive systematicity is Classical architecture. [Fodor and Pylyshyn 1988 p.49]

The thing to notice about this claim of theirs is that they claim it is an *advantage* that classical theories and their attendant LOT simply guarantee systematicity; thus removing all possible surprise at systematic behaviour. The above constraints on an evolutionary account, though, suggest that it might even be a *disadvantage* that systematicity and productivity are guaranteed. A theory of the functioning of minds which allowed more or less systematicity to appear according to how they evolve, without that property being guaranteed by the basic architecture, would make the mind's ætiology more credible. In much the same way, a theory of the structure of trees which shows how it is possible to trim them to look like giraffes is going to be more illuminating than an account of the structure of a giraffe-tree on which its giraffe shape is guaranteed, even though a trivial theory of that kind is to be had for the asking—or at least the measuring.

5 Explanation II: Implementation and Levels of Explanation

It seems that Fodor believes that taking a connectionist approach to the explanation of our apparently productive and systematic capacities is to make a kind of mistake about explanatory levels. I take it that this kind of objection might also be leveled at my claim that an ætiological explanation will go a fair way towards being sufficient for the explanation of these properties.

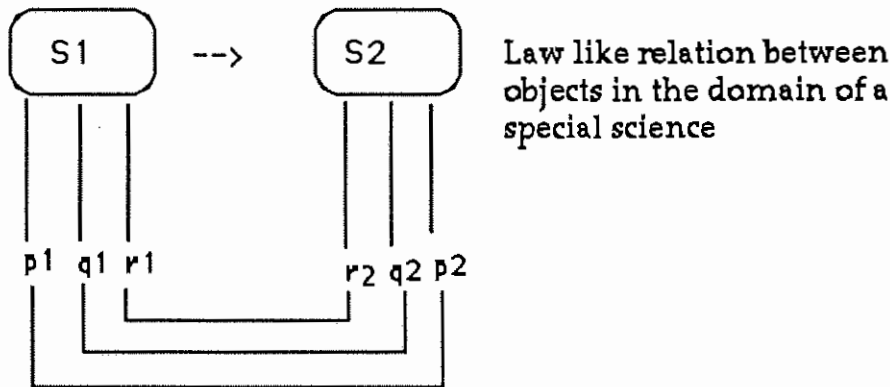
The idea is that there are lots of different levels of explanation. Of course the changing states of the brain can be explained by some neurophysiological story, and perhaps some ætiological account can be given of why certain high level regularities appear in these state changes when viewed from some high level. But neither of these is a *psychological* explanation of the supposed productivity and systematicity of thought, for that would have to be at the psychological level: and the only explanation going at *that* level is the Language of Thought. As Fodor and Pylyshyn write:

It seems certain that the world has causal structure at very many different levels of analysis, with the individuals recognized at the lowest levels being, in general, very small and the individuals recognized at the highest levels being, in general, very large. Thus there is a scientific story to be told about quarks; and a scientific story to be told about atoms; and a scientific story to be told about molecules... ditto rocks and stones and rivers... ditto galaxies. And the story that scientists tell about the causal structure that the world has at any one of these levels may be quite different from the story that they tell at the next level up or down. The methodological implication for psychology is this: if you want to have an argument about *cognitive* architecture, you had better specify the level of analysis that's supposed to be at issue. [Fodor and Pylyshyn 1988 p 9]

This, then, is the Fodor and Pylyshyn doctrine about levels of explanation. We can agree that the world is organized at many levels, many of which are scientifically (or otherwise) interesting. But notice that for Fodor and Pylyshyn there must be *causal structure at many levels*. The Big things at the high levels cause things to happen to other Big things, just as the little 'uns cause things to happen to other little 'uns, while the Big things are composed of the little things to tie the whole story down respectably. I think that it is this requirement that *causation* proceeds at every level which, in a very subtle way, commits Fodor and Pylyshyn to the Language of Thought before all the evidence is in. The arguments of pt. 1 of this thesis, though,

will allow us to let *explanation* proceed at many levels, while leaving it open at what level the real causal processes of the universe proceed.

It is perhaps timely to consider a diagram made famous in Chapter 1 of Fodor's *The Language of Thought* [Fodor 1975 Ch. 1] which describes the relationship of the special sciences to physics.



Lawlike causal relations between each of the things at a lower level which can realize S₁, and member of the class of realizers of S₂

There are two ways of reading a diagram like this. The lawlike relations which hold between S₁ and S₂ can be taken to be *causal* laws that justify the claim that S₁ *caused* S₂ or else in a less orthodox way that S₁ may in some circumstances *explain* S₂. In this case S₁ would explain S₂ in virtue of its being a good description of genuine regularities in the world, and in virtue of the fact that in general one of its realizations will *cause* one of the realizations of S₂.

As I argue in ch. 5, there are good reasons for preferring an account in which the high level properties do not do the causing. Firstly, if there is some lower level causal interaction going on which is sufficient for the state changes of the system, then the high level causal interactions are idle. What purpose would overdetermination of the causal history have? Secondly, and relatedly, the higher level properties are related to the lower level ones by

relations of supervenience and multiple instantiability. But in each particular case only the actual instantiation of the higher level property is present; so to claim that the higher level property is causally efficacious seems to rely on the other possible but non-actual realizations doing some causal work. But in fact, it does not matter to the particular case what is non-actual.

The motivation for Fodor's claims about causation proceeding at the high level is, of course, to avoid reductionism. He wants to preserve the special sciences as genuine fields of inquiry, not reducible to physics. This can be achieved, though, by noting that regularities can be observed at high levels which may not be observable at lower levels. And if high level entities can *explain*, partly in virtue of the fact that they token the existence of a causal process at a lower level, that may be enough.

5.1 Why High Level Causation Brings on the Language of Thought

Fodor and Pylyshyn believe that the scientifically explanatory higher level properties are causally efficacious ones. Psychological properties are certainly high level; and they take it as uncontroversial that they are causally efficacious. The methodology is this. Look for high level generalizations, and if one is found which looks like it has the desirable properties of simplicity and power, then find the causally efficacious items in that domain which explain phenomena at that level. In the psychological case, you *start* with content (which even Fodor admits, is an extrinsic property, and is causally neutral with respect to behaviour) and then look for the things at a high level which *have* the content—the semantic tokens which are realized in the brain. Now there are various behavioural regularities which need to be explained, and they are regularities when seen from the perspective of

content. Content is not causally effective, so it must be whatever *has that content* which is causally significant. So a taxonomy of the mind is given in which it is mapped, at some high level, in a way which *mirrors* [see Fodor 1987 pp 12–17 and 1985 pp 93–94] the content.

So far so good. In fact at this point we could have our Language of Thought, although it would be the too easy one described in § 2. It would be an *extrinsic* property of the brain; it would be the mere mapping of our semantics on to the brain to create a stipulatory syntax. The crucial move comes when it is assumed that high level true powerful generalizations must have causal structure. As soon as this is required, Fodor must insist that his too easy Language of Thought is pretty close to the Real Thing; because if our stipulatory syntax is to have *causal* structure then it had better not be an *extrinsic* property of the brain; no constitutively extrinsic properties are going to have causal powers over narrow behaviour. If one of our syntactic states is required to *cause* another, then it just *has* to have intrinsic causal powers.

This may be what underlies Fodor and Pylyshyn's insistence that ætiological explanations or connectionist explanations are making a level mistake. If causation proceeds at every level, then a causal explanation at one level will not nearly exhaust our requirement to explain causally. And if explanations at the psychological level are bound to be causal, and if the psychological level mirrors our semantics in the way Fodor thinks it does, then intrinsic entities which can enter into the right kind of causal relations

specified by that schema are required.⁹ If we had to have those, maybe a strong Language of Thought would indeed be required.

This requirement can be sidestepped if, as above, we do not require that all the high level properties be causally efficacious ones. Removing that requirement allows us to assess unblinkeredly whether or not the kinds of token syntactic states postulated by the Language of Thought are likely to feature as part of the intrinsic furniture of the mind.

5.2 The Fallacy of the Implementation Fallacy

It is the concern about levels of explanation that I believe underlies Fodor and Pylyshyn's argument that connectionism could only be a theory of cognitive *implementation*, not of cognitive architecture. They take it that even if it turns out that the kinds of networks that connectionists hypothesize actually exist, this would be no argument for connectionism as a *cognitive* architecture. Rather, they take it that this would merely show that the classical architecture (complete with Language of Thought) was *implemented* in a connectionist network.

⁹If you have a sufficiently weak view of supervenient causation, then high level properties might be able to be causally efficacious without having intrinsic structure, just so long as there is intrinsic structure picked out by the basic causal properties on which they supervene. Thus you might be able to buy into my account while believing with Jaegwon Kim [Kim 1979] in supervenient causation at high levels. You might also be able to have high level causation and yet do without the need for intrinsic structures at that level if you agree with Peter Menzies' more freewheeling account of high level causation in [Menzies 1988] in which he argues for the causal efficacy of relational properties, unhampered by the requirement that they be reducible to the causal properties of a base state. For arguments against these two positions, however, see ch. 5.

Fodor and Pylyshyn make a lot of the fact that different computer programs are able to be emulated by others, or that machines of one kind can be emulated by machines of another kind. The actual physical instruction set of one machine can be emulated by operations which consist of a series of operations in the instruction set of the other machine. The emulated machine is said to be a *virtual* machine.

Just as in the language and grammar case, Fodor and Pylyshyn take as certain views from an area where the philosophical problems abound. After all, under what conditions are assertions about the real logical architecture of a machine assertable? If the architecture of one machine is emulated on another, then in virtue of what is the logical architecture of the emulated machine the 'true' architecture of the machine? There are certainly pragmatic considerations that make the notion of the true architecture useful, and perhaps explanatory. A number of considerations might be brought to bear.

- 1) It is when the machine's output is described as the output of the logical architecture—not as the indescribably complex and apparently patternless output of the machine architecture—that the machine is intelligible to us.
- 2) The machine is *designed* to be input/output equivalent to the actual machine that the virtual machine is emulating.
- 3) Perhaps a little stronger: the algorithm which underlies the virtual machine's design, or even the design specification of the virtual machine's architecture, features *causally* in the creation of the

machine emulation. The programmer looked at the design of the first machine, and the causal process which made the emulation was mediated by that design. Even here, though, this doesn't guarantee that anything is left *intrinsically* after the programs have been compiled and all trace of the original structure other than its input/output equivalence is lost. Perhaps the design's having featured causally is a good criterion for asserting that the virtual architecture is the true architecture, but that wouldn't guarantee the kind of intrinsic causally efficacious states that Fodor demands.

None of these provides very strong or robust criteria for being intrinsic causal realists about the states in computers which map onto high level languages—the folk languages, if you like. Even if you think they *are* immensely powerful criteria, they don't seem to have plausible or sufficiently powerful analogues in the case of psychology. No-one programmed us by following a Language of Thought implementation manual, so the Language of Thought wasn't instrumental in our being intrinsically structured the way we are. In case (2) it is not true that, even if we do act in a (more or less!) input/output equivalent way to the Language of Thought that it is the *only* architecture to which we act in an input/output equivalent way, and (1) is just too weak to give us the Fodorian goods.

More than any of these points, though, I want to stress that this talk of connectionism (or any other architecture) being just an implementation doesn't solve any problems about the Language of Thought, it merely sweeps them into another area. In fact I think that it remains almost *exactly*

the same problem in its new cybernetic home. It just reemerges as the problem of the causal efficacy of representational syntactic states in complex and abstract computing environments. A definitive philosophical answer to the one problem will certainly help the other, but it does no good at all to just pretend that the problem is solved in one arena, and apply it in another.

The philosophical diagnosis of why it is that Fodor and Pylyshyn think *it is* so obvious that the problem is solved in the machine case—i.e. that virtual machine architecture is the true architecture at that level—is much the same as in the Language of Thought case proper. By insisting that high level generalizations are true in virtue of causal connexions between high level entities, they become committed to a real, intrinsic causal structure at that level, even in the absence of any independent intrinsic structural motivation.

This does not mean that I think having features which are directly involved in causation is a necessary condition which true structural architectures must meet; but rather that in the absence of this condition we are owed an account of why we should favour one architecture over another in intrinsic terms.

6 Conclusion

I see the points made above as having a significant bearing not just on the Language of Thought hypothesis, but on traditional functionalism as a whole. Certainly you can have a functionalism at the highest level—that of inputs and outputs under a certain description—but this, with a suitable story about the real functions of particular classes of outputs and inputs in the environment, will give you the external functionalist story about how we might go about vindicating belief desire psychology, not Fodor's. It is when the taxonomies created at this high level are then turned back on the brain,

and it is assumed that there are structural features which implement the high level story in a way which mirrors the taxonomy that the Language of Thought is born, and the worries set in.

I end with a conjecture. Just what counts as a sufficient motivation for something's being a bona fide intrinsic structural property is unclear, the suggestions in ch. 4 offer only the beginnings of an answer. But I think that the following is at least a pretty good methodological heuristic. When investigations with very different interests end up taxonomizing things in similar ways, it is not a bad bet that there is some structural property at work. Physiology, for example, is often concerned with relatively abstract function; but often physiologically significant taxonomies produce objects which are realized by objects which also appear in anatomical taxonomies, with their quite different interests and guiding motivations. This is perhaps enough to say that, for all that 'heart' is a functional term, it is in fact structurally realized (see ch 4).

So too for psychology and the Language of Thought. With a proper conception of what needs explaining, and how it is to be explained the fact that the Language of Thought is a neat functional specification of (some of the capacities of) minds is not enough. To give us the strong version of the hypothesis we need an independently motivated taxonomy of the mind—perhaps from the neurosciences—to come up with taxonomies the objects of which turn out as a matter of fact to realize the syntactic tokens beloved of Fodor. I await with interest the results of that enterprise.

Appendix

Explanation and the Reference of Kind Terms

Table of Contents

1	Why-Questions and Kind Terms.....	292
	1.1 Pragmatics.....	295
2	The Programme	296
3	Causal Theories of Natural Kind Terms and the Sameness Relation	297
4	Natural Kinds and Natural Kind Terms.....	302
5	Reference Shift and the Bohr Atom	302
6	Some Twin-Earth Cases.....	306
	6.1 Functional Kinds and Microstructural Kinds.....	311
7	Final Remarks.....	312

"Why did those brigands relieve Alexander of those lumps of rock when he was ambushed in Egypt?" asked one of his lieutenants on seeing that none of the spears had been taken.

"Didn't you know?" replied another, "They were gold".

The children are gathering materials for their end of year chemistry project: a big wall chart of the periodic table, with samples of as many elements as can be easily and safely gathered together. After some hesitation the head of the Chemistry Department removes a ring from her finger and puts it in a glass container.

"Why are you putting your ring on the wall chart, please teacher?" asks a perplexed student. "Because it is gold" she replies.

"Why can that kissing gourami survive in such muddy water?" asks a disgusted parent staring into the foetid depths of what passes for little Jo's aquarium.

"Because it is an anabantid" she replies smugly.

"And why does it have those ugly little barbels at the mouth, then?" The smug smile spreads intolerably, and she pronounces "Because it is a cyprinid."

1 Why-Questions and Kind Terms

The point of these little stories, and countless others that could be told, is to suggest that there is a purpose to grouping things in kinds, and that, further, a particular can be grouped in many different kinds, depending on the purpose of the grouping or the interests which underlie it.

I hope that this is not a controversial point, as I want to use it to make other claims about the reference of kind terms. The idea is that in these stories, the invocation of kind membership has an explanatory rôle. Answers to why-questions are provided by assertions of kind membership. It is because they were *gold* that Alexander's 'rocks' were taken, and it is

because it was gold that the teacher put her ring on the wall chart. Being an anabantid accounts for a gourami's capacity to breath air, but it is being a cyprinid (a higher-order state of affairs) which accounts for the barbels.

My claim is that the rôle which kind-terms have in languages is having some explanatory function: an explanatorily useless kind term is a useless kind term. Whether explanation figures significantly in the ætiology of kind terms is a different matter: I suspect it does, but not always. It is certainly true that being able to recognize different samples of H_2O as members of a kind is something that has great survival value, but that is a value which is pre-linguistically significant. Even in linguistic communities, while it may be important to be able to talk about water and communicate its whereabouts, there may well be no need for explicit explanations to be offered by language users which feature claims of kind membership as their explanantia. So the claim that having an explanatory rôle is constitutive of kind terms is a dispositional one: a term is a useful kind term only if it *could* be used in explanations.

The next thing to notice about these stories is that the way in which kind membership is invoked is different from case to case. The first two stories both feature the punchline 'it is gold' as an explanans, but it is different features of gold which seem to do the work. In the first case it is the *value* of gold which seems to be significant. To that extent the kind term may not be one which, at that time, was grounded in intrinsic properties at all. More plausibly the colour or ductility of the material may have been sufficient and necessary grounds for these people's concerns with it. The other case has the atomic number well to the fore as a grounds for grouping. As a good chemistry teacher, our teacher would no doubt have little concern with colour of her ring so long as it has an atomic number of 79. As for the fish, which feature of the fish one is trying to explain determines whether it

is the fishes order or family which is the explanatorily relevant kind term. This last relationship—between different explanatory functions of taxonomical levels has been explored by Mark Platts in his *Explanatory Kinds* [Platts 1983].

The claim is that this way of viewing the kind relation makes the motivations for various connexions among kind terms more apparent. Take, for instance, some object which is both a citrus fruit and a lemon. It appears to be included in a number of natural kinds—why? Because they have different explanatory functions. In particular there is a trade-off between explanatory power and explanatory range. Its being a fruit explains what it has in common with a melon, and explains why they both have seeds. Its being a citrus fruit explains the segmentation—which its being a fruit does not, though at the cost of ignoring what is in common with fruits which are not citrus. The more general kind term—fruit—has a wide explanatory range, giving a more comprehensive generality to its explanatory function. It points to what the object has in common with a wide range of other things, and places it in a large and important class whose membership criteria must be involved in any successful picture of the botanical world. As we descend to the family and species level this generality is sacrificed for more *power*. More fine detail of a lemon can be explained by the fact that it is a *citrus* fruit, than can be explained by its being just a fruit, and more detail still (at the cost of generality) by the fact that it is a lemon.

Serving as explanations seems to be an important rôle of kind terms. Pointing out that some ϕ belongs to a kind can function as an explanation of its characteristics, or of the effect it has on other things. Assertions of kind membership are appropriate when membership of that kind subsumes the characteristics of the individual under the general characteristics of the relevant kind, and in those cases under any theory of explanation we have

discussed it can be asserted that it is *because* the individual is a member of the kind that it has the characteristics it does.

Subsumption of particulars under kinds plays a crucial rôle in the most elementary attempts to explain action in a normalizing way—‘He put the liquid in the stew *because (inter alia)* it was *water*’—and in the crudest to the most sophisticated attempts to regularize people’s surroundings.

1.1 Pragmatics

The importance of pragmatics in the determination of the exact rôle which kinds have in providing answers to why-questions can be seen by considering some questions which might invoke kind membership as an answer.

- (1) Why was this liquid consumed by thirsty Fred?

As it stands the question is in need of disambiguation by pragmatic analysis. Suppose we provide contrast classes such that we get:

- (2) Why was this liquid [rather than: A, B, C]
consumed by thirsty Fred]?

Let us further provide a couple of possible relevance relations:

R1: relevance in respect of some phenomenal property of the liquid which makes it desirable to Fred.

R2: relevance in respect of some microstructural fact which figures in the relative potability of the liquids.

Both these relations allow of the same answer to the question: the liquid is water, only in each our attention is being directed to different properties of water which do the explanatory work. In the first case it is the functional rôle at the level of sensory perception that counts, in the second it is the actual details of the interaction that H_2O has with the body that are being invoked. If our concern is with the specific chemistry of water understood rigidly, then when we answer “water” to the question, the kind which serves as that answer will be one construed as a compound. If our concern is with the phenomenal properties that water has, then anything else which has the right properties would fall under the umbrella of the level of explanatory generality required to answer the question. In fact I take it that nothing but H_2O actually does have these phenomenal properties, but if we imagine one of the worlds envisaged in Mellor’s [Mellor 1977], where the natural laws are different and the “water rôle” is filled by a substance of different microstructural constitution, then the point becomes clearer

2 The Programme

My goal in using the explanatory account of kind terms is threefold. First, to develop an account of the reference of kind terms which will preserve stability of reference where I find it plausible, whilst allowing reference to drift when it clearly has. Second, to develop a theory of kinds of kinds, varying along lines of explanatory interest, and which provides a plausible answer to the problem of reference stability. Third, to gesture towards a return to the separation of semantics and metaphysics and stop insisting that whatever the correct physics or metaphysics of the case may be, that this determines meaning.

The problem of reference stability I take to be this: purely descriptive theories of reference¹, amongst other shortcomings, make the reference of kind terms shift with the winds of whatever beliefs are held about the kinds, and indeed often ensure that the terms do not refer at all. The case of the Bohr atom is often cited as a paradigm. Almost every description believed by Bohr to be true of atoms is false, so how under a description theory could he be said to be referring to the same entities that we refer to in using the word. Pure causal theories, on the other hand, have the opposite vice. Reference is stable where the intuitions of someone trying to use languages would make one think otherwise. If it is a Zebra that is ostended to by an elder of some community in dubbing the term, then the other equally edible (but now extinct) stripey quadrupeds that were brought back by the hunters were brought back in error. But try taking the non-zebra steaks of the revellers and see what happens.

3 Causal Theories of Natural Kind Terms and the Sameness Relation

An account of natural kind terms that has a lot of currency stems from [Kripke 1972] and [Putnam 1975c], and involves grounding natural kind terms in an act of baptism. A paradigm case is ostended to, and the extension of the kind is supposed to be everything that is the same as it. Well, no, *that* can't be right, only one thing can be the same as it—the thing itself. Now I am not suggesting for a minute that Kripke, Putnam or Sterelny in his [Sterelny 1983] fall into that trap, but this is where the fun begins because we have to decide *in what respect* something has to be similar to the

¹As will become apparant, the account of reference at work in this account of kind terms is a brand of *causal descriptivism*. See [Kroon 1987] for a general defence of this view.

paradigm case in order to count as being in the same kind. Plainly the identity relation won't do.

One valiant attempt might be mounted using the distinction between intrinsic and extrinsic properties. If we take spatio-temporal location to be an extrinsic property, then perhaps sameness of intrinsic property will capture what we are looking for when we are trying to establish what sorts of things something else must be similar in respect of so as to count as a member of the same natural kind.

This won't do either, however. Hugh Mellor in his [Mellor 1977] cites the case of isotopes to claim that it may always have been the case that what appeared to be a natural kind was in fact two kinds. His example is the two most common isotopes of chlorine, which were presumably included in the extension of chlorine before their discovery. Mellor's claim is that chlorine has been found not to be a natural kind at all but rather a mixture of kinds. This I find unconvincing; whatever the theory of natural kinds that does the best explanatory work in chemistry, it had better make elements natural kinds, and in any case there is no problem with a substance belonging to two natural kinds. The point remains interesting, however, in that it shows that sameness of intrinsic properties won't do as a criterion for picking out the relevant sense of sameness for the baptism to go through. Not only can there be intrinsic differences between members of a kind, but they can be ones which are non-trivial from the point of view of a theory of kinds. If it were possible to restrict the variations of intrinsic property to ones not relevant to kind designations (such as mass of the particular sample, or degree of admixture of other substances) then perhaps the intrinsic properties theory might work. What the isotope case shows is that there can even be variation amongst members of a natural kind which are important for the establishment of kindhood. In which case we are reduced to saying

that the extension on the kind term is determined by those intrinsic properties which constitute the criteria for membership of that kind, which begins to undermine the purely causal account of natural kinds.

If sameness of intrinsic properties won't do, then what will? Putnam introduces sameness relations which seem to include in them some ideas about kindhood. He says of water, for example, that the extension of water includes whatever bears the *same_{liquid}* relation to a paradigm sample of water.

Now, what are we to make of such a relation? The first point is that even were the relation itself unproblematic and not interest-related, its introduction is certainly related to the explanatory interests of the introducer. The introduction of the *same_{liquid}* relation presupposes that, for example, we are not interested in the class of bodies in the same physical state. It is *water* that we are baptising, not liquids in general.

The stipulation that we should look at the *same_{liquid}* relation rather than the *same_{physical state}* relation corresponds to differences of contrast class in the various explanation requests that can be made about the substance. Take for instance Platts' Ur-question "What are these?", and insert a contrast class giving us "What are these? (rather than those)". If the ostended substance is in fact water, and the contrast class pointed to as 'those' includes other liquids, then we can be certain that the *same_{physical state}* relation is not appropriate. If the contrast class includes only non-liquids (perhaps a sample of solids, gases and some constrained plasma!) then it is more likely that the *same_{physical state}* relation is the relevant one. If the contrast class includes some samples of water, then the *same_{liquid}* relation is once again counterindicated: if it contains only samples of water, then some similarity relation at a finer level of microstructural detail may be

called for, or else grosser details like types of impurity in the solution could be appropriate, depending on the purpose of the comparison. Of course the contrast class alone cannot hope to completely determine the appropriate relation. If the contrast class is small and includes only non-liquids, then it might be possible to argue that the relevance relation is radically underdetermined by the contrast class, since both water and liquids in general are in contrast with this class. This ambiguity could be undermined by appeal to a convention that the relevant contrast is in respect of the least general features not ruled out by the contrast class. Thus if there are no liquids in the contrast-class, and water vapour is included, then the *same_{physical state}* relation may be the least general relation which distinguishes the sample from the contrast class. This convention itself, though, would have its motivation in the pursuit of systematic taxonomy, and different explanatory interests of a more pragmatic (in the vulgar sense) nature would easily vary which relation is appropriate.

The relations themselves seem in need of being spelt out. With fixed *same_x* relations we have seen that we need pragmatic considerations to sort out exactly which one is appropriate, but how are we going to fix the relations anyway? Even with an account of liquidhood, we don't automatically have an account of what it is to be the *same* liquid. It certainly won't be the same intrinsic property story. No two samples of water—and very few pairs of water molecules—have exactly the same intrinsic properties. Two water molecules may well have electrons at different energy levels, but surely they belong to the same natural kind.

One attempt at a solution is to say that we do not *need* an account of the *same_{liquid}* relation, or more generally of the same kind relation. Not only is it an empirical discovery that, given a same kind relation, two samples are

tokens of the same kind, it is also an empirical discovery what the same kind relation is. In [Sterelny 1983] Kim Sterelny wrote:

Crucially, the nature of this relation is only discoverable by scientific research. It follows that users of a natural kind term need not, and normally will not, have a necessary and sufficient condition for membership of that kind.

It is difficult to see how this avoids a kind of regress. Certainly, given a same kind relation, whether it holds between two substances is a matter for empirical discovery. Equally, given a theory of theories of same kind relations, it will be an empirical discovery whether two pairs of substances share the same kind of kind relations. But the crucial interest dependence is still there in the determination of what kind of kind relation is called for.

An example: suppose that one inquirer is interested in microstructural properties which are necessary parts of sufficient but not necessary conditions for the production of the phenomenal properties of certain paradigm instances of a substance. Answers in terms of these properties are eligible to be answers to her why-questions given her relevance relation. Another is interested only in phenomenal properties, or perhaps functional properties.

Now, *given* the first interests, certain kinds of sameness relation have clear advantages. This is something that can be discovered, so it is reasonable to say that, for example, the same kind relation in the case of elements could be cashed out as the same_{atomic number} relation, and it is certainly a matter for empirical discovery whether any two pairs of substances are linked by the same kind of kind relation: whether, say, there are microstructural similarities between members of one pair, but only phenomenal or functional similarities between the members of another.

4 Natural Kinds and Natural Kind Terms

Perhaps I should come clean here about what I think about *natural* kinds. I am a good child of my times: I am interested in the microstructural determination of the causal powers of things. I want my explanations to be useful both technologically and in the natural-scientific explanation of the world. To search for a minimal set of kind groupings that will regularize the operations of the natural world, and, if deployed, be sufficient to figure in explanations of that type is an important and admirable pursuit—even if it occasionally has unfortunate consequences. I have not got a lot to say about how one should characterize this kind of explanation: and that is what is required for a substantive *theory* of natural kinds. In other words, the problem of natural kinds reduces to the problem of what properly constitutes natural explanation as we understand and practice it. Natural kinds will be the kinds that are invoked in natural explanations.

Without trying to give an account of natural explanation this idea can be used to sketch an account of how one can allow the contextual interest-relatedness of explanation to play a rôle which allows beliefs and dispositions to play a more determining rôle in reference of kind terms than pure causal theories would allow, and still preserve a plausible amount of continuity of reference over theory change, thus avoiding one of the major pitfalls of description theories.

5 Reference Shift and the Bohr Atom

Consider the problem of the Bohr atom, which Putnam cites in his [Putnam 1973]. The problem is basically that so many of the things which seemed to count as necessary conditions for falling into the extension of ‘particle_{bohr}’ ‘electron_{bohr}’ and therefore ‘atom_{bohr}’ are not met by anything (let alone particles, atoms or electrons), that Bohr’s terms seem to fail to refer. Yet it

would be nice to be able to say that the theories that we now have about atoms are better theories about atoms, rather than theories about quite other things which, we hope, exist.

Let me first say that if Bohr himself were to have insisted that the beliefs—even those that look like definitions—that he had about atoms were indeed necessary conditions for inclusion in the extensions of his terms, *and was not disposed to change his mind in the light of experimental evidence*, then in the sociolect of Bohr and likeminded people the terms wouldn't have referred. This doesn't worry me, however, because then he wouldn't have been doing science. It would mean that the kinds of interests he had, and the kinds of dispositions to change his mind in the light of theoretical and experimental change precluded the degree of reference stability needed for natural explanation.

The model of intensionality that seems to stem from Frege, in which intensions are a list of properties which give necessary and sufficient conditions for falling into the extension of kind terms, is far too simple. Even a simple cluster theory like Strawson's in [Strawson 1950, 1959] is far too simple. One can agree with causal theorists that indexicality plays an important rôle in kind terms both natural and non-natural; but the importance of indexicality is in picking out the object of belief. It ensures that if the extension of 'gold' for us is fixed, at least in part, by the beliefs and dispositions of physicists, it is their beliefs and dispositions about actual gold and not something which lurks in some remote possibilium that counts. The beliefs, and dispositions to change belief, have an important determining rôle, but so does indexicality and causation, since which result will be realized, given a complex set of dispositions to change belief in differing circumstances, will depend on the actual nature of the paradigm sample. So in every so-called baptism—and in the multiple regroundings in

[Sterelny 1983]—both descriptions, dispositions to change description, and the actual nature of the causal interactions which take place play a part.

What moderates these indexical beliefs, though, is a set of other beliefs and dispositions which determine what the salient kinds of similarity are in the case of baptisms. These beliefs can be weighed against each other, and so determine which beliefs should be scrapped and which retained in the light of empirical discoveries. These dispositions to change beliefs are what makes it possible for it to seem that *all* the beliefs and dispositions which one can have about something can turn out to be false. The case of cats which turn out to be Martian agent spy robots is one of these. This example of Kripke's is supposed to show that all our beliefs could indeed turn out to be false. The idea is that we are supposed to imagine that there have never been any members of the species *Feline Domesticus*, instead all purported cat-sightings have been caused by Martian spy-robots with complex hologram generators and presumably some method of inspiring tactile and olfactory sensations. Kripke's contention was that it would be plain that there could be no disputing that there have always been cats—cats must be whatever it was that caused cat-impressions—it's just that all our beliefs about cats have turned out to be comprehensively false. I am not sure that it is so obvious that a community is bound to decide that there have always been cats, but let that go. Even if it did turn out that a community decided that 'cat' referred to the robots, it would be in virtue of certain beliefs being prioritized—such as cats are the things which chase after the mice in the back yard. What is more, such an example shows how fractured a speech-community can be. While the person in the street may indeed decide that cats are robots, biologists in virtue of different explanatory interests may well decide that the *biological* term 'cat'—which they thought co-extensive with the common one—did not refer and never did.

To return to stability of reference: it is the explanatory concerns which prioritize beliefs—typically, say, in favour of specification by causal powers—and the explanatory interests which give rise to the dispositions that govern the transformation of the actual clusters of first-order beliefs, which give us the required stability. If we stipulate that the intension of a term includes these sorts of higher-level dispositions, then intensions rule again—or at least they link together terms throughout theoretical change.

If in fact Bohr didn't insist that there would be no electrons unless they had a determinate position and momentum (or at least was disposed to withdraw this insistence in the light of evidence) there is nothing to stop his terms referring to the same things as our more recent ones, and this in virtue of the higher-order beliefs in the cluster associated with the Bohrian model. So long as the higher level explanatory interests remain constant, and are manifested in the theoretical practice of natural science, then successor theories can be said to be about the same entities.

What, then, is the rôle of a causal theory of natural kind terms? Well, if we allow the term 'natural' to be appropriated by the natural sciences (remembering that we cannot always read the term as used in historical discourse in this way) and their distinctive explanatory interests, then something resembling a causal account, with a recognition of the moderating rôle played by beliefs and dispositions, is probably going to be a fairly good theory. What this can be analysed out as, though, is that in natural scientific practice, because of the particular explanatory interests that govern it, *included in the intensions of its kind terms are many of the beliefs that we associate with causal theories of reference and natural kinds*. So the causal theory doesn't tell us what natural kinds are *ex nihilo*. It *stipulates* what they will be, given certain interests.

The determination of kinds by explanatory interests, then, gives an account of reference stability consistent with a partially description-based theory of reference by locating the stability in the explanatory interests of a speech community, expressed in the higher-level beliefs and dispositions located in the cluster. But what does it have to say about desirable non-continuity of reference? What, in particular, will be the result applied to the Twin-Earth examples of Putnam [Putnam 1975c] and their ilk?

6 Some Twin-Earth Cases

In these familiar examples, we are invited to imagine that Earth and everyone who lived on it had a twin. It is supposed to be a phenomenologically identical twin, but one which has microstructural differences. In particular, what goes under the name 'water' on Earth is H_2O , and on the twin what is called 'water' has a different microstructure. Putnam calls it XYZ.

This is supposed to show that meaning is not determined by the narrow psychological states of speakers; Earth people have the same concepts associated with the term 'water' as do the Twin-Earthers—indeed my *doppelgänger* may be in exactly the same narrow psychological state as I—but Earthlings are referring to H_2O , and Twin-Earthlings are referring to XYZ, all by means of causation: the reference is determined by the stuff they are in actual causal contact with. More than this, all and only H_2O falls into the intension of 'water' for Earthlings, and all and only XYZ falls into the intension of 'water' for XYZ.

This works perfectly well if we assume that the intension, in the broad sense outlined above, includes the explanatory interests that characterize natural science—or in the case of individuals, deference to those who do share those interests. The higher-level beliefs about the importance of micro-

structural similarity relations and so on will then, *together* with the indexical situation, determine the meaning of 'water_e' and 'water_{te}' even if the actual microstructures were not known. The reference is determined by the interaction of the environment and an intentional system whose functional states involve, in effect, a theory of kinds.

Putnam, however, makes a further claim. He thinks that even if we roll the clock back to 1750 it is still true that the meanings of 'water'_e and 'water'_{te} are different. I want to argue that an understanding of the possible changes in explanatory interests of communities can make this latter claim less plausible, by giving a non-*ad hoc* account of how the reference change and meaning shift from then to now might have taken place.

We saw above that, on my account, some criterion for the sorts of similarities which govern the same_{liquid} relation had to be given. This was given in terms of the beliefs about the importance of natural explanation in terms of microstructural features, or at least deference to experts with those beliefs, amongst the contemporary inhabitants of Earth and Twin-Earth. Is there any reason for supposing that these beliefs were prevalent in 1750? If they were, then what about 1550? This sounds like work for an historian of ideas.

If, as seems quite likely for early dates, or indeed for some contemporary cultures, these higher level beliefs were not prevalent, then the reference of the terms would not be governed by the explanatory interests of natural science. In which case we can assume that it was certain important phenomenological or functional properties which were important in governing what counts as similarity to a paradigm sample. In these situations, there seems little reason to suppose that the reference of 'water' in each world is different, since from the perspective of these worlds there is

nothing importantly different between H₂O and XYZ. Perhaps a test: if a time traveller could outline the microstructural differences, would they have cared?

Sterelny, in [Sterelny 1983], has two arguments against this. The first is that the fact that if Earth people were to have had commerce with Twin-Earthlings in 1750 they would have called XYZ 'water' is, while true, no argument that XYZ was always in the extension of 'water_e'. His claim is that what is being described is a change—a change in the reference of 'water_e' from H₂O to both H₂O and XYZ.

Is it a change? Not, I think, on my theory. On this account, things fall into the extension of a term if, were the appropriate interactions to take place between the intensional system and the environment, the thing would be taken to fall into the terms extension. This being criterial, if the beliefs and dispositions brought about by the explanatory interests of our 1750 terrestrials were such that they treat phenomenal properties as kind markers, XYZ and H₂O would indeed have been in the extension of 'water_e' before the interaction since the conditional was presumably true then. Sterelny says, though, that there would still be a change if commerce happened, and that it therefore cannot be used to warrant the conditional—and that the change is that 'water' would have ceased to be a natural kind term. This is a case of arguing from your conclusion back to it: if the reference of water now is unchanged from how it was in 1750, and it is now a natural kind term, then it must have been a natural kind term then, so any commerce after which it was no longer a natural kind term must have changed its reference. But why suppose that it was a natural kind term in our sense? Indeed if the kinds of beliefs and dispositions which fix the reference were such as to allow the inclusion of XYZ in the extension post commerce, then it could not have been a natural kind term in our sense. The

fact that in the real 1750, sans Twin-Earth, most of what fell into the extension of 'water' indeed constituted a natural kind, does not guarantee that water_{1750} a natural kind *term*.

Sterelny's second argument is that the reference of 'water' in our language does not depend on facts about other languages: it is not because Twin-Earthlings call XYZ 'water' that we would have done so had we had contact with them. For the conditional to be true, it must have been the case that we would have done so. But we know that we now would not do so, so it could not be the case that we then would have done so unless the reference of 'water' has changed from 1750 to the present. This, he claims, would be *ad hoc*, since there doesn't seem to be any obvious intuitive reason for supposing that we mean something different by 'water' now than we did in 1750.

Since there isn't and indeed never has been any XYZ, I doubt if there has been much if any change in the extension of 'water' since 1750. But suppose that there had been a deep well of XYZ lurking in the inner recesses of Uluru. Even had no-one found it, I do not think it would be *ad hoc* to suppose that the extension of 'water' would have changed from 1750, at any rate if the beliefs and interests of eighteenth century people did fix on phenomenological or functional properties as salient. (In fact, even if natural science could be said to have been fully under way, it is unlikely that many people would have deferred to scientists on so common a matter as what counts as water). Of course, this has changed; today most people defer to experts for their views on substancehood, and the experts' explanatory interests are causal-taxonomic. If it is announced in *Nature* next week that it is generally agreed amongst researchers that much of what we had previously taken to be water was something quite other, then our speech community would probably bow its collective head and accept in puzzled

tones that the boffins have spoken and mutter about the wonderful and strange discoveries made by science.

Sterelny has another swipe at the matter: suppose that Earth and Twin-Earth have some biochemical differences. In particular, XYZ is poisonous and foul tasting to Earthlings, and similarly H₂O to Twin-Earthlings.

Certainly this is a case where the extensions of 'water'_e and 'water'_{te} are different. Sterelny's claim is that this difference could not be one connected with descriptions or beliefs, since it is a difference which would only be effective with causal connexions between individuals and the relevant foul counter-water. But why does this requirement insulate the difference from being one which is partially determined by beliefs and dispositions? After all, whatever water was considered to be in 1750, standard water must have been paradigmatically potable! (I am not sure about earlier periods—perhaps in the sixteenth century potability might have been more usually predicated of alcoholic beverages, at least in urbanized areas!). Of course this need not be determined by psychological state viewed solipsistically—to that extent the slogan about meanings not being in the head is right. But the belief involved is that water must be potable to me (or us). Of course there is an indexical element to this belief; but it is a partially determining belief none the less. If the belief were that water must be potable to people, where the required similarity relation between the paradigm person and the term's extension were itself phenomenological in a superficial way (it must be potable to anything that looks and talks like a person, say), then the extension of 'water'_e would indeed include both H₂O and XYZ. Why, in any case, should taste be any more problematic a phenomenological property than appearance? The partial determination of reference by descriptions whose relative importance is determined by explanatory interest is only problematic if the descriptions

are supposed to be non-indexical: as though earthlings are required only to be concerned with how things taste or appear to twin-earthlings.

6.1 Functional Kinds and Microstructural Kinds

This debate about the extension of 'water' is sometimes seen as part of a debate about whether natural kinds are functional kinds, or whether they group things together according to the causally relevant microstructural properties which determine their functional properties. Hugh Mellor, for example, has argued *contra* Putnam that heat is whatever fills the heat rôle, rather than being necessarily what fills that rôle in our possible world.

This seems to me a fruitless argument which may be dangerously terminological. Let us agree on some terminology. Let 'natural kind' designate the kind of kind term that does the best explanatory work relative to the explanatory interests of the natural sciences as they are presently constituted. Exactly what these kind terms are is a question for a substantial theory of natural kinds; I do not have an answer here. But we should be careful, when considering controversies like that between Mellor and Putnam, to distinguish two possibilities. Are they arguing about whether the terms used in language and are normally called natural kind terms, are functional or microstructurally specific in their constitution, or are they arguing about whether functional or microstructurally specific terms do the best job in science? The answers to neither of these questions is clear; my guess is that in both cases both kinds of kinds have a rôle, but in very different ways. The case of the common English terms is likely to be particularly complex. There is nothing to stop there being functional kind terms, fixed immutably by descriptions, coexisting with natural kind terms (which may or may not have a functional element) and their being coextensive at a given time. And it may be very unclear which kind of kind

term is meant by a given speaker at a given time. Regrettable indeterminacy, perhaps, but one which mirrors accurately the confusions people get into when asked to specify their meanings carefully.

7 Final Remarks

Of course explanatory interests of communities are enormously diverse, I am even sympathetic to the claim in [van Fraassen 1981] that interests vary to the point where what counts as the *essence* of something, and is therefore a condition of counting as that thing for a given explanatory purpose, varies with context frequently. But this does not mean that we cannot find out how communities prioritize and balance explanatory interests and beliefs about kinds. For Alexander's lieutenant, it is an extrinsic property of gold that explained the action—its value. It may even have fixed it: who cared what other properties (within reason) it had so long as it paid the bills? But there were lots of other beliefs and interests at work in the community, and to give an account of the semantics of 'gold' in the community is to see how it functions systematically, and to see how individual beliefs and interests interact with each other. This is the rôle of semantics, and ever since the first publication of 'Naming and Necessity' in *The Semantics of Natural Language* [Davidson and Harman 1972] this rôle and the rôles of the scientist and metaphysician have been confused with it. In pointing out the inadequacy of theories of meaning being always *only* in the head, the causal theorists of everything have ignored the interactive nature of reference determination. Let us no longer confuse theories of what 'gold' meant or means, with a theory of what 'gold' is.

References

- [Achinstein 1983] Achinstein, P., *The Nature of Explanation*, Oxford University Press, N.Y. 1983.
- [Belnap and Steel 1976] Belnap, N.D. Jr., and Steel, J.B. Jr., *The Logic of Questions and Answers*, Yale University Press, New Haven 1976.
- [Bigelow and Pargetter 1987] Bigelow, J., and Pargetter, R., 'Functions', *The Journal of Philosophy* 84:1987 pp 181–196.
- [Bohm 1957] Bohm, D., *Causality and Chance in Modern Physics*, RKP, London, 1957.
- [Braithwaite 1953] Braithwaite, R.B., *Scientific Explanation*, Cambridge University Press, Cambridge 1953.
- [Brand and Harnish 1986] Brand, M. and Harnish, H., ed., *The Representation of Knowledge and Belief*, University of Arizona Press, in the series 'Arizona Colloquium in Cognition', Tucson, Arizona 1986.
- [Brand and Walton 1976] Brand, M. and Walton, D., (eds.), *Action Theory*, Reidel, Dordrecht 1976.
- [Bromberger 1966] Bromberger, S., 'Why-Questions', in [Colodny 1966] pp. 86–111.
- [Cartwright 1979] Cartwright, N., 'Causal Laws and Effective Strategies' *Noûs* vol. 13 1979. reprinted, with changes, in [Cartwright, 1983].
- [Cartwright 1983] Cartwright, N., *How the Laws of Physics Lie*, Oxford University Press, Oxford 1983.

[Chomsky 1965] Chomsky, N., *Aspects of the theory of Syntax*, MIT Press, Cambridge Massachusetts 1965.

[Chomsky 1980] Chomsky, N., *Rules and Representations*, Blackwell, Oxford 1980.

[Churchland 1981] Churchland, Paul, 'Eliminative Materialism and Propositional Attitudes' *The Journal of Philosophy* 78:1981.

[Churchland 1986] Churchland Patricia M., *Neurophilosophy*, MIT Press, Cambridge Massachusetts 1986.

[Collingwood 1940] Collingwood, R.G., *An Essay on Metaphysics*, Oxford University Press, Oxford 1940.

[Colodny 1966] Colodny, R. ed., *Mind and Cosmos*, Pittsburgh University Press, Pittsburgh 1966.

[Cummins 1984] Cummins, R., *The Nature of Psychological Explanation*, MIT Press, Cambridge Massachusetts 1984.

[Davidson 1980] Davidson, D., *Essays on Actions and Events*, Oxford University Press, Oxford 1980.

[Davidson 1980a] Davidson, D., 'The Individuation of Events' in [Davidson 1980] pp 163–80.

[Davidson 1980b] Davidson, D., 'Causal Relations' in [Davidson 1980] pp. 149–62.

[Davidson 1980c] Davidson, D., 'Mental Events' in [Davidson 1980] pp. 207–204.

[Davidson and Harman 1972] Davidson, D., and Harman, G., (eds.), *Semantics of Natural Language*, Reidel, Dordrecht 1972.

[Davies 1974] Davies, P.C., *The Physics of Time Asymmetry*, Surrey University Press, Leighton Buzzard and London 1974.

[Dennett 1978] Dennett, D., 'Current Issues in the Philosophy of Mind' *American Philosophical Quarterly*, 15:1978 pp. 249–262.

[Dennett 1978a] Dennett, D., *Brainstorms*, MIT Press, Cambridge Massachusetts 1978.

[Dennett 1978b] Dennett, D., 'Intentional Systems', in [Dennett 1978a] pp. 3–22.

[Dennett 1984] Dennett, D., 'Cognitive Wheels: The Frame Problem in AI' in [Hookaway 1984].

[Dennett 1986] Dennett, D., 'The Logical Geography of Computational Approaches: a View from the East Pole', in [Brand and Harnish 1986] pp 59–79.

[Dennett 1987] Dennett, D., *The Intentional Stance*, Bradford Books, Cambridge Massachusetts 1987.

[Dennett 1987a] Dennett, D., 'True Believers', in [Dennett 1987] pp. 13–35.

[Dennett 1988] Dennett, D., Review of [Fodor 1987], *The Journal of Philosophy* 1988 pp. 384ff.

[Devitt and Sterelny 1987] Devitt, M., and Sterelny, K., *Language and Reality*, MIT Press, Cambridge Massachusetts 1987.

- [Dretske 1972] Dretske, F., 'Contrastive Statements' *Philosophical Review* 1972 p411.
- [Eldredge and Gould 1972] Eldredge, N., and Gould, S.J., 'Punctuated Equilibria: an Alternative to Phyletic Gradualism' in [Schopf 1972].
- [Fodor 1968] Fodor, J., *Psychological Explanation*, Random House, New York 1968.
- [Fodor 1976] Fodor, J., *The Language of Thought*, Harvester Press, Sussex 1976.
- [Fodor 1981] Fodor, J., *Representations: philosophical essays on the foundations of cognitive science*, Harvester, Brighton 1981.
- [Fodor 1981a] Fodor, J., 'Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology' in [Fodor 1981] pp. 225–253.
- [Fodor 1985] Fodor, J., 'Fodor's Guide to Mental Representation: The Intelligent Aunty's *Vade Mecum*, *Mind* 94:1985 pp. 76–99.
- [Fodor 1987] Fodor, J., *Psychosemantics*, MIT Press, Cambridge Massachusetts 1987.
- [Fodor 1987a] Fodor, J., 'Why There Still Has to Be a Language of Thought', in [Fodor 1987] pp. 135–154.
- [Fodor and Pylyshyn 1988] Fodor, J., and Pylyshyn, Z., 'Connectionism and cognitive architecture: A critical analysis', *Cognition* 28:1988 pp. 3-71.
- [Fodor forth.] Fodor, J., 'A Theory of Content Parts I and II' semifinal draft, forthcoming in 1989.

- [Garfinkel 1981] Garfinkel, A., *Forms of Explanation*, Yale University Press, New Haven 1981.
- [Globus, Maxwell and Savodnik 1976] Globus, G., Maxwell, G., and Savodnik, I., *Consciousness and the Brain*, Plenum Press, New York 1976.
- [Goldman 1970] Goldman A. I., *A Theory of Human Action*, Princeton University Press, Princeton 1970.
- [Gould 1980] Gould, S. J., 'Is a New and General Theory of Evolution Emerging?', *Paleobiology*, 6:1980 pp. 119-130.
- [Hamblin 1958] Hamblin, C.L., 'Questions', *Australasian Journal of Philosophy*, 36:1958 pp. 159-168.
- [Hampshire 1966] Hampshire, S., ed., *Philosophy of Mind*, Harper and Row, N.Y. 1966.
- [Healey 1981] Healey, R., *Reduction, time and reality*, Cambridge University Press, Cambridge 1981.
- [Heathcote 1988 forth.] 'A Theory of Causality: Causality Interaction (As Defined by a Suitable Quantum Field Theory)', unpublished MS, 1988.
- [Hempel 1965] Hempel, C.G., *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, Macmillan/The Free Press, New York and London 1965.
- [Hempel 1968] Hempel, C.G., 'Maximal Specificity and Lawlikeness in Probabilistic Explanation', *Philosophy of Science* 1968 pp. 116-133.
- [Hempel and Oppenheim 1948] Hempel, C.G., and Oppenheim, P., 'Studies in the Logic of Explanation', *Philosophy of Science*, 15:1948 pp.135-75, reprinted 1965 in [Hempel 1965].

[Hookaway 1984] Hookaway, C., ed., *Minds, Machines and Evolution*, Cambridge University Press, Cambridge 1984.

[Jackson and Pettit 1988] Jackson, F and Pettit, P. 'Functionalism and Broad Content' *Mind* 97:1988 pp. 381–400.

[Jackson, Pargetter and Prior 1982] Jackson, F., Pargetter, R., and Prior, E.W., 'Functionalism and Type-Type Identity Theories' *Philosophical Studies* 42:1982.

[Kalke, 1969] Kalke, W., 'What is Wrong with Fodor and Putnam's Functionalism' *Nous* 3:1969 pp. 83–93.

[Kim 1973] Kim, J., 'Causation, Nomic Subsumption, and the Concept of an event' *The Journal of Philosophy*, 70:1973 pp. 217–66.

[Kim 1974] Kim, J., 'Epiphenomenalism and Supervenient Causation' *Midwest Studies in Philosophy*, 9:1984 pp. 257–70.

[Kim 1976] Kim, J., 'Events as Property Exemplifications' in [Brand and Walton 1976] pp. 217–236.

[Kim 1978] Kim, J., 'Supervenience and Nomological Incommensurables', *American Philosophical Quarterly* 15:1978 pp.149–156.

[Kim 1979] Kim, J., 'Causality, Identity and Supervenience in the Mind–Body Problem' *Midwest Studies in Philosophy*, 4:1979 pp. 31–49.

[Kim 1984] Kim, J., 'Concepts of Supervenience', *Philosophy and Phenomenological Research*, 1984 pp. 153–176.

[Kitcher 1980] Kitcher, Patricia, 'How to Reduce a Functional Psychology?', *Philosophy of Science*, 47:1980 pp. 134–140.

- [Kitcher and Salmon 1987] Kitcher, P., and Salmon, W., 'Van Fraassen on Explanation', *The Journal of Philosophy* 84:1987 pp. 315–330.
- [Kripke 1972] Kripke, S., 'Naming and Necessity', in [Davidson and Harman 1972] pp. 253–355.
- [Kripke 1980] Kripke, S., *Naming and Necessity*, (a revised and expanded version of [Kripke 1972]), B.H. Blackwell, Oxford 1980.
- [Kroon 1987] Kroon, F.W., 'Causal Descriptivism' *Australasian Journal of Philosophy*, 65:1987 pp. 1–17
- [Lewis 1983] Lewis, D.K., 'New Work for a Theory of Universals' *Australasian Journal of Philosophy*, 61:1983 pp. 343–377.
- [Lewis 1983a] Lewis, D.K., *Philosophical Papers, Vol 1*, Oxford University Press, Oxford 1983.
- [Lewis 1986] Lewis, D.K., *Philosophical Papers, Vol 2*, Oxford University Press, Oxford 1986.
- [Lewis 1986a] Lewis, D.K., 'Events', in [Lewis 1986] pp. 241–269.
- [Lewis 1986b] Lewis, D.K., 'Causation', in [Lewis 1986] pp. 159–171.
- [Lewis 1986c] Lewis, D.K., 'Postscripts to Causation', in [Lewis 1986] pp. 172–213.
- [Lewis 1986d] Lewis D.K., 'Causal Explanation' in [Lewis 1986] pp. 214–270.
- [Lewis 1986e] Lewis D.K., *On the Plurality of Worlds*, Backwell, Oxford 1986.
- [Lycan 1974] Lycan, W.G., 'Mental States and Putnam's Functionalist Hypothesis' *Australasian Journal of Philosophy* 52:1974.

- [Lycan 1987] Lycan, W.G., *Consciousness*, MIT Press, Cambridge Massachusetts 1987.
- [Mackie 1965] Mackie, J.L., 'Causes and Conditions', *American Philosophical Quarterly*, 2:1965 pp. 245–64.
- [Mackie 1974] Mackie, J. L., *The Cement of the Universe*, The Clarendon Press, Oxford 1974.
- [McCulloch 1986] McCulloch, G., 'Scientism, Mind and Meaning', in [Pettit and McDowell 1986] pp. 60–94.
- [Mellor 1976] Mellor, D.H., 'Probable Explanation', *Australasian Journal of Philosophy* 54: 1976 pp. 231–241.
- [Mellor 1977] Mellor, D.H., 'Natural Kinds' *British Journal of Philosophy of Science*, 28:1977 pp. 299–312.
- [Mellor 1981] Mellor, D.H., *Real Time*, Cambridge University Press, Cambridge 1981.
- [Menzies 1988] Menzies, P., 'Against Causal Reductionism', forthcoming in *Mind* October 1988.
- [Millikan 1984] Millikan, R., *Language, Thought and other Biological Categories*, MIT Press, Cambridge Massachusetts 1984.
- [Nagel 1961] Nagel, E., *The Structure of Science: Problems in the logic of Scientific Explanation*, Routledge and Kegan Paul, London 1961.
- [Nagel 1974] Nagel, T., 'What Is it Like to be a Bat' *Philosophical Review* 1974 pp. 435–450. Reprinted in [Nagel 1979].

- [Nagel 1979] Nagel, T., *Mortal Questions*, Cambridge University Press, Cambridge 1979.
- [Pearce and Maynard 1973] Pearce, G., and Maynard, P., *Conceptual Change*, Reidel, Dordrecht 1973.
- [Pelletier 1982] Pelletier, J., 'Completely Non-Clausal, Completely Heuristically Driven Automatic Theorem Proving' *Technical Report 82-7*, Department of Computing Science, University of Alberta 1982.
- [Perry 1979] Perry, J., 'The Problem of the Essential Indexical' *Noûs* 1979 pp. 3-21.
- [Pettit 1986] Pettit, P., 'Broad Minded Explanation and Psychology' in [Pettit and McDowell, 1986] pp. 17-58.
- [Pettit and McDowell 1986] Pettit, P., and McDowell, J., *Subject, Thought and Context*, Clarendon Press, Oxford 1986.
- [Platts 1983] Platts, M., 'Explanatory Kinds' *British Journal of Philosophy of Science* 34:1983 pp. 133-148.
- [Pollock forth.] Pollock, J. 'Interest Driven Reasoning', read to the Department of Philosophy, RSSS, ANU on 31/5/88.
- [Popper 1934] Popper, K., *Logik der Forschung*, Springer Verlag, Vienna 1934.
- [Popper 1959] Popper, K., *The Logic of Scientific Discovery*, Hutchinson, London 1959. Translation of [Popper 1934].
- [Price forth.] Price, H., 'Time and Explanation in Cosmology', preprint, Dept. of Philosophy, Research School of Social Sciences 1988.

- [Putnam 1973] Putnam, H., 'Explanation and Reference' in [Pearce and Maynard 1973] pp. 199–221 reprinted in [Putnam 1975] pp. 196–214.
- [Putnam 1975] Putnam, H., *Mind Language and Reality*, Cambridge University Press, Cambridge 1975.
- [Putnam 1975a] Putnam, H., 'Minds and Machines' in [Putnam 1975] pp. 362–385.
- [Putnam 1975b] Putnam, H., 'The Mental Life of Some Machines' in [Putnam 1975] pp. 408–428.
- [Putnam 1975c] Putnam, H., 'The meaning of 'meaning'' in [Putnam 1975] pp. 215–271.
- [Putnam 1978] Putnam, H., *Meaning and the Moral Sciences*, Routledge & Kegan Paul, London 1978.
- [Putnam 1981] Putnam, H., *Reason, Truth and History*, Cambridge University Press, Cambridge 1981.
- [Ramsey, Stich and Garon 1988 forth.] Ramsey, W., Stich, S., and Garon, J., 'Connectionism, Eliminativism, and the Future of Folk Psychology' preprint, University of California at San Diego 1988.
- [Reichenbach 1956] Reichenbach, H., *The Direction of Time*, University of California Press, Berkeley 1956.
- [Richardson 1979] Richardson, R.C., 'Functionalism and Reductionism' in *Philosophy of Science* 46:1979 pp. 533–558.
- [Richardson 1982] Richardson, R.C., 'How Not to Reduce a Functional Psychology' *Philosophy of Science* 49:1982 pp. 125–137.

- [Rorty 1965] Rorty, R., 'Mind-Body Identity, Privacy and Categories' *The Review of Metaphysics* 1965 reprinted in [Hampshire 1966] pp. 310–63.
- [Ryle 1949] Ryle, G., *The Concept of Mind*, Hutchinson, London 1949.
- [Salmon 1971] Salmon, W.C., *Statistical Explanation and Statistical Relevance*, University of Pittsburgh Press, Pittsburgh 1971.
- [Salmon 1978] Salmon, W.C., 'Why and "Why"' *Proceedings and Addresses of the American Philosophical Association* 51:1978 pp. 683–705.
- [Salmon 1984] Salmon, W.C., *Scientific Explanation and the Causal Structure of the World*, Princeton University Press, Princeton 1984.
- [Salmon 1985] Salmon, W.C., 'Conflicting Conceptions of Scientific Explanation', *The Journal of Philosophy*, 82: 1985 pp.651–653.
- [Schopf 1972] Schopf, T., ed., *Models in Paleobiology*, Freeman Cooper, San Francisco 1972.
- [Scriven 1975] Scriven, M., 'Causation as Explanation', *Nous*, vol. 9 pp. 3–16 1975.
- [Smolensky 1988] Smolensky, P. 'On the Proper Treatment of Connectionism', forthcoming in *The Behavioral and Brain Sciences* 11:1988 [references in the text are to a preprint from the University of Colorado 1986].
- [Stalnacker 1984] Stalnacker, R., *Inquiry*, MIT Press, Cambridge Massachusetts 1984.
- [Sterelny 1983] Sterelny, K., 'Natural Kind Terms', *Pacific Philosophical Quarterly* 64:1983 pp. 110-125

- [Sterelny and Godfrey-Smith forth.] Sterelny, K. and Godfrey-Smith, P. 'Semantic Psychology', forthcoming.
- [Stich 1983] Stich, S., *From Folk Psychology to Cognitive Science*, MIT Press, Cambridge Massachusetts 1983.
- [Strawson 1950] Strawson, P.F., 'On Referring', *Mind* 65:1950 pp. 320–44.
- [Strawson 1959] Strawson, P.F., *Individuals*, Methuen, London 1959.
- [Toulmin 1961] Toulmin, S., *Foresight and Understanding: an enquiry into the aims of science*,. Indiana University Press, Bloomington 1961.
- [van Fraassen 1980] van Fraassen, B., *The Scientific Image*, Clarendon Press, Oxford 1980.
- [van Fraassen 1981] van Fraassen, B., 'Essences and Laws of Nature' in [Healey 1981] pp.189–200.
- [van Fraassen 1985] van Fraassen, B., 'Salmon on Explanation' *The Journal of Philosophy*, 82:1985 pp. 639–650.
- [von Wright 1971] von Wright, G.H., *Explanation and Understanding*, Cornell University Press, Ithica N.Y. 1971.
- [Watson 1925] Watson, J.B., *Behaviourism*, Kegan Paul, London 1925.
- [Wimsatt 1976] Wimsatt, W., 'Reductionism, Levels of Organization, and the Mind–Body Problem' in [Globus, Maxwell and Savodnik 1976] pp. 205–293.
- [Zwart 1967] Zwart. P.J., *Causaliteit*, van Gorcum 1967.