This unpublished manuscript was completed c.1995 and is a distant descendant of my PhD thesis, *Meaning Change and Theory Change* (Columbia University, 1991).

CONCEPTUAL CHANGE IN SCIENCE

Muhammad Ali Khalidi

Table of Contents

Introduction

> I define a nose as follows--entreating only beforehand, and beseeching my readers, both male and female, of what age, complexion, and condition soever, for the love of God and their own souls, to guard against the temptations and suggestions of the devil, and suffer him by no art or wile to put any other ideas into their minds, than what I put into my definition.--For by the word Nose, throughout all this long chapter of noses, and in every other part of my work, where the word Nose occurs,--I declare, by that word I mean a Nose, and nothing more, or less.
> Laurence Sterne, Tristram Shandy

Often, in the course of a long discussion, the parties gradually realize that they do not mean the same thing by a certain key word. For instance, imagine that you are arguing with Socrates about whether aristocracy is the best form of government, and discover that you mean different things by the term 'aristocracy'. While Socrates uses it to mean the rule of the best, you use it to denote the rule of a select few noble families. When this terminological difference is uncovered and ironed out, you may well find that you are in accord on all the facts of the case. You might agree to talk about 'aristocracy$_1$' and 'aristocracy$_2$' and proceed on your merry way.

On closer inspection, two different possibilities can be distinguished. It may transpire that both parties to the debate have both concepts but that they use different terms for them. Alternatively, it may turn out that one of you lacks one of the two concepts (and perhaps that the other lacks the other). The first case is relatively straightforward; it is merely a disagreement over the sounds uttered or marks made. To be sure, it is perplexing when a meaning or concept associated with one term by one agent is associated with another by another agent and vice versa, but at least there are no conceptual gaps involved. Term-swapping among concepts often occurs; indeed, in the course of intellectual history, opposite concepts have been known to exchange terms.[1]

---

[1] That is what took place with the terms 'subjective' and 'objective' (and their cognates) in some European languages between the medieval and early modern periods. Raymond Williams explains: "The normal scholastic distinction between subjective and objective was: subjective--as things are in themselves (from the sense of subject as substance); objective--as things are presented to consciousness ('thrown before' the mind)." (1976,

But it is the second type of case that generally seems more problematic, in which one party to a discussion has a concept that the other party simply lacks; the first may even insist that the second's concept is inappropriate and should be discarded. In these cases, terms are not merely switched around, since at least one of the parties has a conceptual deficit--and one that may not be perceived as such. These are the paradigmatic instances of conceptual difference or change. They are particularly prevalent in science, for the introduction of a new scientific theory typically carries with it at least some new concepts. The history of science can therefore be expected to exhibit examples of conceptual difference between successive scientific theories. How can such differences be identified? And are they so widespread as to pose a threat to the comparison of pairs of theories? In tackling these questions, I will be assuming that the distinction between conceptual difference and the seemingly more enticing phenomenon of conceptual change is a practical one: 'conceptual difference' is more appropriate when talking about two agents at the same time, while 'conceptual change' is suitable mainly in discussing one or more agents at two different times. For these philosophical purposes, the synchronic and diachronic situations will be treated in fundamentally the same way.

Before philosophizing, we are inclined to say that there will be a conceptual difference between two agents just in case there is a difference in their definitions. That is even suggested by the way in which the Socratic example was explicated, since a brief definition was supplied for each concept to show how they differed. However, in the context of inquiry, erstwhile definitions seem to be no holier than other strongly held tenets and are equally subject to being revoked. The definition is sometimes given up when the associated concept seems to be retained, and that is precisely what makes the phenomenon of conceptual change so problematic. An appeal to definitions does not work for fixing the meaning of concepts, because what is considered definitional can change even where there is conceptual stability. Despite this difficulty, there seem to be such things as differences in meaning or concepts between two agents, which are not the same as differences in the theories or beliefs held by those agents. In this book, I will argue that it is

---

260-1) Since the modern use is diametrically opposed to the scholastic one, this case is particularly confusing. In commenting on this transformation, Williams goes on to raise an important point: "It is not that the terms were at all quickly clarified in this way; any such distinction is a much later summary." (1976, 261) It is worth pointing out that in this work, I will consider theories after they have gelled rather than when they are in flux; static theoretical snapshots will be compared rather than fluid dynamical systems.

important to distinguish them and will propose a particular way of doing so, with special reference to scientific discourse. In so doing, I will talk indifferently about the totality of an agent's beliefs or about an agent's theory of the world. I am assuming that one can think of any agent's set of beliefs as that agent's theory, and equally, any theory can be thought of as a set of beliefs held by a particular agent.

The problem of conceptual change has come to haunt contemporary philosophy of science and to threaten theorizing about scientific theories. Beginning in the early 1960s, the writings of Thomas Kuhn and Paul Feyerabend proposed a set of now famous theses inspired by a historical study of science. Among these was the claim that the meanings of scientific terms change in the course of the history of science in such a way as to make the comparison of successive theories problematic at the very least. This is the infamous "incommensurability thesis." It has since become evident that this thesis is not unique to a reading of science such as the one that Kuhn and Feyerabend gave. Even some of the interpretations that their work was meant to supplant, those given by logical empiricism (e.g. Rudolf Carnap's views), implicitly contained a "meaning-change view", though it was not thought to threaten the possibility of directly comparing theories.

Among those who have rejected the claim of the incommensurability of scientific theories, a number of responses can be discerned, none of which is wholly satisfactory. Some philosophers have suggested that the study of meaning has no place in the philosophy of science and that philosophers should not think of scientific theories as linguistic entities. In response to Kuhn and Feyerabend, Dudley Shapere wrote that "in view of the fact that that term ['meaning'] has proved such an obstruction to the fulfillment of this purpose, the wisest course seems to be to avoid it altogether as a fundamental tool for dealing with this sort of problem." (1966, 57) However, this defeatist conclusion often seems to be drawn in response to the perceived intractability of the problem and philosophers usually reach it after despairing of a satisfactory solution. Thus, Shapere adopts the position after admitting, somewhat paradoxically, that it might yet be "very valuable" for some purposes to formulate a precise "criterion" of meaning change. Soon afterwards, a number of philosophers attempted to do just that. But criteria that seemed to work for some case studies were open to counterexamples, so that attempt was short-lived.

Another view of the problem of incommensurability presumes that it is capable of being solved using the causal theory of reference, which was first expounded in the early 1970s by Keith Donnellan, Saul Kripke and Hilary Putnam. Although the problems that that theory was first designed to solve concern the reference of proper names, some philosophers have assumed that the causal theory is suitable for export to the realm of scientific discourse. These writers hold that general terms in science can be dealt with in

much the same way as proper names in natural language. I will argue that the detailed attempts to apply the causal theory to this problem are open to devastating objections.

A third, entirely different attitude towards the question of incommensurability and theory-comparison was put forward by Donald Davidson. Based on W.V. Quine's thought experiment of radical translation and his own truth-theoretic account of meaning, Davidson has argued that all languages are inter-translatable. He has used this argument to undermine Kuhn's claim that scientists who operate with different theories (or "paradigms") work in different "worlds". The argument is brief and proceeds at a purely general level; while it has convinced some, it has left many utterly dissatisfied. Some of those who have written explicitly about it are unable to see how it helps to defeat particular cases of alleged incommensurability among scientific theories.

A new attempt would seem to be in order, but the task might appear to be a thankless one, for a reason which has already been hinted at, namely the impossibility of coming up with indefeasible definitions for scientific terms.[2] In the absence of such definitions, a criterion for meaning change is not likely to be forthcoming. Moreover, the very yearning for unrevisable definitions (and hence for such a criterion) seems inimical to the spirit of corrigibilism inherent in scientific inquiry. This difficulty will be overcome by adopting a way of comparing scientific theories that is not committed to the existence of definitions for scientific terms. To this end, I will take the Quinean-Davidsonian program of translation or interpretation as my starting point, modifying it to suit these purposes. Crucially, I will augment it with a number of interpretive maxims and principles. The resulting methodological framework, which I call the "interpretive approach," will be used to show how successive theories can be compared to determine where they agree and disagree, thus restoring the possibility of rational choice in the sciences (without guaranteeing it, of course). Two notable attempts have been made that have tried to use Quine's and Davidson's insights about language and interpretation to tackle the problem of comparing scientific theories, David Papineau's Theory and Meaning and Peter Smith's Realism and the Progress of Science. Both works can be regarded as precursors to this one.

---

[2] The claim that there can be no fixed definitions in the context of inquiry is sometimes thought to be tantamount to a denial of the analytic-synthetic distinction. But since analyticity is understood differently by different writers, I will refrain from putting the claim in these terms. However, I will take it as fairly uncontroversial that the history of science has shown that putative definitions or 'meaning postulates' are revisable diachronically in the context of inquiry.

However, Papineau concludes that "there is no possibility, or need, of anything to play quite the traditional role of meanings in respect of scientific generalizations." (1979, 118) While professing realism, he rejects the idea that we can say that rival theories get particular things wrong and other things right, and ends up by saying instead that our theories as wholes are more or less successful attempts to picture reality. (1979, 128) As for Smith, he agrees with the present view, that terms of rival scientific terms get translated piecemeal into terms of our theories, and he looks at three of the case studies discussed in this work. However, Smith focuses on reference, which I will argue is the wrong place to look if one is interested in comparing scientific theories, and he employs a descriptional theory of reference, which I will criticize in later chapters. More importantly, he ends up conceding that translation may be indeterminate to some degree (even though reference is not).

It is not enough to argue that all languages are inter-translatable in order to refute the claim of incommensurability among scientific theories. Much more needs to be said about the translational or interpretive strategies that are to be employed when translating one scientific theory into another and the difficulties that one typically encounters. When the issue is recast in this mold, individual translational decisions become paramount. Such problems of detail will be addressed by working through a number of case studies, some of which have also been approached by other writers. By looking at particular cases, the all-important minutiae will be addressed and it will be shown how interpretive decisions are made. These include decisions concerning when to rule that a term from another theory fails to correspond to one of our own terms, when to introduce a neologism rather than use one of our already existing terms, whether to translate a term differently on different occurrences, and so on. To be sure, no entirely general answer to such questions is possible, but certain translational principles can be culled from the case studies that serve as constraints on the project of interpreting scientific theories. The solubility of these questions in specific cases can be demonstrated by tackling thorny problems head-on. The point is not to seek a criterion of meaning change but to formulate certain interpretive maxims or principles which will enable us to rule on meaning change in particular cases. One revisionist message that will emerge is that conceptual change is much rarer than theoretical change. Many changes that philosophers, historians, and others have traditionally characterized as conceptual changes turn out to be merely theoretical. There is a tendency to view any considerable theoretical difference among theories as a conceptual one. But that blurs the distinction between mere changes in belief and changes that result in a modification of the conceptual repertoire, ones that involve conceptual innovations or extinctions. So, not only are there no conceptual ruptures that render different theories incommensurable, conceptual differences between theories are rarer

than is usually supposed.

It should be clear from the discussion thus far that this work considers scientific theories to be expressible in natural language and advocates studying them in this form. That may sound like a return to the heyday of logical empiricism and to a discredited obsession with language in all areas of philosophical inquiry. However, it will soon become evident that the views of both language and science being put forward here are quite different from those once advocated by the logical empiricists. The stand taken in this work is also in direct opposition to a kind of glossophobia that has spread among philosophers of science in the past quarter century, which has driven many of them to hold that questions of language have no place in the foundational study of science. In recent years, this fear of language has been accompanied by a number of treatments of scientific theories that consider them to be something other than linguistic structures, but I would argue that a linguistic treatment is still indispensable. That is not to say that these other approaches should not be pursued and further developed, but they should proceed alongside rather than displace a linguistic account.

The non-linguistic views of scientific theories can be classified into three main approaches or schools of thought. The first is the "structuralist approach", represented by writers such as Patrick Suppes, Joseph Sneed, and Wolfgang Stegmüller, who consider scientific theories to be expressible in mathematical terms and adopt a model-theoretic approach to the study of scientific theories. The second, the "semantic view" of scientific theories is also model-theoretic, but the approach is more meta-mathematical. This method is advocated mainly by Bas van Fraassen and Frederick Suppe (based on the work of Evert Beth). The third, more recent trend, is to study scientific theories as they might be represented in the mind or brain, chiefly by considering computational models of these theories. Some advocates of what might be called the "cognitivist" or "computationalist" account are Ronald Giere, Paul Churchland, Paul Thagard, Nancy Nersessian, and others.

There are a number of reasons for continuing to propound a linguistic view of scientific theories in the face of these prominent alternatives. First, that is how scientific theories are regularly expressed, manipulated, discussed, defended, attacked, and so on. Since so much scientific traffic involves linguistic vehicles, that provides one reason for continuing to pursue the linguistic conception of scientific theories. This is not a view brought a priori by philosophers to scientific theories; it is a view very much inspired by scientific practice itself. Not only does language figure noticeably in the daily work of science, the place of language in science is not about to be taken over by any other representational medium. Specifically, neither the model-theoretic nor the computational modes of representation are likely to depose the linguistic, at least not in the foreseeable

future. Some of the advocates of the structuralist approach have noticed this point. For example, Richard Grandy ends an essay on the structuralist view by saying that while some theories in mathematical physics lend themselves to a structuralist treatment, it remains to be determined whether the approach fits geology, biology, or psychology "without serious distortion." (1992, 230) There is an asymmetry here, for while the linguistic conception can accommodate mathematical apparatus by embedding it within a larger linguistic framework (in what van Fraassen has called "mathematical English"), a mathematical or meta-mathematical approach cannot do the opposite. To demonstrate this point, I will show (in Chapter 4) how some of the central concepts of two mathematical scientific theories, relativistic and classical mechanics, can be matched up using the linguistic approach.

The second reason for pursuing a language-based conception of scientific theories has to do with the continuity of the scientific enterprise with the everyday, and the continuity amongst different scientific disciplines. Grandy's observation raises the issue of the relation of physics to the other natural sciences and the behavioral sciences. One need not be committed to a strong version of the unity of science thesis to pursue an approach that does not treat theories in mathematical physics altogether differently from theories in the other sciences. Although mathematical physics may not have the same methods and standards as the other sciences, it would be surprising if there were no common medium within which to represent theories in mathematical physics and those in the other sciences. This particular consideration does not favor the linguistic view over the computational view, for the latter also claims to be capable of representing theories in the other sciences as well as everyday systems of beliefs. Thagard, for example, has drawn his case studies from a number of sciences, including biology and psychology. But I will argue in Chapter 6 that at least some of the cognitivist literature still relies on a linguistic system of representing theories. Although such devices as "prototypes", "mental models", and "neural networks" have been used in a limited way, there is as yet no full-blown alternative to a sententialist medium in the cognitivist literature on science.[3] Moreover, psychologists themselves are increasingly attracted to a theory-based account of concepts influenced in no small measure by sententialist philosophers of science such as Quine (as will be seen in

[3] Despite the claims of Churchland (1992), the connectionist paradigm, which uses neural nets to model cognitive functions, is as yet unable to model higher cognitive processes such as those involved in devising, applying, and testing scientific theories. For further discussion of this point, see Chapter 6.

sections 6.6. and 6.7.). It is ironic that philosophers of science are looking to cognitive science for accounts of concepts just as cognitive psychologists are poaching on philosophical territory.

The third reason for reviving the linguistic view is less direct. It concerns the fact that the view seems to have been abandoned on insufficient grounds. Without proffering a sociological study of some recent trends in the philosophy of science, it can be argued that the reasons for the decline in the fortunes of the linguistic approach have been circumstantial and not fully reasoned. As I suggested above, there was increased concern about the problem of conceptual change among philosophers of science beginning in the early 1960s, and a despair over the prospects for solving it. At the same time, the logical empiricist treatments of science were coming under general attack, and so the linguistic view of scientific theories was thrown overboard together with other positivist baggage.[4] The linguistic view of scientific theories went out with the theory-observation distinction, the analytic-synthetic distinction, and other theses considered central to logical empiricism. Indeed, a reason commonly given for abandoning the linguistic view is the problem of conceptual change. Many philosophers have come to the conclusion that if scientific theories are expressed in language, the prospects for finding a way to compare theories are slim. It has become fashionable to say that the whole enterprise was misguided and philosophers have begun experimenting with other ways of representing theories. Obviously, when scientific theories are not represented by sentential structures, one need no longer bother about the meanings of scientific terms. But if the main reason for shelving the linguistic approach was the supposed obstacle posed by the problem of meaning change, a proposal for solving that problem should constitute something of a reason for reconsidering the linguistic approach.

This book proposes a way of identifying conceptual change in science by way of giving a semantics for scientific theories; but there is also another way to take it. In recent years, there has been considerable discussion of the nature and status of "folk psychology", our everyday practice of ascribing mental attitudes to human agents (also known as "propositional-attitude psychology" or "common-sense belief-desire psychology"). This work can be seen as an attempt to theorize about one particular aspect of our common-sense psychologizing, specifically the way we interpret theoretical and explanatory systems of belief in the natural and social sciences. Scientific theories hold an interest in their own

[4] In one of the seminal works of the structuralist approach, Stegmüller (1979) considers it an important doctrine of logical empiricism, calling it the "statement view".

right and the possibility of comparing them over time is relevant to the rationality of science and the question of conceptual relativism. But scientific theories are also useful as examples of belief systems held by psychological agents who are subject to interpretation by other agents. Moreover, they are relatively unproblematic as belief systems go, for the following four reasons (some of which will be further justified in later chapters). First, they are usually exhaustively articulated and can be gleaned explicitly from utterances (whether spoken or written) rather than implicitly from actions and other indirect means. Second, the psychological agents who hold them are the experts on those theories and are generally well-versed in the theories they profess to hold, unlike many of us who use concepts without a complete knowledge of the theories from which they derive. Third, such theories are explanatory and usually contain fewer unnecessary concepts that do not earn their keep, as well as fewer cases of obvious inconsistency. Fourth, although they often contain non-literal and rhetorical elements, they tend to be more literal than most portions of ordinary discourse. These factors make it easier to analyze such theories using philosophical tools. They also make it easier to emerge with certain general pronouncements about the canons of interpretation and belief-ascription than when one is dealing more generally with ordinary psychological agents.

These special characteristics of scientific discourse also mean that any lessons that one might draw will not necessarily be applicable to folk psychology in general. But then, folk psychology and the theory of meaning should not be thought of as a monolithic theory deployed for a single unified purpose, but one that may be as diverse as the folk themselves. Noam Chomsky has written that "an interest in intelligibility in scientific discourse across time is a fair enough concern, still it is hard to see why it is a basis for a general theory of meaning; it is, after all, only one concern among many, and not a central one for the study of human psychology." (1989, 16) While I would not be so rash as to claim that this particular application of folk psychology should serve as a basis for the theory of meaning, it does seem that the considerations just cited make the interpretation of scientific discourse a relatively straightforward application of our folk theory. For this reason, I suspect that it is less adulterated than other uses of folk psychology and that a semantic theory for scientific terms and a method for the interpretation of scientific theories might play the role that certain simplified and idealized systems generally play in empirical inquiries, but I will not try to justify this suspicion further.

* * * *

The first two chapters of this book take a critical look at some of the previous work

done on the issue of conceptual change. The following four chapters furnish the bulk of my own positive account, which I call the "interpretive approach". The last chapter poses some questions about realism, which are raised by the interpretive approach to the comparison of scientific theories.

The problem of the meaning of theoretical terms in science is traced back to the work of N.R. Campbell in **Chapter 1**, since he was one of the first to discuss it in any detail. The account he gave was influential for the logical empiricists and provides the background for their successive accounts of theoretical terms. Without pursuing these accounts in detail, I take a closer look at Carnap's later work on this issue. His later view entails that the meaning of all theoretical terms change with every change in the theoretical tenets of a scientific theory. That is because he subscribes to a kind of extreme holistic theory of meaning for theoretical terms, though not for observational ones. In Feyerabend's work, there is a similar commitment to extreme holism. When accompanied by a denial of the distinction between theoretical and observational terms, it leads to the conclusion that all terms change in meaning with every change in theory. A kindred view is sometimes suggested in Kuhn's early work, but his later statements of incommensurability focus on local translational difficulties whereby specific terms or clusters of terms resist translation from one theory to another. I also look at Israel Scheffler's response to Kuhn and Feyerabend, which shares some of their assumptions, but appeals to the notion of reference to avoid the conclusion of incommensurability. However, Scheffler does not present a fully developed account of reference, a task that was left for the causal theorists of reference, to be discussed in the following chapter.

Chapter 2 introduces another influential account of the meanings of scientific terms, which was developed in the wake of the incommensurability thesis: the causal theory of reference advocated by Putnam, Donnellan, Kripke, and others. The perceived successes of the causal theory may be partly responsible for the impression that the problem of conceptual change in science has already been solved. Although the causal theory was not developed primarily as an account of the reference of scientific terms, some philosophers went on to use it for this purpose. Many others simply assumed that it could be relied upon to give such an account and proceeded as though it already had. But a closer look at the attempts in this direction and at the details of the causal theory itself reveals that such confidence is misplaced. The causal theory of reference is not equipped to give a plausible picture of reference change and, more importantly, is unable to allow proponents of scientific theories to compare those theories and defeat the claim of incommensurability. It also presupposes an incorrect account of scientific taxonomy and makes inordinate demands on the mechanism of ostension. These three criticisms render it inapplicable to

scientific terms, even when supplemented by a second component of meaning.

In **Chapter 3**, I outline the philosophical framework for the alternative account of conceptual change and continuity in science. I begin with a presentation of Davidson's argument against the possibility of wholly or partially incommensurable conceptual schemes, based on his interpretive account of meaning. The main objections confronting the interpretive approach involve the indeterminacy of translation and holism about meaning. The first charge would have it that the interpretive approach is only able to defeat incommensurability at the expense of succumbing to its irrationalist ally, indeterminacy. However, I argue that indeterminacy does not have debilitating implications for the comparison of scientific theories, since there are constraints on the interpretation of one theory in terms of another which will yield an optimal translation in each case. The second objection, which has been made recently by Fodor and Lepore, claims that holism leads inevitably to the anarchic scenario that every change in scientific theory leads to a change in meaning of all the terms involved in the theory. To counter this objection, I distinguish the "moderate holism" involved in the interpretive approach from the "extreme holism" that was associated with some of the accounts explored in Chapter 1. In this chapter, I also address other concerns associated with translatability (especially ones raised by Kuhn) and reject the possibility of local incommensurability among scientific theories.

A number of case studies are analyzed in **Chapter 4**, which are designed to show how a translation is arrived at in practice and how conceptual change and continuity can be identified. The first case study compares classical mechanics to relativistic mechanics. A stumbling block in this case appears to be Newton's term 'mass', but there are compelling reasons to match this term with Einstein's term 'rest mass'. When this is done, one of the main differences between the two theories turns out to be over Euclidean space versus Minkowskian space-time. In the comparison of Priestley's phlogiston theory and the post-phlogiston theory of Lavoisier and others, two of the main problems are the translation of the terms 'phlogiston' and 'dephlogisticated air'. I show how to decide that the former fails to correspond to any of our terms (a genuine case of conceptual difference) and that the latter should be translated as 'oxygen'. Another case study involves comparing Dalton's atomic theory with that of Avogadro and others. Here, one problem is to interpret Dalton's term 'elementary atom', which seems to presuppose that all elements are composed of atoms (rather than molecules) in their natural state. The evidence militates for translating 'elementary atom' as 'atom', thereby enabling us to say that Dalton shared our concept, but was wrong in this important belief. Aristotle's views on motion are also interpreted. One of his concepts seems indeterminate between instantaneous speed and average speed; I

argue that it should be interpreted as the latter. A final case study examines a few examples drawn from the history of political and social theory.

Some general translational principles are outlined in **Chapter 5** based on the specific examples encountered in the previous chapter. These act as constraints on translation or interpretation and enable us to produce an optimal mapping between two scientific theories. The Principle of Conceptual Charity calls on the interpreter to maximize agreement in concepts (rather than beliefs). The Principle of Uniformity recommends that the same term from the source theory should be substituted for a given term from the target theory on each occurrence, barring equivocality. The Principle of Simplicity distinguishes between composite expressions that stand for simple concepts with their own entry in the lexicon and those that do not; it provides a way of identifying such expressions and calls for translating the former by simple expressions. The Principle of Warranty enjoins us not to ascribe a concept without sufficient warrant and provides some guidelines as to the type of warrant required. The Principle of Undefinability points out that there are no unrevisable definitions for scientific terms and shows how concepts may be shared despite the fact that supposed definitions may differ. The Principle of Neologization specifies the conditions under which one coins new terms to translate a scientific theory; some of the concerns surrounding this practice are addressed. Finally, the Principle of Literality instructs the interpreter to concentrate exclusively on literal meaning in interpreting scientific theories; I argue that some alleged translational difficulties arise because of a failure to do so. Each of these principles is given a justification which is in keeping with the framework of the interpretive approach.

The purpose of **Chapter 6** is to explicate what concepts are and to justify the claim that one compares scientific theories by way of their concepts. I begin by drawing a distinction between an innocuous notion of extension and a metaphysical realist notion of reference. After arguing that there is room for a straightforward notion of extension within the interpretive approach and explicating such a notion, I criticize the rival, metaphysical realist notion of reference. This account of reference (of which the causal theory encountered in Chapter 2 is one variety), is inappropriate as a means of comparing scientific theories. I also include a brief discussion in this chapter of some of the examples that have been used to motivate a metaphysical realist theory of reference, namely the Twin Earth example and related cases. In some of these cases, inclusive concepts are shared among agents; in others, we simply consider their concepts to be parasitic on those of the experts in their communities. This sociolinguistic phenomenon ("linguistic division of labor") is not of primary concern here, since the concepts of the experts themselves are under investigation in this work, but this discussion allows us to say something about the

relation between expert and lay concepts. Then, in light of what I call the "failure of transitivity" in the ascription of concepts, I respond to a worry that the interpretive approach is anti-realist about concepts and go on to discuss the nature of concepts. Finally, I relate my account of concepts to some recent work by cognitive psychologists and find considerable agreement; I also point to compatibility with recent work in artificial intelligence. Along the way, I criticize some of the cognitivist work in the philosophy of science for simply presupposing a way of comparing scientific theories, and I criticize some of the work in developmental psychology for being too quick to come to a conclusion of incommensurability among the belief systems of children and adults.

The whole discussion of conceptual change in science and the meaning of scientific terms raises larger issues about realism. In **Chapter 7**, I consider the charge that the interpretive approach is not realist about scientific theories. One ground for the charge of anti-realism might lie in the breakdown of transitivity for the interpretation of concepts, which was discussed in Chapter 6. However, I show that no such implications follow from this result, since the explanatory efficacy of concepts is ultimately what tethers them and prevents a kind of conceptual drift. Terms are not tied individually to entities in the world by a kind of metaphysical anchor, but that does not mean that they float free from the world. Another ground for the charge of anti-realism is based on my claim that scientific taxonomies are crosscutting. It may be thought that crosscutting theories are incommensurable, but I argue that such theories are capable of coexisting in our total theory of the world. Far from being incommensurable rivals, they coexist because they pertain to different interests. I end by referring to the debate about scientific realism and argue that some realists have been too complacent in accepting incommensurability as a fact of life and not recognizing it for the anti-realist threat that it is. Rather, one needs to defeat incommensurability, by offering an account of the meaning of scientific terms and of the phenomenon of conceptual change in science.

Chapter 1: **Meaning**

> It is easy to assent to the statement "in the beginning was the Word". This view underlies the philosophies of Plato and Carnap and of most of the intermediate metaphysicians.
>
> Bertrand Russell, An Inquiry into Meaning and Truth

1.1. Language and Science

The language of science has been a subject of philosophical speculation at least since Aristotle, and it seems to have come under scrutiny ever since philosophers began to theorize systematically about the world. But the topic of the meaning of scientific terms seems to have aroused special interest with the outbreak of what came to be known as the "Scientific Revolution" in early modern Europe. Francis Bacon included the "Idol of the Marketplace" among the Four Idols that he believed skewed and distorted the practice of science. According to Bacon, worship of this Idol amounted to undue veneration of existing language, terminology, and jargon, which in Bacon's time was primarily Aristotelian and Scholastic in origin. Bacon regarded Scholastic terminology as unsuited to the latest scientific research and he advised natural philosophers to reform their language in accordance with recent discoveries. He also urged them to begin their inquiries with definitions, though he realized--to his credit--that this would not really solve the problem: "Yet even definitions cannot cure this evil in dealing with natural and material things, since the definitions themselves consist of words, and those words beget others." (1620/1985, 56-7) Bacon's admirer Kant was also sceptical about the possibility of coming up with definitions for empirical terms, though his rationale is rather different. It is not that definitions involve other terms which need to be defined in their own right. Rather, the definitions are themselves subject to revision in the course of inquiry. Kant notes that empirical concepts such as gold and water cannot really be defined, since "new observations remove some properties and add others; and thus the limits of the concept are never assured." (1787/1933, A728/B756)

The problem being discussed here is related to these traditional ones. It is one that has been intensively scrutinized in the twentieth century, with the advent of the "linguistic

turn" among Anglophone philosophers. The issue consists in determining whether two scientific terms mean the same thing, as used by different scientists operating with different scientific theories. Equivalently, it consists in determining when scientific concepts have changed and when they have remained fixed in the face of significant theoretical change. At first glance, the solution might seem straightforward. A term means whatever its users say it means and when in doubt, they can supply a definition. But definitions do not help, because different definitions are sometimes given of terms that have the same meaning (which is a version of Kant's point), and because definitions merely defer the burden to other terms (which is the observation that Bacon so astutely made). Moreover, we cannot stop the process of deference simply by singling out whatever is under the magnifying glass, in the cloud chamber, or on the petri dish, especially when theoretical entities, such as acids, genes, magnetic fields, or neurological disorders, are involved.[1] Hence the "problem of theoretical terms", as it is sometimes dubbed. Twentieth-century philosophers of science have been concerned with finding a way of specifying the meanings of scientific terms, so that they can determine when scientific concepts have changed and when they have remained constant. The story of their successive attempts recapitulates some of the developments that occurred in the more general project of devising a theory of meaning for all terms, whether scientific or non-scientific. But it is a story that bears telling anew. In this chapter, the problem of the meaning of scientific terms will be tracked through some of its most important incarnations in twentieth-century philosophy of science. The main way-stations on this road are the positions of Campbell, Carnap, Feyerabend, Kuhn, and Scheffler.

1.2. Campbell: Subjective Meaning

[1] The problems with grounding the meaning of scientific terms in ostension will be mentioned in the course of this chapter, and will be criticized more directly in section 2.5.

One of the earliest analytic discussions of the problem of conceptual change in science occurs in Campbell's The Foundations of Science (1919/1957).[2] This work anticipates several moves made by later contributors to the debate and contains an interesting attempt to come to terms with the problem. In a chapter on the nature of scientific laws, Campbell argues that the use of certain scientific terms presupposes that certain scientific laws are true, and that any statement containing those terms is meaningless if those laws are not true. (1919/1957, 42) He goes on to assert:

> These words include most of the technical terms of science, but the laws on which they depend for their meaning are often not explicitly recognised as such. It will be convenient to have a name for such words and they will in the future be called concepts. A concept is a word denoting an idea which depends for its meaning or significance on the truth of some law. (1919/1957, 45)

Setting aside the throwback to classical empiricism in this passage ("a word denoting an idea"), Campbell seems to be putting forward a definitional theory for scientific terms. According to such a theory, the meaning of each scientific term would be tied to a single scientific law. However, later remarks suggest that this might not be quite right.

Campbell goes on to discuss the meaning of the term 'silver', considered as a technical scientific term. He contemplates the view that the term is meaningless unless "the proposition asserting the uniform association of the properties of silver" is true. Therefore, the meaning of 'silver' depends on the truth of a proposition that specifies all of its properties, and not just on a single law. But, according to him, that would imply that every time the word 'silver' is used, one has already assumed the truth of, say, the proposition that silver melts at 960° C, which is one of the properties of silver included in the more comprehensive proposition. However, he rejects this claim as obviously false, since it would lead to the conclusion that the statement that silver melts at 960° C is "a mere truism, like the statement that a black cat is black..." (1919/1957, 45) Far from being

---

[2] First published in 1919 as Physics: The Elements and reissued in 1957 as The Foundations of Science.

a tautology, such a statement is an empirical truth in Campbell's opinion, and this shows that something is wrong with this view.

To avert the conclusion that scientific statements are mere tautologies, Campbell considers changing things such that the meaning of 'silver' does not depend on the property that silver has a melting point of 960° C. But then, he says, the same difficulty can be raised with another property, say the density. When we say that silver has a density of 10.5, this will now be a truism, just as the statement concerning the melting point was previously. Campbell continues by saying that if we omit the density from the definition, we can be driven to omit all properties from the definition of silver one by one, which will force us to admit that by 'silver' we mean nothing at all. If, instead, we reinstate the melting point when we omit the density from the definition, it can be said that a different definition is employed on each occasion, "again a conclusion we cannot admit." (1919/1957, 46)[3] Campbell has constructed a dilemma. The first horn is that 'silver' has a different definition on each occasion of use, depending on the property that is being asserted of it. If a statement is made concerning the density, the density is omitted from the definition, and if it is made concerning the melting point, the melting point is omitted in turn. The second horn is that 'silver' becomes a meaningless term because its definition is emptied of all the distinctive properties of silver.

In order to find a way out, Campbell considers dividing the properties of silver into two groups, those that give the meaning of the term 'silver' and those that do not. He asks: "Are there properties of silver which simply define what we mean by silver and such that, if they were altered, the substance would not be silver; and are there on the other hand non-defining properties, such that they might be changed without affecting the fact that the substance in question is silver?" (1919/1957, 47) He answers this question in the negative, based on the fact that there is no principled distinction between the defining properties

---

[3] Compare Kant: "Thus in the concept of gold one man may think, in addition to its weight, colour, malleability, also its property of resisting rust, while another will perhaps know nothing of this quality." (1787/1933, A728/B756) By contrast with Campbell, this problem leads Kant to give up on the attempt to define an empirical concept, as mentioned in the previous section.

and the non-defining properties.[4] After failing to block the dilemma in this way, by preventing it even from arising, he eventually embraces the first horn: that on different occasions, we may mean different things by the term 'silver'.

It is not that Campbell is unaware of the difficulties with his view. Indeed, he seems to realize fully the consequences of his position. If one accepts the idea that the term 'silver' is polysemous, many arguments that involve the term would be rendered invalid on grounds of equivocation. In fact, Campbell thinks that we are led into error "if from two propositions involving the word silver we deduced a third which would follow only if it were certain that the word was used in exactly the same sense in each of them." (1919/1957, 51) However, his response to this apparently serious problem is unconvincing. He distinguishes between the use of words in science and their use in logic and declares logical standards to be out of place in science: "I believe that all important scientific thought is illogical, and that we shall be led into nothing but error if we try to force scientific reasoning into the forms prescribed by logical canons." (1919/1957, 52)

It is not clear what Campbell is denying at this point, whether he doubts that sameness of meaning is required for deductive arguments to go through, or whether he is implying that deductive arguments play no role in science. If it is the former, then some other criterion must be offered. But Campbell does not advance one; indeed it is not even clear that he feels that one is needed. As for the latter, it flies in the face of scientific practice and theorizing about science. While there is an illustrious tradition that denies that induction plays a role in science, it is difficult to find any theorists who deny the role of

---

[4] One particularly interesting aspect of Campbell's treatment of this issue is the way in which it anticipates later discussions. He notes that a similar objection might be raised for non-scientific statements such as 'William Smith lives next door to me'. But he argues that one can distinguish between defining and non-defining properties in this case, the defining property being that William Smith is the son of John Smith and Eliza (formerly Jones). (1957, 46-7) Despite this proto-Kripkean attitude towards proper names, Campbell explicitly resists giving a similar treatment for scientific terms.

deduction.[5] Consider the following simplified example. Suppose that a materials scientist is looking for a solution to an engineering problem that requires manufacturing a piece of equipment that would be exposed to temperatures of 600°C and has a density of less than 12. In determining whether a solution exists to this problem, a scientist might argue as follows: "This task requires a metal with a melting point over 600°C and a density of less than 12. Silver melts at 960°C. Silver has a density of 10.5. Therefore, silver is a suitable metal for this task." According to Campbell's account, 'silver' means something different in the first and second occurrences. In the first occurrence, the melting point is not part of the meaning of the term but the density is; in the second occurrence, the reverse is true. Hence, the argument is simply invalid. However, this is a simplified version of an argument that a scientist or engineer might make in seeking a heat-resistant metal to serve a particular engineering task. It is the kind of argument that an account of the meaning of scientific terms cannot afford to dismiss as Campbell appears to do.

Campbell's unsatisfactory position may stem partly from a problematic conception of meaning. He characterizes meaning as follows: "The meaning of a proposition... is simply the set of thoughts which it calls to mind; the meaning of two propositions is different if they call up different thoughts." (1919/1957, 52) To be sure, he is speaking of propositions here, but he would say something similar when talking about the meanings of terms, since he writes that "our words are perfectly effective in calling up the thoughts we desire..." (1919/1957, 53) Although his conception of meaning has affinities to Fregean "sense" (Sinn), Campbell departs from Frege by adopting a thoroughly subjectivist notion of meaning in talking about the thoughts or ideas that are conjured up by words in the minds

---

[5] Francis Bacon might be thought to be an exception. But even his fiercest attacks on Aristotelian logic do not imply that logic has no place in science, but only that it is powerless to supply the inquirer with first principles: "For logical invention does not discover principles and chief axioms, of which arts are composed, but only such things as appear to be consistent with them. For if you grow more curious and importunate and busy, and question her of probations and invention of principles or primary axioms, her answer is well known; she refers you to the faith you are bound to give to the principles of each separate art." (1620/1985, 79)

of language users. This may be what leads him to make light of the conclusion that the term 'silver' has a different meaning on different occasions of use. On a subjectivist conception of meaning such as the one he advocates, comprehension and communication are precarious at best, since the ideas that a word conjures up in the mind of two different language users will not generally be the same. As he puts it: "Meaning... is something individual and personal; it is something which depends on the qualities of my mind and is present in my mind whether or no it is present in the minds of others; a proposition may have meaning for me even if it has meaning for nobody else; and it is not certain ever that its meaning for me is the same as its meaning for anyone else." (1919/1957, 219) Since Campbell thinks that communication does not depend on the coincidence of subjective ideas anyhow, it is perhaps not surprising that he does not consider it a problem for scientific terms to have different meanings in different contexts.

This subjectivist conception of linguistic meaning is coupled with an extreme semantic holism, and the two conspire to issue in Campbell's overall position. Campbell thinks that the meaning or significance of a scientific statement depends on the entire theory in which it is embedded, in such a way that any change in the theory alters the meaning of the original statement.[6] He writes: "If we consider any law very carefully we shall find that there is somehow involved in it a reference to any other law, and that its significance would be changed to some small degree if any other law whatever ceased to be true or if any new law were discovered." (1919/1957, 50) It is safe to assume here that Campbell is using "significance" more or less interchangeably with "meaning", for he states directly before that "statements about silver, mercury, and lead are not independent statements; each depends for its meaning... on the truth of all the remainder." (1919/1957, 50) This extreme holism seems to affect, not just the highly theoretical statements of a theory, but the observational ones as well. For Campbell, laws (as opposed to theories) are couched in observational vocabulary. Thus, even statements describing the

---

[6] In Chapter 3, I will distinguish this extreme brand of holism from a more moderate variety that I will argue does not have this consequence. But throughout this chapter, when I speak of "holism", I will mean the extreme variety that does have this consequence.

macroproperties of elements like silver and lead are subject to this extreme variability in meaning as new statements are added to or subtracted from the theory. It is therefore not surprising that Campbell doubts that any two agents mean the same thing by their terms. The chances are that no two agents will hold exactly the same set of statements in their respective theories, so the meanings of their terms will generally be different. Not only does the definition of 'silver' change on each occasion of use, the meanings of the terms contained in each definition are not likely to be held in common by different scientists who hold different overall theories. Even a single proposition about silver as used by two different scientists is liable to have a different meaning because of this extreme holism.

A third, and perhaps the most influential, aspect of Campbell's view of scientific theories concerns his resolution of a scientific theory into two parts or two groups of propositions, the "hypothesis" and the "dictionary": "One group [the hypothesis] consists of statements about some collection of ideas which are characteristic of the theory; the other group [the dictionary] consists of statements of the relation between these ideas and some other ideas of a different nature." (1919/1957, 122) Campbell refrains from alluding specifically to observational and theoretical statements, preferring instead to make the distinction based on less controversial grounds. Thus, the hypothesis contains concepts that are "characteristic" of the theory, presumably those concepts that are introduced by the theory in question and are not given antecedently by another theory. These concepts are termed "hypothetical ideas" by Campbell, as distinct from the "concepts" proper, which are featured in the dictionary. Campbell acknowledges that some of the hypothetical ideas might be linked directly to concepts by means of the dictionary, but maintains that hypothetical ideas and concepts will nevertheless differ in meaning. He thinks that although a theory may be logically equivalent to a set of experimental statements, it <u>means</u> something quite different. A theory is valuable only if it evokes ideas that are not contained in the laws that it explains. (1919/1957, 132) Campbell insists that even when the theoretical concepts are linked directly to experiment by way of the dictionary, they differ from them in meaning. He is able to say this because he thinks that their meaning is given by their place in the theory as a whole.

To sum up, Campbell considers making the meaning of a scientific term such as 'silver' dependent on the totality of laws in which it features or on all the properties that

are shared by samples of silver. But that would make any statement about the properties of silver true by virtue of meaning alone. Since he regards this to be an intolerable conclusion, he proposes that for each term in an empirical statement, one leaves out of its definition on that occasion the property that is being asserted in that statement. A statement about the melting point of silver would leave that property out of the definition of 'silver', so as to avoid making it true by virtue of meaning alone. However, this implies that different occurrences of the term 'silver', even within a single theory (let alone across theories) have different meanings. This is a conclusion that Campbell professes himself willing to live with, and although it might seem surprising, it is perhaps less so given his extreme subjectivist theory of meaning. However, the fact that it would render all interesting deductive arguments invalid, even within a single theory, is sufficient reason to look for a different account. If the account cannot guarantee intra-theoretic stability of meaning, it is obviously powerless when it comes to inter-theoretic stability.

1.3. Carnap: Meaning Variance

Aspects of Campbell's account of the meaning of scientific terms can also be found in the work of some of the logical empiricists. In fact, two of the three main elements of Campbell's view as it was characterized in the previous section (the theory-observation distinction and extreme holism) seem to have been shared by Rudolf Carnap, at least in his later work. Carnap's treatment of the meaning of scientific terms will be examined by looking at two of his later papers, but it will also be useful to begin by alluding to the way in which his views evolved.

The first feature that Carnap's theory shares with Campbell's is the distinction between the theoretical and observational vocabulary of a scientific theory. As is well-known, in early articulations of Carnap's position, he held that each theoretical term was strictly defined by means of observational terms alone. Two aspects of this view were later abandoned: the requirement that every theoretical term should be so defined, and the requirement that these definitions should be strict. Instead, in the first paper to be discussed, he allowed that the definitional rules (or "correspondence rules", or "C-rules") could only be given for some theoretical terms and that the definitions that linked them to observational terms could only generally be partial. While he still believed that

correspondence rules could be used to link some theoretical terms to the observational vocabulary, he now thought that there were other theoretical terms that could only be linked by certain theoretical postulates to the first set of theoretical terms, instead of directly to observational ones. This was a concession made to the way in which theoretical terms are commonly introduced and understood in scientific practice. The relaxation of the second condition and the move from strict definitions to partial definitions was made in response to the well-known problem of dispositional terms. Such terms ('soluble', 'fragile') were found to defy a strict definitional treatment for reasons that need not be rehearsed here. Suffice it to say that their relation to observation was found to be considerably looser than Carnap first postulated and could not be given in the form of necessary and sufficient conditions. Subsequently, this treatment was found to be more plausible for theoretical terms in general.[7]

As an example of a correspondence rule that links a theoretical term to an observational term, Carnap gives the following for the theoretical term 'temperature' and the observational predicate 'warmer than': "If u is warmer than v, then the temperature of u' is higher than that of v'." Here, u and v are material bodies (observable at locations u and v) and u' and v' are the coordinate regions corresponding to u and v, respectively. Carnap points out that such examples show that the C-rules effect a connection only between certain sentences of a very special kind in LT (the theoretical vocabulary) and sentences in LO (the observational vocabulary). As a direct consequence of this, he acknowledges that the definition of meaningfulness must be relative to a theory T, because the same term may be meaningful with respect to one theory but meaningless with respect to another. (1956, 48) Inter-theoretic stability of meaning cannot generally be guaranteed if one cannot give a strict definition of each theoretical term by way of observational terms. Since Carnap abandoned the latter feature, he could not satisfy the former demand.

If the entire set of theoretical postulates (T) plays a role in giving a meaning-specification of the theoretical terms, then the possibility arises that a change in the

---

[7] For some reasons why this treatment was found to be superior for theoretical terms generally and not just for disposition terms, see Hempel (1963, 689).

theoretical postulates will ensue in a change of meaning of the terms. This leads directly to the second important feature of Carnap's later view, and one which he shares with Campbell: extreme holism. Carnap considers the following objection to the view[8]:

> Perhaps the objection might be raised that, if significance is dependent upon T, then any observation of a new fact may compel us to take as nonsignificant a term so far regarded as significant or vice versa. However, it should be noted first that the theory T which is here presupposed in the examination of the significance of a term, contains only the postulates, that is, the fundamental laws of science, and not other scientifically asserted sentences, e.g., those describing single facts. Therefore the class of the terms of LT admitted as significant is not changed whenever new facts are discovered. This class will generally be changed only when a radical revolution in the system of science is made, especially by the introduction of a new primitive theoretical term and the addition of postulates for that term. (1956, 50-51)

Carnap does not say how large a change a "radical revolution" might be, or how to distinguish such a revolution from a meaning-preserving change. He does indicate, however, that any change that involves the introduction of a new theoretical term is sufficient to constitute a meaning-altering change. It is not immediately obvious whether the introduction of new theoretical postulates for an existing theoretical term would also have the same effect. But upon reflection, it becomes clear that this would have to be the case because the meaning of a theoretical term is given for Carnap by its theoretical postulate. Changing the theoretical postulate in any way would therefore change the meaning of the term, so the same theoretical quantity would no longer be in play. Thus, any real theoretical change must be a meaning-altering change. As Jane English has argued, since every postulate of the theory is represented in the theoretical postulate together with the correspondence rules, any theoretical change will involve a difference in meaning

---

[8] In the following passage and elsewhere, Carnap uses the terms "significance" and "empirical significance" instead of "meaning" or "meaningfulness". He seems to use these expressions interchangeably however. Or, more correctly, he thinks of "significance" as the more precise counterpart, after philosophical explication, of the pre-theoretic term "meaning".

conventions, according to Carnap's later view. Hence, she concludes that on this view, every theoretical change leads to a change in meaning of at least some theoretical terms. (1978, 67)

This leads naturally to a question about the scope of the meaning change. Would the addition of one new theoretical tenet to a theory alter the meanings of all the theoretical terms of that theory? Consider what happens when a theoretical tenet is revised or a new theoretical tenet introduced. Since the meaning of theoretical terms is given by the theoretical postulates of the theory and not just by the correspondence rules as before, the meanings of all the theoretical terms featured in the revised theoretical tenets will change with the revision. The change of meaning of these terms in turn changes the meaning of all other theoretical terms linked to them by way of other theoretical tenets. In light of this, it is not surprising that Carnap does not talk about changing the meaning or significance of a certain term or a certain set of terms with an alteration in the theory. Instead, in the passage quoted above, he speaks of taking "as nonsignificant a term so far regarded as significant or vice versa." That is, a term that was once significant can become nonsignifcant or a new term that had no significance can become significant. It is not that a term changes in significance, in the sense of having one significance and acquiring another. That is because the old terms that are affected by the change in theory will not generally correspond to any of the terms in the new theory. From the point of view of the old theory, significant terms have ceased to be so, while from the perspective of the new theory, altogether new significant terms have been introduced. A change in significance might suggest that a term from the new theory might have the same meaning or significance as some other term in the old theory. But since significance depends on the whole theory, or at least on all the postulates that contain the term in question, this cannot come about on Carnap's mature view.

In a later paper, a reply to Carl Hempel, Carnap goes even further. Rather than holding that significance or meaning is given by TC, the set of purely theoretical postulates plus the correspondence rules, he holds that purely observational postulates (e.g. empirical generalizations) can also convey significance to the theoretical terms. He now decomposes a theory TC into two parts. First, he constructs the theory's Ramsey sentence RTC by conjoining all the theoretical tenets, replacing each theoretical term with a predicate

variable, and existentially quantifying over all the variables. The resulting (unwieldy) sentence can be regarded intuitively as reading: there exist such entities as stand in such and such relations to observation (without commitment as to which particular entities they are). Then, he forms the conditional RTC -> TC. English suggests that one think of this conditional as saying, "If anything stands in these relations to observation, then let us call them 'T1',...'Tn'." (1978, 69). In this way Carnap ensures that the two components (RTC and RTC -> TC) together logically imply the original theory, TC. This formulation also has the advantage of bringing out the fact that the theory can be decomposed into two components, where RTC is the factual content of the theory and RTC -> TC is the meaning postulate.

This means that Carnap has further relaxed things so that the addition of a new purely observational postulate can also affect the significance of a theoretical term. On this view, changes in one postulate of the theory may have repercussions for the interpretations of any of its theoretical terms. If two observationally compatible theories make different stipulations in their sentences using a term t1, the interpretation of t1 is changed. But then what about some other theoretical term t2 that is not explicitly mentioned? Will it also be affected by a revision of one of the observational postulates? Carnap does not address this question, but since the sole meaning postulate (R -> TC) for the theory has also been changed, the significance of t2 and every other theoretical term is thereby altered. Since every postulate of the theory is represented in TC, any disagreement, whether theoretical or observational, leads to a difference in meaning conventions (i.e. for all the theoretical terms, not just as before for only the theoretical terms that are involved in the theoretical change). Any change, however small, is reflected in a change in the theory's Ramsey sentence, and will lead to changes in meaning of all the theoretical terms. On Carnap's revised view, even minor changes in the theory lead to changes of meaning of all the theoretical terms.[9]

---

[9] English claims that this leads to an account of meaning change more extreme than Kuhn's; she is right to point out that Carnap's later views are in some ways more extreme than Kuhn's on the subject of meaning change. But, as we shall see in the following paragraph and in section 1.5., there is at least one way in which Carnap's views continued to be more

Despite the extreme variability in the meanings of theoretical terms, Carnap wanted to claim that the meanings of observational terms escape such scientific changes unscathed because their meanings are given independently of the theory involved. It is not clear how Carnap would assign meanings to the observational terms, but there is an assumption that their significance can be specified without reference to the scientific theory in which they happen to play a part. In the later paper, he writes: "It is assumed that the terms of VO [the observational terms] designate directly observable properties or relations, and that their meanings are completely understood." (1963, 959) Carnap may have thought that their meanings were given directly by ostension, though the problems with this naive view could not have been unknown to him. But even the fixity of observation terms could not guarantee inter-theoretic comparisons of meaning for theoretical terms, since Carnap had abandoned strict definitions of theoretical terms in observational vocabulary.[10]

To sum up, Carnap shares two of the three characteristics identified in Campbell's account: extreme holism, and the theory-observation distinction. But there is one crucial

---

moderate than Kuhn's: the distinction between observational and theoretical terms. Affinities between Carnap and Kuhn, as well as Carnap's talk of "revolutions" in science may seem surprising to anyone brought up on the idea that their accounts of scientific change were diametrically opposed. But that standard interpretation is beginning to be questioned; see for example John Earman (1993), and Gürol Irzik and Theo Grünberg (1995). Irzik and Grünberg criticize Earman for failing to realize that Carnap subscribed to the thesis of semantic holism.

[10] Irzik and Grünberg argue that even observational terms do not escape Carnap's holism, and that they are equally subject to meaning change. They point out that all that Carnap assumes is that "the meanings of observation sentences are nonproblematic in the language community," not that their meanings are given by ostension. (1995, 292) They claim further that observational terms get part of their meanings from meaning postulates and are theory-laden. However, they do not refer explicitly to Carnap's reply to Hempel, from which my interpretation of this issue is derived. Be that as it may, if they are correct, this would make Carnap's view even closer to Feyerabend's and (perhaps) Kuhn's, and would make the need for an account of the meaning of scientific terms all the more pressing.

distinction, namely that the holism is tempered by the fact that observation terms are not affected by the sweeping changes that afflict theoretical terms. When it comes to the distinction between theoretical and observational terms, at least in his later work, Carnap had a meaning-change view: with every substantial change in either the theoretical or the observational sentences of the theory, all the theoretical terms change in meaning.

1.4. Feyerabend: Incommensurability and Extreme Holism

In this revisionist and very selective history of twentieth-century philosophy of science, the transition from Carnap's views to Feyerabend's and Kuhn's should not be seen as an epistemic rupture. After Carnap claimed that every substantial change in a theory leads to a change of the meaning of all the theoretical terms, it is perhaps not surprising that someone should have taken things a step further. It remained for Feyerabend to deny the distinction between theoretical and observational terms, and go on to assert that with every substantial change in a theory, all the terms of that theory (without exception) change in meaning. Feyerabend usually allowed that this meaning change did not transcend the boundaries of the particular theory in question to infect the language as a whole (though, presumably, some of the observational terms would also be deployed in other parts of the total language). Still, if all the terms of the new theory are different in meaning from the terms of the old, then the two theories cannot be compared in the most natural and immediate fashion. This was the notorious claim of "incommensurability", which Feyerabend and Kuhn hit upon independently and justified in somewhat different ways, as I shall argue in this section and the next.

One of Feyerabend's most detailed attempts to illustrate the notion of incommensurability involves the medieval European impetus theory and Newtonian classical mechanics. He claims that the concept of impetus, as fixed by the usage established in the impetus theory, cannot be defined in a reasonable way within Newton's theory. (1962, 66) On the basis of this and other considerations, he holds that when a transition is made from a theory T' to another theory T, which covers all the phenomena covered by T' as well as some new phenomena, something more radical happens than the simple incorporation of T' into T. He explains: "It is rather a replacement of the ontology of

T' by the ontology of T, and a corresponding change in the meanings of all descriptive terms of T' (provided these terms are still employed)." (1962, 68)

On several occasions Feyerabend explains the reasons for incommensurability by saying that there are certain "universal rules" or "principles of construction" that govern the terms of one theory and that are violated by the other theory. Since the second theory violates such rules, any attempt to state the claims of that theory in terms of the first will be rendered futile.[11] "We have a point of view (theory, framework, cosmos, mode of representation) whose elements (concepts, 'facts', pictures) are built up in accordance with certain principles of construction. The principles involve something like a 'closure': there are things that cannot be said, or 'discovered', without violating the principles (which does not mean contradicting them)." (1975, 269) After terming such principles "universal", he proposes that a discovery, statement, or attitude is incommensurable with a theory if it suspends some of that theory's universal principles. As an example of this phenomenon, consider two theories T and T', where T is classical celestial mechanics, including the spacetime framework, and T' is general relativity theory. About these theories, Feyerabend claims:

> The classical, or absolute idea of mass, or of distance, cannot be defined within T'. Any such definition must assume the absence of an upper limit for signal velocities and cannot therefore be given within T'. Not a single primitive descriptive term of T can be incorporated into T'.... the meanings of all descriptive terms of the two theories, primitive as well as defined terms, will be different: T and T' are incommensurable theories... (1965b, 115)

Such principles as the absence of an upper limit for signal velocities govern all of the terms in celestial mechanics and these terms cannot be expressed at all once such principles are violated, as they will be by the general theory of relativity.

---

[11] Notice that this seems to imply that at least those rules themselves are expressible in terms of the two theories--otherwise, it is not clear how Feyerabend can tell they are not shared by those theories.

The reason that these universal rules infect the meanings of all the terms of the theory that contains them is to be found in Feyerabend's theory of meaning, which he calls a "contextual theory of meaning". He uses this contextual theory to define "strong alternatives" to a given scientific theory: theories that can be considered true competitors to a dominant theory, as opposed to those that are mere variants. One of the main properties of strong alternatives is that they disagree everywhere if they disagree at a finite number of points. (1965b, 115) In other words, one sign that a theory is substantively different from another is that the differences between them infect the meanings of all terms; otherwise, Feyerabend implies, the rival theory is not a genuine alternative but a mere variant. All such "strong alternatives" are incommensurable. According to Feyerabend, the meaning of a term is not an intrinsic property of it, but is dependent on the way in which the term has been incorporated into a theory. (1962, 74) This is the gist of what he calls a "contextual theory of meaning". It also accords with his ridicule of what he calls the "hole theory" or the "Swiss cheese theory" of meaning, which holds that the conceptual cavities in a theory or language can be plugged without displacing the meanings of any of the existing terms. "According to the hole theory every cosmology (every language, every mode of perception) has sizeable lacunae which can be filled, leaving everything else unchanged." (1975, 266) The idea seems to be that the meaning of every term is affected by the general principles governing the theory, and that the principles change with every substantial theoretical change, so that the meaning of every term also changes. But Feyerabend concedes that large parts of our total theory of the world remain constant across some scientific theory changes. "It may be readily admitted," he writes, "that the transition from T to T' will not lead to new methods for estimating the size of an egg at the grocery store..." (1965a, 100) And he says that the transition from Newtonian mechanics to the general theory of relativity has left the arts, ordinary language, and perception unchanged. (1975, 271)[12]

---

[12] However, on one occasion, he writes of the conceptual disparity between classical mechanics and special relativity theory as follows: "This conceptual disparity, if taken seriously, infects even the most 'ordinary' situations: the relativistic concept of a certain shape, such as a table, or of a certain temporal sequence, such as my saying 'yes', will differ

One significant feature of Feyerabend's view is that he does not think that incommensurability is incomparability tout court. He countenances, and indeed recommends, alternative modes of comparison. Feyerabend says that "the use of incommensurable theories for the purpose of criticism must be based on methods that do not depend on the comparison of statements with identical constituents. Such methods are readily available." (1965b, 115) But although he mentions a number of methods, he does not explicate them in full and they remain promissory notes. For example, he says that theories can be compared using the "pragmatic theory of observation", according to which you attend to causes of the production of a certain observational sentence, rather than the meaning of that sentence. (1962, 93) And he insists that there may be empirical evidence against one theory and for another theory without any need for similarity of meanings. (1965b, 116) He does not elaborate further, but these claims are difficult to uphold given his insistence that even the meanings of "descriptive terms" are different in incommensurable theories. He also argues that "when making a comparative evaluation of classical physics and of general relativity we do not compare meanings; we investigate the conditions under which a structural similarity can be obtained." (1965a, 102-3) In addition, he maintains that it is possible to use incommensurable theories for the purpose of mutual criticism, adding that this removes "one of the main 'paradoxes' of the approach" that he suggests. (1965b, 117) Again, it is not clear how a "structural similarity" is apprehended or how "mutual criticism" can take place. Finally, Feyerabend uses an analogy also used by Kuhn to explain a scientist's ability to learn a new theory, that of a child learning a new language. Rather than translating between languages, "We can learn a language or a culture from scratch, as a child learns them, without detour through our native tongue..." (1987, 266) However, he does not elaborate further on the nature of this

---

from the corresponding classical concept also." (1970, 222) The apparent discrepancy might be resolved by saying that some purportedly observational terms change in meaning but not all; the ones that remain fixed in meaning are those that are deployed only in contexts far outside that of the scientific theory being examined (thus suggesting a contextualized theory-observation distinction).

language-learning process and, in the absence of a concrete proposal, it is difficult to assess his repeated insistence that theories can be compared without translation.[13]

For Feyerabend, the claim that the meanings of scientific terms from one theory are all different from the meanings of the terms in another theory rests on the premise of extreme holism. The (partly implicit) line of reasoning is that if certain scientific rules or principles are modified or abandoned in the course of the history of science, they somehow affect the meanings of all the terms in a theory, rendering the new theory unfit for a direct linguistic comparison with the old. This appears to follow from an extreme version of holism about meaning, the thesis that a substantive theoretical change cannot fail to affect the language as a whole, that a rupture in the network of concepts will displace many of the nodes, making for a radical mismatch between the old network and the new, and rendering them incapable of translation one into the other, or both into a common language. This kind of extreme holism will be countered and distinguished from a more moderate version of holism to be expounded and defended in Chapter 3.

1.5. Kuhn: Incommensurability and Untranslatability

As we shall see in this section, Kuhn's account of incommensurability changes over time and it is difficult to attribute to him a single theory of meaning, however sketchy. Moreover, it takes some time for Kuhn's views on the meaning of scientific terms to gel. In early writings, a number of distinct reasons for the extreme variability of meaning among scientific terms can be discerned; indeed, there are some indications in his early writings that incommensurability does not have much to do with meaning change at all. But that impression is corrected in his later writings on the subject of incommensurability, which revert to framing things in terms of meaning change, albeit of a local rather than global character.

---

[13] On a sarcastic, though revealing, note Feyerabend states: "Of course, some kind of comparison is always possible (for example, one physical theory may sound more melodious when read aloud to the accompaniment of a guitar than another physical theory)." (1975, 232)

In the Structure of Scientific Revolutions Kuhn often puts incommensurability in terms of change of meaning or change of concept. He writes that the referents of the Einsteinian concepts space, time, and mass, are not the same as the Newtonian concepts that bear the same names, adding that the need to change the meaning of established and familiar concepts is central to the revolutionary impact of Einstein's theory. Kuhn refers to this revolutionary change from classical to relativistic mechanics as a "displacement of the conceptual network". (1970a, 102) In the "Postscript" to the text, he reiterates the view that incommensurability involves differences in meaning between two agents espousing incommensurable theories: "Two men who perceive the same situation differently but nevertheless employ the same vocabulary in its discussion must be using words differently. They speak, that is, from what I have called incommensurable viewpoints." (1970a, 200)

However, in the same text, he sometimes suggests that translation is possible between two incommensurable theories or paradigms. At one point, he affirms that the participants in a communication breakdown can recognize each other as members of different language communities and then become translators, resorting to "shared everyday vocabularies" in doing so. (1970a, 202) If this is carried out successfully, Kuhn thinks, then: "Each will have learned to translate the other's theory and its consequences into his own language and simultaneously to describe in his language the world to which that theory applies. This is what the historian of science regularly does (or should [do]) when dealing with out-of-date scientific theories." (1970a, 202)

Since Kuhn sometimes suggests that translation is indeed possible between two incommensurable scientific theories, how are we to understand the claim of incommensurability? At some points in the "Postscript" to the text, he hints that it is a claim about the impossibility of a more general assessment of two scientific theories. This second construal of the notion of incommensurability, that it precludes a neutral way of appraising scientific theories, seems to rest on a different assumption, namely that scientific theories or paradigms contain within themselves their own standards for success or criteria of appraisal. Not only do scientific paradigms differ "about the population of the universe and about that population's behavior," Kuhn writes that they are also "the source of the methods, problem-field, and standards of solution accepted by any mature scientific community at any given time." (1970a, 103) These "non-substantive differences" are an

integral part of incommensurability, which is demonstrated by the fact that adherents of two scientific paradigms "will inevitably talk through each other when debating the relative merits of their respective paradigms...," since "each paradigm will be shown to satisfy more or less the criteria that it dictates for itself and to fall short of a few of those dictated by its opponent." (1970a, 109-110)

However, in later developments of Kuhn's view, less emphasis is placed on what might be called "evaluative incommensurability" and more on "linguistic incommensurability". Indeed, by 1983, Kuhn appears to have moved away from evaluative incommensurability entirely by saying that speaking of differences in "methods, problem-field and standards" is "something I would no longer do except to the considerable extent that the latter differences are necessary consequences of the language-learning process." (1983a, 684n.3) And, in later work, Kuhn states quite baldly: "Incommensurability thus equals untranslatability..." (1990, 299) By way of explanation, he adds that his original discussion concerned non-linguistic forms of incommensurability in addition to linguistic ones, but that he simply failed to realize how much the apparently non-linguistic component was invested in language.

Not only does Kuhn, in his later work, take incommensurability more explicitly to be the denial of translatability, he also states that this version of the claim was the same as the "original version" of the incommensurability thesis, which he characterizes as follows: "The claim that two theories are incommensurable is then the claim that there is no language, neutral or otherwise, into which both theories, conceived as sets of sentences, can be translated without residue or loss." (1983a, 670) Therefore, if incommensurability equals untranslatability, what is it about scientific paradigms that precludes translation into a single common language, so that their claims can be set side by side and their points of agreement and disagreement isolated? Moreover, how does this claim square with Kuhn's earlier claim (in the "Postscript") that historians of science can and do translate out-of-date scientific theories?[14]

---

[14] Some commentators on Kuhn have regarded this as the supreme irony of his work, that he denies translatability while at the same time serving as an articulate expositor of historical scientific theories.

The resolution of this tension lies in what Kuhn says after equating incommensurability with untranslatability: "...what incommensurability bars is not quite the activity of professional translators. Rather, it is a quasi-mechanical activity governed in full by a manual that specifies, as a function of context, which string in one language may, salva veritate, be substituted for a given string in the other." (1990, 299) Such a "quasi-mechanical translation" cannot be effected because of certain concrete problems posed by the translation of a scientific theory by a translator who does not share that theory. Kuhn claims that the problems of translating a scientific text into a foreign language or a later version of the same language are very similar to the problems of translating literature. In an illuminating passage that is worth quoting in full, he comments on the translational difficulties that are shared among literary and scientific discourse:

> In both cases the translator repeatedly encounters sentences that can be rendered in several alternative ways, none of which captures them completely. Difficult decisions must then be made about which aspects of the original it is most important to preserve. Different translators may differ, and the same translator may make different choices in different places, even though the term involved is in neither language ambiguous. Such choices are governed by standards of responsibility, but they are not determined by them. In these matters there is no such thing as being merely right or wrong. The preservation of truth values when translating scientific prose is as delicate a task as the preservation of resonance and emotional tone in the translation of literature. Neither can be fully achieved; even responsible approximation requires the greatest tact and taste. In the scientific case, these generalizations apply, not only to passages that make explicit use of theory, but also and more significantly to those their authors took to be merely descriptive. (1990, 300-1)

Kuhn does not clarify the specific translational difficulties involved here, but in other work, certain specific obstacles emerge. Although he does not always distinguish them clearly, two can be singled out for special attention. I will present them as neutrally as possible

here, but will discuss them further and respond to them in section 3.4., after I have enunciated a theory of meaning for scientific terms.

The first kind of translational difficulty implicated in incommensurability is the problem of clusters of interdefined terms. Kuhn uses the example of the eighteenth century chemical term 'phlogiston' to illustrate his point. He says that the term cannot be translated into terms of later chemical theory because of its relation to a number of other terms in the phlogiston theory, like 'principle' and 'element'. Together with 'phlogiston', Kuhn explains, "they constitute an interrelated or interdefined set that must be acquired together, as a whole, before any of them can be used, applied to natural phenomena." (1983a, 676) He acknowledges that one can introduce a neologism for a term from a previous scientific theory that is no longer part of the current scientific vocabulary. However, he suggests that when there are whole clusters of such interrelated terms, translation is no longer possible, presumably because each neologism needs to be explicated in terms of the extant vocabulary, making whole clusters of them resist such explication.

Another translational problem is that of conceptual disparity among terms. Kuhn brings this out by adverting to an example drawn from non-scientific discourse. He explains that the French term doux does not correspond to any single term of English. It "can be applied, inter alia, to honey ('sweet'), to underseasoned soup ('bland'), to a memory ('tender'), or to a slope or a wind ('gentle'). These are not cases of ambiguity, but of conceptual disparity between French and English." (1983a, 679-80) He emphasizes that doux is a unitary concept for French speakers and that English speakers have no single equivalent. English paraphrases for this French term provide no substitute because of their clumsiness and because the term must be learned together with other parts of the French vocabulary. (1983a, 685n.12) While he acknowledges that a translation manual is adequate to deal with cases of straightforward ambiguity, Kuhn argues that the examples he uses are not to be seen in this light and should be distinguished from standard examples of ambiguous words, such as 'bank' or 'cape'. The reason seems to be that it is crucial for French speakers, as opposed to English speakers, that there is a single concept at play, rather than a single term that happens to stand for a number of distinct concepts. Thus, a translation that substituted a different English term for doux depending on context would

be misleading. Though he does not explicitly say so, a scientific example of this phenomenon might be found in Kuhn's discussion of one of Aristotle's physical concepts, which he says contains "two disparate criteria", the first giving rise to our concept "average speed" and the second to our concept "instantaneous velocity". (1977, 246-7) However, Aristotle himself never made the distinction and employs what he would consider to be a unitary concept.

There are two things to note about Kuhn's views, which provide important contrasts with Feyerabend. The first is that Kuhn's variety of incommensurability is less global than Feyerabend's and can be localized in the vicinity of a cluster of terms. Feyerabend holds that fundamental changes of theory lead to changes of the meanings of all the terms in a particular theory, while Kuhn does not. The other significant difference between them, which is closely related to the first, concerns the reasons for incommensurability. While Feyerabend's variety of incommensurability seems to result from a kind of extreme holism about the nature of meaning itself, Kuhn thinks that incommensurability stems from specific translational difficulties involving problematic terms.

One point of agreement between Kuhn and Feyerabend is that both deny that incommensurable theories cannot be compared at all. Kuhn says that some comparisons will involve concrete measurements of phenomena, presumably ones described in terms shared by the two theories. He states that "proponents of different theories can exhibit to each other, not always easily, the concrete technical results achievable by those who practice within each theory." (1977, 339) Although he claims that the Ptolemaic theory and Copernican theory are incommensurable because of such problematic terms as 'planet', "The quantitative superiority of Kepler's Rudolphine tables to all those computed from the Ptolemaic theory was a major factor in the conversion of astronomers to Copernicanism." (1970a, 154) But he also advances other criteria for comparison; for example, "there are arguments... that appeal to the individual's sense of the appropriate or aesthetic--the new theory is said to be 'neater', 'more suitable', or 'simpler' than the old." (1970a, 155) Grounds for comparison remain despite incommensurability, including "accuracy, scope, simplicity, fruitfulness, and the like." (1970b, 261) Still, a complete translation is impossible and the tenets of the two theories cannot be directly compared. As for the

indirect methods, they are not fully explicated and it is not clear that they can be implemented in the absence of linguistic commensurability.

According to Kuhn's mature view, it is not possible to phrase all the claims of two scientific theories in a single language so that they can be put side by side and their exact differences pinpointed. Kuhn thereby denies the possibility of the most direct and natural method of comparing two scientific theories. As a result, choices between scientific theories are not based on a point-by-point comparison. Scientists who learn a new theory do not merely translate the new terms into the old terms; rather they begin from scratch in the way that learners of a natural language do. A language learner, Kuhn states, will not always "be able to translate from his newly acquired language to the one with which he was raised." (1990, 300) Since Kuhn's later work takes incommensurability to be the denial of translatability, I will argue against this claim in Chapter 3 based on the conception of translation that I will elucidate in that chapter. Specifically, I will show that the problem of interpreting scientific discourse is more manageable than that of interpreting literary discourse and show how the translational obstacles mentioned above (under the rubric of local incommensurability) can be dealt with.

1.6. Scheffler: Reference and Observational Terms

One of the earliest responses to Feyerabend's and Kuhn's claims of incommensurability is contained in Israel Scheffler's Science and Subjectivity. Scheffler attempts to restore some degree of objectivity to science by countering, among other things, what he calls the "paradox of common language," which he understands as the claim that there can be "no intelligible converse between scientists of differing theoretical persuasions," since, "There are perhaps common sounds but no common meanings." (1967, 16) As a result, each scientist is "effectively isolated within his own system of meanings..." (1967, 17) In response, Scheffler tries to rehabilitate the notion of communication across the theoretical divide but without reverting to all the assumptions made by the "standard view" of the logical empiricists. This last he refers to as the "two-tier view", an allusion to the distinction that the logical empiricists made between theory and observation. Scheffler does not think it possible in light of the work of Kuhn, Feyerabend and N.R. Hanson, to maintain an invariant boundary between the theoretical vocabulary and the observational

vocabulary. In addition, he does not think that meanings are unchanging and can be fixed in the face of all theoretical changes. Still, he holds that it is possible to make science safe for cumulativists who see scientific change as a gradual and rational process. He does this by making use of the sense-reference distinction, as well as by introducing a kind of contextual theoretical-observational distinction. His overall aim is to combine the strengths of both views into "a new and coherent objectivism". (1967, 54)

According to Scheffler, it is the reference of observational terms that provides a means of comparison among theories. These observational terms are not fixed for all time; rather, certain terms pass into and out of the observational realm as science advances. But at any juncture, one can identify the observational ones, and hence weigh the claims of competing theories side by side. These points are not absolutely explicit in Scheffler's text, but they can be inferred. He claims that "for the purposes of mathematics and science, it is sameness of reference that is of interest rather than synonymy, in accordance with the general principle that a truth about any object is equally true no matter how the object is designated."[15] (1967, 57) He goes on to say that the reference of terms that feature in experimental laws can be determined independently of their sense, since experimental laws are largely observational in nature. However, Scheffler insists that, "The relative independence of observation from theory must not be taken to imply that there is some single descriptive language, fixed for all time, within which science must forever fit its experimental accounts of nature." (1967, 65)

Scheffler's position is complicated by the claim that the reference of these terms can be independent enough to allow theory-comparison, yet not so independent as to float free of theory. His remarks on the subject of meaning sometimes evoke extreme holism:

> In general, terms possess meanings not as isolated attachments, but rather as organic qualities which accrue in virtue of systematic function within a framework of linguistic use and intent... The meaning of a category name is thus dependent

[15] In saying that "a truth about any object is equally true no matter how the object is designated", Scheffler seems to be assuming that scientific theories do not involve intensional contexts, such as belief ascriptions and modal statements. While this is a justifiable claim, Scheffler does not enunciate it explicitly.

> upon the language in which it has a place. To alter this language is to alter its relative location, and so its very meaning. (1967, 45-46)

However, although "connotative meaning" (or sense) is generally theory-dependent, Scheffler thinks that it may remain constant across specific instances of theory changes. This is all the more true of reference, since identity of reference does not imply identity of sense, and reference "is relative to language though shareable by theoretical opponents..." (1967, 60) Thus, experimental laws "may retain their referential identities throughout variations of theoretical context." (1967, 61) But Scheffler does not give a precise way of determining sameness of reference of the terms in these laws. If the meaning (both sense and reference) of theoretical terms is drastically theory-dependent, and if sense is also variable for observational terms, then one needs to specify a means of determining reference for observational terms in order to assure a means of comparison.

The solution to this problem is not explicitly provided by Scheffler. Since the terms with constant reference are all observational ones, perhaps he thinks that reference can be determined by way of ostension or ostensive definition. This is suggested by the fact that Scheffler alludes with approval to Ernest Nagel's doctrine of the "meaning-independence of experimental laws". (1967, 48) However, Scheffler breaks with Nagel in denying that meaning-independence applies to the sense of observational terms and holds only that it pertains to reference. (1967, 63-4) Nagel's view follows the standard logical empiricist line, at least in its mature version. When it comes to theoretical terms, Nagel says that their meanings change when the theory does. For example, "though the same word 'electron' is used in pre-quantum theories of the electronic constitution of matter, in the Bohr theory, and in post-Bohr theories, the meaning of the word is not the same in all these theories." (1960/1979, 88) By contrast, Nagel claims that an observational term (or, as he puts it, a nonlogical term featured in an experimental law) retains a meaning that can be formulated independently of the theory. (1960/1979, 86-87) The meaning of each observational term "is associated with at least one overt procedure for predicating the term of some observationally identifiable trait when certain specified circumstances are realized." He goes on to say: "The procedure associated with a term in an experimental law thus fixes a definite, even if only a partial, meaning for the term." (1960/1979, 83) While Scheffler does not talk about overt procedures that fix the meaning (or, in his case, the reference) of

observational terms, it appears as if he would have to take this or a similar line in order to use the reference of observational terms to effect the comparison between theories, as he requires.

But this does not solve Scheffler's problem. He cannot consistently follow Nagel and adopt certain overt procedures for determining the reference of observational terms. The reason is that he has denied that the reference of observational terms remains constant for all time, as it would if there were uniform ostensive procedures for determining reference. His replacement of the logical empiricists' rigid distinction between theory and observation with a more contextual distinction implies that he cannot rely on something like Nagel's overt procedures. Scheffler takes all the terms that have a shared reference in two theories to be the observational ones and the others to be theoretical. This pushes us in the direction of reformulating the theoretical-observational distinction in the following way: those terms that have a common reference in two theories should be considered observational <u>for those two theories</u>, while those that are not held in common will be the theoretical terms. In this way, the old distinction is relativized to pairs of theories, making it more contextual and basing it on a more mundane and straightforward criterion. This is an appealing proposal and foreshadows some philosophical discussions subsequent to Scheffler's.[16] But then the question arises, how are we to determine that such terms share a reference in the first place? Scheffler might say that the two theories will share ostensive procedures and that these can be used to effect a comparison. But this seems just to assume that for any pair of theories, the ostensive procedures will be the same, or that there will be enough shared procedures to effect a comparison. Once one has given up on the old theory-observation distinction and on a timeless way to determine the reference of observational terms, it becomes implausible just to assume that there will be shared ostensive procedures that will enable any two theories to be compared (even any two successive theories). The problem of demarcating observational terms and the problem of

[16] See, for example, David Lewis' position on theoretical terms, to be discussed in section 6.3.

determining their reference are of a piece, and Scheffler does not supply a replacement to the logical empiricist criterion that would enable us to do both.

1.7. Extreme Holism and Inter-Theoretic Assignments of Meaning

If there is a culprit in this story of successive attempts to secure the meaning or reference of scientific terms, it would seem to be what I have been calling the doctrine of extreme holism. The idea that a change at one point in the system jostles everything else (or a sufficiently large number of things) has been the spoiler for a feasible account of meaning change in science. Campbell accepted it, saying that stability of meaning was not even necessary within a single theory, thereby implying that deductive arguments had no place in science. Carnap also endorsed it, although it was mitigated in his case by the (untenable) assumption that observational terms got their meanings unproblematically, perhaps by ostension. In the case of Feyerabend, and perhaps the early Kuhn, this conclusion was embraced willingly for all terms of a theory, theoretical as well as observational, and led to the notorious claim of incommensurability. Finally, Scheffler accepted it for everything but the reference of observational terms, and tried to circumvent it there by appealing to a referential model of meaning that was not supposed to be subject to extreme holism. But he failed to specify how reference was to be fixed. In later chapters, I will distinguish the thesis of extreme holism from another variety of holism about linguistic meaning that does not have this consequence. In Chapter 3, I will claim that one can be a holist about the meaning of scientific terms without accepting the claim that every change in theory leads to a change of meaning of all (or many) of the terms involved. This also involves rejecting an atomistic view of meaning according to which terms get their meanings piecemeal, say by way of direct relations with physical determinants.

There are other problems with the accounts of the meanings of scientific terms surveyed in this chapter. To prepare further for the positive account that I will be proposing, I will point to one that is very prevalent. None of these authors makes a clear distinction between inter-theoretic and intra-theoretic assignments of meaning. For Carnap, the two are apparently linked, since the idea is that one gives the meaning (or "empirical significance") of a theoretical term by way of observational terms, so theoretical terms from different theories might be linked up indirectly via the common observational

terms. But intra-theoretic definitions are best abandoned in this context, in favor of inter-theoretic assignments of meaning. Whenever there is a demand for giving the meaning of a theoretical term, this should be understood as a demand for providing an equivalent term from another theory. Within a single theory, a demand for giving the meaning of a term can always be transformed by "semantic descent" into a demand for a theoretical explication of some kind. On this way of doing things, one does not give the meaning of a term from one's own theory in terms drawn from that very theory. A request for the meaning of our term 'mass' by someone who shares our theory can always be transformed into a request for an explanation of what mass is, of its properties or its connection to other theoretical entities. Since there are no definitions in the context of inquiry, in the sense of unrevisable theoretical tenets, there may be various different ways of supplying such explanations, although there may be a particularly perspicuous one relative to a certain presentation of a theory or a certain context. There may also be situations in which two terms within a single theory are exactly interchangeable (at least synchronically), so that one of them can be furnished to explicate the other. In later chapters, one member of such a pair will be considered a "redundant" term, but supplying the other member of such a pair is not to be confused with giving inter-theoretic meaning assignments, which are of a different order altogether. Inter-theoretic equivalences constitute part of an overall mapping or translation function that can be constructed between the terms of two scientific theories, and that is how assignments of meaning are to be understood in the rest of this work.

Finally, at least since Scheffler's discussion, the distinction between sense and reference (or connotation and denotation, or intension and extension) has been prominent in discussions of scientific change and has been invoked to show how relative fixity of subject matter can be maintained in the face of radical change in theory. The idea is superficially appealing: substantive revisions in theory can be correlated with changes in the sense of scientific terms, while constancy of ontology can be explained by adverting to sameness of reference. But on closer inspection, this formula is merely a promissory note. First, one needs a way of determining reference that does not just rely on ostensive procedures, particularly since in many experimental situations, a whole range of entities, properties, and relations may be picked out by simple pointing or even by more sophisticated perceptual identification (as will be argued in section 2.5.). Second, if

reference is not to be determined by the precise content of a theory (which might raise the threat of radical reference change), neither should it be completely theory-independent, since we are interested in the reference of the terms of the particular theories that we need to compare. This latter point will become clearer in the course of the following chapter, where I will turn to an examination of an account of reference that purports to solve the problem and provide a theory of reference for scientific terms. After rejecting that account, I will proceed for the next three chapters as though the proper way of comparing scientific theories is by concentrating on their concepts or the meanings of their terms, rather than the reference of those terms. This central assumption will be further justified in Chapter 6.

Chapter 2: Reference

> Our speech, like everything else, has its defects and weaknesses. Most of the world's squabbles are occasioned by grammar! Lawsuits are born from disputes over the interpretation of laws; most wars arise from our inability to express clearly the conventions and treaties agreed on by monarchs. How many quarrels, momentous quarrels, have arisen in this world because of doubts about the meaning of that single syllable Hoc [this]. Michel de Montaigne, Apology for Raymond Sebond

2.1. Appeals to Reference

A common rejoinder to the claim of the incomparability or incommensurability of successive scientific theories makes an appeal to the reference of scientific terms. Some authors roundly claim that meaning is a problematic notion and proceed to focus on reference instead. Some such appeals to reference or extension were encountered in the previous chapter, although the authors discussed did not introduce a full-blown theory of reference so much as rely on an intuitive grasp of the notion. Several later appeals to reference are united by their commitment to a theory of reference first proposed by Keith Donnellan, Saul Kripke, and Hilary Putnam, usually known as the "causal theory of reference" (sometimes also the "causal-historical theory" or the "new theory"). The aim of this chapter will be to detail the problems that the theory faces when it tries to explain how scientific terms refer, thus demonstrating that it is inadequate to resolve the problem of theory comparison.

Although it may not have been designed for the purpose, the causal theory of reference appeared on the philosophical scene in the wake of discussions of incommensurability and was quickly embraced by many philosophers of science as the solution to their problems.[1] In Putnam's writings on the subject, one of the causal theory's

[1] The version of the causal theory that I will be relying on is culled from Hilary Putnam (1973a), (1973b), and (1975), Keith Donnellan (1974), and Nathan Salmon (1981). It has been pointed out that the "causal theory of reference" is something of a misnomer, but the name has gained more currency than the "new theory" or the "causal-historical theory". Significantly, Kuhn seems to think that the causal theory is the main rival to his account of the meaning of scientific terms. He has made some criticisms in his (1990), but they are different from, and more restricted in scope than, the ones that I make in this chapter. His main quarrel is with the essentialism associated with the causal theory, although he is briefly concerned with the problematic nature of re-baptism, which is to be discussed in

main attractions seems to have been the promise it held out for solving the problem of the meaning or reference of scientific terms. Even philosophers who are uneasy with one or the other aspect of the causal theory have considered it a potential solution to this problem. Ian Hacking might be taken as a typical representative of this attitude: "I do not literally believe Putnam, but I am happy to employ his account as an alternative to the unpalatable account in fashion some time ago."[2] (1984, 159) Subsequent loss of interest in the problem of the meaning or reference of scientific terms seems to have been encouraged by the perception that the causal theory provided a solution. This attitude persists despite the fact that some of the problems with the attempt to adapt the causal theory for this purpose have become fairly well known. This chapter will reiterate some of these problems and bring up some new problems as well. I will conclude that the causal theory is unsuitable as an account of the reference of scientific terms.

2.2. The Causal Theory and Science

In this section, I will outline the way in which the causal theory of reference is meant to give an account of scientific terms. This will not only be useful for the criticisms I intend to make, it is also something that has not been done in detail by other writers who have written on this topic and there may be independent interest in it for that reason. The causal theory is committed to a conception of reference, according to which the reference of a term should not be determined by the descriptions of the properties that are

---

section 2.2, below. Two other prominent critiques of the causal theory as applied to science are found in Dudley Shapere (1982) and Arthur Fine (1975). But the former also concentrates on essentialism, and the latter puts the criticisms in very broad terms. Fine does draw attention to the difficulty of accounting for reference change within the context of the causal theory, but does not seem to countenance the possibility of re-baptism. When it comes to essentialism, Salmon demonstrated in (1981) that essentialism did not follow from the theory of reference as such, but I will argue in section 2.5. that the causal theory presupposes a particular view of scientific taxonomy which is controversial and which also enters into the derivation of essentialism. See Khalidi (1993a) for further details.

[2] Other authors have been similarly ambivalent. While making some harsh criticisms of the causal theory when applied to science, Richard Boyd claims continued allegiance to it. By way of criticism, he writes: "The kind to which a term refers is determined by the role that term plays in socially coordinated inquiry, rather than by any particular features of an introducing ceremony, or the intentions of the speakers who first introduced it." (1979, 386) Why then does Boyd continue to espouse the causal theory? At one point he states: "A causal theory of reference is true... precisely because a causal theory of knowledge is true." (1979, 380) I do not think that it follows, but this is not the place for a full discussion.

associated with the referent. In science, the descriptions of the properties in question are given by the theory to which the term belongs. The effect of the account is therefore to make reference theory-free. A term from one theory can be matched up with a term from another without regard to the theories to which those terms belong.

But the causal theory of reference goes beyond saying that the reference of some terms is theory-free. The causal theory also says that certain terms are rigid designators, that is, that they denote one and the same entity in all possible worlds. On this picture, reference cannot be mediated by way of an agent's beliefs about the properties of the referent, or the descriptions associated with the referent, or the theoretical tenets which concern the referent, because these may be false as applied to that referent in another possible world and might pick out something other than the true referent.[3] The causal theory goes on to propose a particular way of ensuring that certain terms are rigid designators by proposing a mechanism of reference. How is the referent of a term to be identified in the actual world, let alone other possible worlds, if full-blown beliefs are to be shunned? More specifically, for our purposes: if scientists use terms to refer independently of the properties that they believe are possessed by their referents, how is the referent of any given term to be pinned down and how are two agents to decide that they are referring to the same thing? The causal theory answers that reference is secured through the intention of the first user of a particular term to refer to an initial event (sometimes known as a "baptism" or "introducing ceremony") at which the referent was causally featured and first identified. On each occasion following the baptism, scientists have an intention to use the term to refer to whatever was referred to by the previous user in the historical chain. There is some reliance here on what intentions scientists have, but these do not involve scientific theories or substantive beliefs about the referent.[4]

---

[3] Of course, it is "narrow" beliefs that are not allowed to determine reference, since "wide" beliefs are themselves supposed to be directly referential. Narrow beliefs capture an agent's conception of things and involve only those discriminations the agent would make, whereas wide beliefs reach out into the world and make the discriminations that are actually out there. In speaking of the causal theory's severance of reference and belief, I will mean beliefs in the narrow sense. For more on this, see section 2.4.

[4] In retrospect, the causal-historical account of reference-fixing can be seen as a mechanism that approximates David Kaplan's 'dthat' operator which is supposed to pick out the same individual in every possible world (though Kaplan himself prefers to consider it a self-standing demonstrative term rather than an operator). Kaplan writes that some have questioned whether these mechanisms, such as the account given by the "historical chain theory" belongs to semantics or "metasemantics", and says that he is "unclear" on this point. (1989, 573)

Not only does the causal theory allow the descriptions that scientists associate with a term to be false as applied to the referent of that term in another possible world, it even allows those descriptions to be false of the referent in the actual world. If all the descriptions that scientists believe to be true about a term's referent are in fact false, they will still be allowed successful reference when they use that term provided they satisfy the conditions outlined above. The descriptions used in baptizing the referent or those associated with it in subsequent dealings can always be false, for they are needed merely as "reference-fixing" not "reference-determining" devices, in the terminology of the causal theorists. A reference-fixing description is one used as a means of specifying the appropriate historical chain, as opposed to a description that the referent actually satisfies. Therefore, the causal theory of reference holds not only that an agent's descriptive beliefs need not be true of the referent in some other possible world, but also that they need not even be true of the referent in the actual world.[5]

It is not merely that the causal theory does not have to rely on descriptive beliefs to determine reference. The theory cannot tolerate a determining connection between a scientist's beliefs about the referent and successful reference to it, because that would upset the basic claim that terms are rigid designators. As explained by Salmon, causal theorists wish to detach "the purely conceptual representation of an object" from "the mechanism by which the reference of the term... is secured and semantically determined." (1981, 12) They insist that there is no guarantee that such conceptual representations or qualitative descriptions will give the correct properties of the referent in this or any other possible world. If one were to rely on such descriptions to determine reference, the associated terms would generally cease to designate the referent in a rigid fashion. Therefore, the causal theory is incompatible with a straightforward descriptional theory of reference.

How then, it might be asked, can one identify the referent of a term for the purpose of comparison? The causal theorists reply that the causal connection at the introducing event is what singles it out. In the case of persons, that claim seems fairly straightforward. But in the case of chemical substances, say, what is the thing that is so pinpointed and how

---

[5] These two claims correspond to what Salmon calls the "modal" and "semantic" arguments for the causal theory. (1981, 23-31) One could have a theory of reference that satisfied the first clause but did not satisfy the second; an example of such a theory would be a rigidified descriptional theory of reference. As shall be seen in due course, some writers hold that the reference of some scientific terms is given by a theory of this kind.

can it be re-identified in the actual world, much less across possible worlds?[6] Their response is that something is a sample of a substance that is denoted by a certain term when it can be determined, by an omniscient observer, to have the same nature (or hidden structure or essence) as the sample first identified at the baptism at which the term was introduced. There are two notions worth examining in this answer: that of an omniscient observer and that of having the same nature.

The notion of an omniscient observer was introduced by Keith Donnellan to overcome the frailties of actual human language users.[7] Although the causal theory places minimal requirements on successful reference, it places a larger burden on anyone interested in identifying the referent of a particular term for purposes of comparison. For it requires one to be able to trace the history of a term along a chain of individual intentions. One may be required to go all the way back to the initial baptism and may even need to determine what kind of substance was featured at that event. But it is obvious that this would generally necessitate acquiring information about past events and about the intentions of individual scientists--information that is often unavailable, or worse, lost for good. The device of an omniscient observer allows one to say that such information, concerning the historical chain and the baptismal event, need only be discoverable in principle.

As for the causal theorists' use of the notion of having the same nature, it ensures that we will be able to say of any given sample or specimen that it is correctly designated by a given term. Even after discovering what kind of substance was featured at the baptism, we may need some way of telling whether the sample at another baptism is a sample of the very same kind of substance (and hence that the terms involved are coreferential). Therefore, the notion presupposes that there is some way, again in principle, of determining whether two things have the same nature. To this end, in speaking about the substance water, Putnam postulates a relation that he calls the "same$_L$ relation", a shorthand for "same liquid as". As he explains it, "the relation same$_L$ is a theoretical relation: whether something is or is not the same liquid as this may take an

[6] I sometimes speak of chemical substances as paradigmatic examples of the referents of scientific terms. This practice follows that of Putnam and other causal theorists, and it should not prejudice their case since the problems associated with substances are probably more tractable than those associated with physical magnitudes or biological species.

[7] Donnellan writes: "I have used the notion of an omniscient observer of history and, of course, we ordinary people cannot be expected to know in detail the history behind the uses of names by those with whom we converse." (1974, 17)

indeterminate amount of scientific investigation to determine." (1973b, 122) The $\underline{\text{same}}_L$ relation implicitly assumes a method for determining whether two samples are samples of the same substance, and therefore, relies on the results of the ultimate scientific theory. Salmon generalizes this notion, introducing the relation of consubstantiality (one can also speak of conspecificity when one is discussing biological species, and so on).

The causal theorist is now in a position to answer the following question: When does a term from theory T_1 have the same reference as a term from theory T_2? When they can both be traced back, if necessary by an omniscient observer, along a historical intentional chain to the same initial baptism at which a single substance was first introduced. Or, failing that, when they can be traced back to different baptisms at which different samples of the same substance were featured, as this would be determined using a criterion of consubstantiality (conspecificity, and so on). If terms from different theories can be matched up in this way, the tenets of the two theories can presumably be directly compared.

2.3. Reference Change

John Searle has amassed some counterexamples to the causal theory's account of the reference of proper names. He has also charged that the theory is not just vulnerable to counterexamples, for it does not even give the right "picture" of how proper names manage to refer. To make this point, he describes an imaginary tribe that uses words in such a way that no proper names are introduced or gain currency in the way posited by the causal theory. The possibility of the existence of such a community is meant to show that the causal theory does not represent the correct way of thinking about the reference of proper names. Searle writes:

> Imagine that everybody in the tribe knows everybody else and that newborn members of the tribe are baptized at ceremonies attended by the entire tribe. Imagine, furthermore, that as the children grow up they learn the names of people as well as the local names of mountains, lakes, streets, houses, etc., by ostension. Suppose also that there is a strict taboo in this tribe against speaking of the dead, so that no one's name is ever mentioned after his death. (1983, 240)

In such a community, there are no historical-intentional chains and proper names are associated with descriptions. Whether or not Searle's criticisms are compelling when it comes to proper names, I think they can be used to make the case that the causal theory does not give the right picture of how scientific terms refer. Analytic philosophers are prone to speak of primitive tribes when they discuss foundational questions, but some of the possibilities Searle raises can be demonstrated closer to home. Searle's natives bear a

distinct resemblance to the scientific community and their use of proper names has certain affinities to the use of general terms in scientific inquiry. Indeed, it can be argued that the linguistic behavior of scientists can depart even more radically from the picture assumed by the causal theory than does the imaginary tribe's.

As in Searle's example, members of the scientific community are regularly taught the use of their terms by ostension in the laboratory and there is rarely reliance on causal-historical chains to determine their referents.[8] While there may be no taboo against speaking of certain things, there are not many equivalents of dead referents when it comes to scientific terms (aside perhaps from extinct biological species, but the vast majority of those were never baptized before they became extinct). Most damaging of all, it is not that baptisms are attended by the entire tribe of scientists, but that there are few events that can be considered baptisms, as the theory requires. Notice that it is not enough that a certain event be taken in retrospect as baptismal; for the causal theory requires that at least one of the agents involved consider the event in question to be a baptism. Only then can one of the agents form an intention to use a term in such a way that it originates in a certain baptismal event. That is the only way to ensure that there is a chain of intentions that leads eventually back to a particular event, and therefore to fix the reference of the term in question.

These criticisms of the causal theory may seem like quibbles and might be shrugged off by saying that the theory is not meant to give an accurate description of scientific practice. However, in the absence of explicit baptisms, a criterion of coreference and an account of reference change are hard to come by. The causal theorists have long been aware of the problem of reference change, so it is worth taking some time to examine the solutions that they propose. One simple case that can be used to illustrate the problem is Putnam's well-known example of the term 'jade'.[9] It turns out that the term was long used

[8] Kuhn has elaborated on the process of learning scientific terms by saying that new terms are acquired "by exposure to examples of their use." He continues: "That exposure often includes actual exhibits, for example in the student laboratory, of one or more exemplary situations in which the terms in question are applied by someone who already knows how to use them... The exemplary situations may instead be introduced by a description in terms drawn from the antecedently available vocabulary, but in which the terms to be learned also appear here and there... Both [processes] include an indispensable ostensive or stipulative element: terms are taught through the exhibit, direct or by description, of situations to which they apply." (1990, 302) Of course, to say that terms are taught (partly) through ostension is not to say that ostensive procedures can be used to define them or to determine their meaning or reference, a view that will be criticized in section 2.5.

[9] The case is mentioned in Putnam (1975, 241).

to refer to (what we now know to be) two distinct substances, jadeite and nephrite, which are superficially indistinguishable although chemically quite different. But it is likely that the term was first introduced in the presence of one or the other of these minerals, although users of the term went on to have contact with both minerals and acquired beliefs that were equally true of both. An orthodox causal theorist, who holds the version of the theory set out in the previous section, would have to insist that the term 'jade' referred exclusively to one of the substances, namely whichever one happened to be featured at the initial baptism, and maintain that the other was no part of the reference of the term. However, many of the causal theorists themselves admit that it is more plausible to allow the other mineral to be included somehow in the reference of 'jade'.

It is not difficult to imagine how such episodes might occur in the history of science. It is sometimes found, in the case of previous scientific theories, that a term introduced with the intention of singling out one type of entity or property, has really been used to pick out two or more kinds or goes on to be used to pick out some other kind (from the perspective of a later theory). The situation seems particularly common in the case of chemical substances or biological taxa, but it can also crop up with other scientific referents. Arthur Fine has suggested that such a situation occurred with the term 'electron'. From 1891 to 1897, 'electron' referred to the unit quantity of electrical charge, but after the charge-to-mass experiments of J.J. Thomson and the increasing acceptance of the particulate nature of electricity, the term was naturally assimilated to Thomson's "corpuscles".[10] It seems crucial to Fine's point that the term was gradually applied to the particles without a conscious decision to withhold it from the unit of charge (let alone an actual baptism). That is, there was initially no explicit intention to apply it exclusively to the particles rather than the unit of charge. The orthodox causal theory does not have the resources to deal with such cases, but there have been numerous attempts to modify it to handle them. In what follows, I will examine four such attempts, due to Salmon, Berger, Nola, and Kitcher, respectively.

(a) *Salmon's account:* Salmon is one author who considers ways of accounting for the reference of the term 'jade' and similar terms. He proposes two principal strategies to deal with this problem, one of which is also used by two of the other three authors to be discussed. The first consists of positing a rebaptism and the second consists of associating a rigidified description with the term. On the rebaptism story, Salmon finds that 'jade'

---

[10] See Fine (1975, 23-26) and references therein.

designates jadeite in some contexts and nephrite in others, and he adds that this kind of equivocation should be understood as semantic ambiguity (cf. 'bank') rather than indexicality (cf. 'you'). (1981, 100n.6) Salmon considers an account according to which 'jade' changes its reference depending, not on the context of utterance like ordinary indexicals, but on the causal-historical chain with which it is associated. Since there are two causal chains, one of which supposedly began with the nephrite baptism and the other with the jadeite baptism, the term will be equivocal. Salmon compares it to the equivocal term 'bank', but unlike 'bank', the agents who use the term 'jade' do not (by hypothesis) know that it has two origins and that it is therefore equivocal. That is why it is implausible to attribute a different reference to the term according to whether it can be traced back to a nephrite baptism or a jadeite baptism. But more importantly, it is problematic to posit two intentional baptisms in these cases, since (again, by hypothesis) the agents themselves are unaware that they are dealing with two distinct substances. Even if one grants that jade was baptized once, it is exceedingly unlikely that it was baptized twice.

Perhaps because of the difficulties with this account, Salmon ventures another way of dealing with terms such as 'jade'. He suggests that earlier inquirers can be taken to refer to the "kind-union... of the two substances in question, i.e. the kind Jadeite or Nephrite." (1981, 100n.6) Notice this is not to say that 'jade' is an inclusive or disjunctive term, for that would make it descriptional. Rather, 'jade' is supposed to refer rigidly to some single metaphysical entity, a kind-union. He does not say very much about this entity, pointing out merely that it involves an analogy to class-union. However, it is implausible to say that scientists refer regularly to "kind-unions" when they themselves are unaware of the existence of such entities. Salmon admits: "The kind-union Jadeite or Nephrite is not one of the kinds of substance ordinarily treated in a chemical theory of minerals." (1981, 100n.6)

(b) *Berger's account:* Another attempt to deal with such terms within the overall framework of the causal theory comes from Alan Berger. He posits that many scientific terms are introduced in much the same way that other causal theorists envisage, by a process he calls "focusing" (such terms are called "F-type terms"). This can include ostension and description, but seems always to involve some initial perceptual encounter. Next, Berger says that a necessary condition for reference change is a genuine re-focusing. He puts forward a hypothetical example involving a twin-earth community that first uses the term 'mass' such that it applies equally to both mass and weight. Then he considers how the term manages to change its reference and come to denote the latter exclusively:

> Now imagine that at some later stage, the twin-earthians develop operational procedures, such as the use of scales, that enable them to measure easily the quantity they take to be the amount of force needed "to get an object going"... Since

> the new procedure is so easy, it can be quickly and widely adapted, it requires merely focusing on meter readings, etc. Thus the new stage in the transmission of the reference of the term 'mass' for twin-earthians takes place through a genuine focusing. (1989, 196)

What is crucial in this variation on the causal theory is the judgment about what to identify as an initial focusing or as a genuine re-focusing, for these determine coreference and are indicators of reference change. But identifying either of these is controversial at best, since a focusing, unlike a baptism, is not intended as such by the agents involved. Since different accounts can be given of the history of a term's usage and of what should be taken as a focusing or re-focusing, this raises the possibility of different pronouncements on the reference of the relevant terms. Indeed, the same story Berger tells might have been cited to bolster a descriptional account. A pure descriptional theorist might say that the term 'mass' changed reference with the invention of scales, not because this constituted a re-focusing, but because the inquirers came to associate new descriptions with the term, ones which fit the property of weight.

This is why Berger's theory represents a genuine departure from the orthodox causal theory sketched in the previous section. A crucial requirement of the causal theory is that there should be a historical chain of intentions that terminates in an agent's intention to refer to the physical magnitude featured at the baptismal event. The importance of this condition is not hard to find, since without it there is generally no definite answer as to which event determines reference. Berger waives the condition that the initiating event be regarded as such by at least one agent involved, presumably because the case he discusses involves scientists who are unaware that two different quantities are being dealt with, and so do not attach particular significance to one event as opposed to another.[11]

---

[11] According to Berger, a necessary condition for reference change for an F-type terms is transmission "by a genuine focusing on a new referent." As he goes on to explain, "Here, current or later focusings can dominate over previous or even initial focusings in determining the term's referent." (1989, 188) A genuine focusing is "a process through which a linguistic community passes an F-type term along a historical chain (of speakers' intentions to corefer) by means of currently focusing on an object--an object that the community takes to be the referent of a term." (1989, 187) If members of the community in question focus on their perceptual encounters with an object (or quantity, or natural kind) and associate a term with these focusings, they thereby manage to change the referent of the term. The change in reference is unintended, but the focusings are obviously intentional actions and are taken as a sign that the term has been given a new referent by the community. Berger acknowledges that the community in such cases may also associate descriptions with the term in question and may also pass it on by relying on

But even if Berger's re-focusings were not problematic, there would be a difficulty with taking the term to have an inclusive reference before the re-focusing or re-baptism. This is what Berger does with the term 'mass' and what Putnam apparently does with the term 'jade'. Berger claims, in his hypothetical story, that the term 'mass' applied equally to mass and weight at first. He thinks that at least some scientific terms are of this sort, so that the term refers (in all possible worlds) to whatever satisfies the initial description in the actual world. He calls such terms "S-type" terms. Since initial descriptions are not likely to endure, such terms (S-type) can turn into the other kind of term (F-type) by re-focusing. In Putnam's case, it involves taking the term 'jade' as used by earlier chemists to refer to samples of mineral that have all and only the macro-properties shared by jadeite and nephrite. Strictly speaking, that option is not open to the orthodox causal theorists because they take the relation of reference to obtain directly between a term and a certain natural kind that appeared at the baptism. Since they assume that there is no such natural kind that shares the macro-properties of jadeite and nephrite, just the natural kinds of jadeite and of nephrite, that option is closed. But both Putnam and Berger effectively amend the causal theory for those terms, and rule them to be rigidified descriptional terms before the re-baptism. That is, they take these terms to refer, in all possible worlds, to whatever satisfies a certain description in the actual world. As Salmon observes, "Putnam's account treats the term 'jade' as if it were descriptional in terms of the usual identifying characteristics for jade, thus univocally designating a certain nonnatural kind or category, perhaps the kind Hard Translucent Stone that is Green or White in Color." (1981, 100n.6)

(c) *Nola's account:* The rigidified description route is also roughly the one taken by Robert Nola. In one of the most historically minded modifications of the causal theory, Nola has argued that ostension at the baptism is an unsatisfactory way to fix the reference of theoretical terms. Therefore, he considers scientific terms to be introduced by means of a reference-fixing description of the following form: "'whatever causes effects O (in some specifiable conditions).'" (1980, 506) However, he finds that such descriptions will pick out events rather than entities, so he proposes to express events as triples of objects,

---

such descriptions (by "mock focusing"). But he assumes that where both descriptions and perceptual encounters are relied upon, the latter will be favored. This assumption seems to beg the question against the description theorist. Such a theorist might insist that the description should determine reference in these cases rather than perceptual encounters. According to orthodox causal theorists it is the agents themselves who decide which intention is primary; one cannot just assume that the community will rely on genuine focusing as the determining factor. Thus, Berger reads certain intentions into their actions.

properties, and times. But to decompose events in this way, he argues, we require prior theoretical beliefs. Thus, theoretical beliefs play a crucial role in picking out a definite theoretical object to which the scientist will attach a newly-coined term. Moreover, Nola says that the description that picks out the theoretical object that scientific baptizers intend to single out will be a priori true (but not analytic or necessary). This is a major point of departure from the orthodox causal theory and one that makes clear Nola's espousal of a rigidified descriptional theory. On his account, the description first used to pick out a theoretical object is satisfied by that object in the actual world; moreover, it cannot be found to be false, on pain of finding that the term fails to refer. As he puts it, "Once 'electricity' has been introduced by a reference-fixing definition, it is not possible, as a result of subsequent developments in physics, to declare that electricity has been discovered not to produce the causal effects attributed to it in the definition and at the same time to make this claim about the very same thing, electricity, that is picked out in the definition." (1980, 515) Either the initial definition is true of electricity or 'electricity' fails to refer. When Nola goes on to give a similar account for the term 'phlogiston', the definition becomes even more elaborate in order to show how 'phlogiston' was found not to refer. He posits that the phlogiston theorists fixed the reference of 'phlogiston' by picking out precisely those effects that were later discovered not to be caused in the way they thought they were.[12]

There are two main defects with Nola's modified causal theory. The first is that it avowedly introduces a priori definitions into scientific inquiry, which is something that scientific practice has taught us to avoid (and it is not even clear how Nola is able to maintain that such statements are not analytic). The second is that it requires us to put elaborate reference-fixing definitions in the mouths of previous scientists for the account to work. Moreover, these definitions must be supplied with the benefit of hindsight for the account to give the correct answers. This also forces us to do violence to the practice of historical agents and to ascribe to them beliefs and intentions that they need not have had. It is not just a matter of supplying a rational reconstruction, for the point is that a variety of possible reconstructions are possible, not all of which will give the desired answer. If the

[12] Actually, it's not even clear that Nola can state this requirement sufficiently neutrally for the account to work. He claims that subsequent chemical theory unearthed the following belief: "Phlogiston's leaving metals does not cause these effects, i.e. calcination; oxygen joining metals does." (1980, 523) What does 'phlogiston' refer to in this claim? It surely does not refer to what was picked out by earlier chemical theory, for that would make the claim self-contradictory given the way 'phlogiston' was defined according to him.

reference-fixing definition of 'phlogiston' is specified differently (for example, 'whatever causes the effects of calcination when metals are heated'), it might come out to refer to oxygen. We have no concrete evidence to indicate that the reference-fixing definition (even assuming there was one) was specified in one of these ways and not the other.

(d) *Kitcher's account:* Yet another attempt to adapt the causal theory to deal with the problem of reference change has been made by Philip Kitcher. Kitcher also discusses the historical example of the phlogiston theory. According to him, as the theory became more sophisticated, particularly with Joseph Priestley, the reference of at least one of the key terms changed. Kitcher proposes that different tokens of the term 'dephlogisticated air' referred differently after a certain point in the development of the theory:

> Priestley's early utterances of "dephlogisticated air" were initiated by an event in which Stahl specified phlogiston as the substance emitted in combustion. After Priestley had isolated oxygen and misidentified it, things changed. His later utterances could be initiated by the event in which Stahl fixed the referent of "phlogiston" or by events of a quite different sort, to wit, encounters with oxygen. (1978, 537)

Thus, Kitcher rules that <u>all</u> early uses of the term 'dephlogisticated air' failed to refer, while <u>some</u> later ones referred to oxygen (and the rest continued to fail to refer).

One can compare Kitcher's modification of the causal theory with Berger's: while the latter takes a re-focusing to determine the referent of a term from thence onwards, the former allows that a re-baptism would only give the referent for some tokens of a term type, the referents of the other tokens still being determined by the earlier baptism. (Compare this also to Salmon's proposal that the term 'jade' designates jadeite with respect to some contexts, and nephrite with respect to others.) Because he opts for re-baptisms rather than rigidified descriptions, Kitcher's problems stem from his inability to show that such events are explicitly intended to determine reference by the agents involved. He needs to show that on each occasion Priestley had an intention to use the term 'dephlogisticated air' either as it originated with Stahl's baptism or as it was re-identified during his own later encounters. But it is unlikely that he had exactly one of these intentions every time he used the term. That is because it is unlikely that he considered both events to be baptisms, since he thought that a <u>single</u> substance was involved on all occasions.

In order to uphold the claim that scientific terms are rigid designators, there are two ways of determining reference: either through a chain of transmission leading to a baptismal event or by means of a rigidified description. This then raises the question of how reference can change, a question to which the authors examined supply various

answers. If reference is determined by an initial baptism, it can change by a re-baptism (Salmon1, Kitcher). If reference is determined by a rigidified description, it can change with the occurrence of a baptismal event (Salmon2, Berger), or by being replaced with a new rigidified description (Nola). This gives rise to three different ways of accounting for reference change. Moreover, having arranged for reference to change in one of these ways, one can allow subsequent occurrences of a term either to refer to what is picked out by each of the two determinants (baptism or description), depending on context (Salmon1, Kitcher), or else allow it to refer exclusively to the more recently introduced determinant (Salmon2, Berger, Nola).

There are a number of problems with these accounts. First, rigidified descriptions, although they can be made to provide the right answers (at least during a limited time period), introduce analytic or a priori definitions into the practice of science. Second, to make the accounts deliver plausible answers, the baptismal or re-baptismal events selected by these philosophers are not typically ones considered as such by the agents involved (that is why Berger talks about "focusing" rather than baptism). They are rather chosen in retrospect and seemingly without a principled criterion. Third, when it comes to subsequent occurrences of a term, there also does not seem to be a principled way of telling which occurrence or utterance of a term refers to which referent. It again involves imputing intentions to the historical agents in the absence of sufficient evidence in that regard. The problem common to the second and third objections is that intentions need to be supplied in an ad hoc fashion to historical agents in order to make the theory supply answers that acccord with our current intuitive judgments as to what they were referring to. These intentions do not necessarily accord with the agents' own predilections but are supplied in such a way as to conform to our present intuitions and may be supplied in a variety of ways. Kripke, in trying to account for cases of reference shift, makes clear that it is the agents' intentions which should determine the change of reference: "a present intention to refer to a given entity... overrides the original intention to preserve reference in the historical chain of transmission." (1980, 163) The problem with that orthodox strategy is that the agents involved do not always have a clear intention to refer to a present entity, perhaps because they do not realize that it is different from the entity at the end of the historical chain. But the attempts at remedying it, which try to account for reference change, are ad hoc and beg too many questions.

2.4. Reference and Belief

The above diagnosis suggests that the causal theory is unable to account for reference change in science because of its detachment of reference from the scientists' conceptions of

things, that is their beliefs, intentions, and other mental states. I will argue in this section that the detachment of reference from belief also renders the causal theory incapable of effecting the comparison of theories or corpora of beliefs. It may be thought that supplying a criterion for coreference automatically guarantees the comparability of theoretical tenets, but this is not actually the case.

Recall Salmon's point about the severance of what he calls the "conceptual representation" associated with a term from the means of determining the referent of that term. In the case of scientific terms, one of the prime motives for this move might seem to be to ensure the comparability of scientific theories. But how exactly does it help to say that reference is shared between two theories if conceptual representations are not? The causal theorist might say that this is a way of making good on the familiar distinction between meaning change and theory change. To see how, simply identify change of meaning with change of reference and change of theory with change of conceptual representation. Meanings can now be said to be shared where theories are different because meaning is made independent of theory. This is the move that seems to provide much of the motivation for using the causal theory as an account of how scientific terms refer. The main objection to this move is that it allows for the possibility that, in some cases at least, theories cannot be compared, although meanings are ruled to be shared. The objection can be illustrated with the aid of the notorious "puzzle about belief" first described by Kripke.[13] In Kripke's story, a Parisian youngster named Pierre is told by his nanny that 'Londres est jolie' and acquires the belief expressed by that French sentence. Later on in life, he travels to London, learns English, and comes to have the belief that 'London is ugly,' but without abandoning his previous belief. Since he does not know that 'Londres' and 'London' refer to one and the same city, he does not knowingly contradict himself.

As Donnellan has observed, Kripke focused in the original statement of the puzzle on linguistic avowals of belief. Instead, Donnellan writes that he is "inclined to protest that the puzzle really is a puzzle about belief, about the psychological state itself..." (1989, 275) But he neglects to emphasize that it is wide (or broad) beliefs that give rise to the puzzle rather than narrow ones. According to a now classic philosophical distinction, wide beliefs are individuated according to their external causes rather than an agent's conception of things. That is why Pierre can come to believe both that 'Londres est jolie' and that 'London is ugly' without noticing that he is contradicting himself. For the causal theorists,

[13] For details, see Kripke (1979).

'London' and 'Londres' refer to the same city in Pierre's idiolect regardless of his conception of things because they have the same external cause or historical origin. This may be a plausible way of ascribing Pierre's beliefs about the referent of the proper name 'London', but it leads to problems if the causal theory of reference is to be used to ensure comparability among scientific theories.

I will now tell a brief story based on the Pierre case which is meant to show why the causal theory of reference does not enable scientists to compare their theories. Consider the case of Paulette, Pierre's teenage sister, who comes to believe the tenets of Bohr's theory of the atom in the way posited by the causal theory. She is taught its principles in the classroom, reads textbooks, and sits for examinations. Naturally, she comes to have a number of beliefs about 'l'électron'. The following year, she moves to London where she enrolls in a university course in physics and is immediately taught some elementary quantum mechanics. After a very short time, she acquires the term 'electron' in a lecture about the Schrödinger equation for the hydrogen atom. Being equipped with the right intentions, she proceeds to refer successfully using the term 'electron' (according to the causal theory). But she fails to connect the two theories and assumes that the two terms apply to different particles or pertain to different realms. Just as Pierre does not notice that 'London' and 'Londres' are translations of one another, Paulette is in the dark about 'l'électron' and 'electron'. It is evident that by adopting the causal theory, one is forced to admit that an agent can unknowingly acquire contradictory beliefs involving what the theory considers to be coreferential terms.

The causal theory allows reference to float free of an agent's conceptions of things and makes it possible for agents to acquire contradictory (wide) beliefs in a completely rational and standard manner. The agents in question have no grounds for coming to believe that the referents of the terms they use are the same, because that would involve tracing back certain historical chains to their origins (again, according to the causal theory). Since that is something rational inquirers are not always in a position to do, they will sometimes end up not being able to compare two such theories or sets of beliefs and determine that they are contradictory. This demonstrates the inappropriateness of the causal account when it comes to dealing with the problem of theory comparison in science. Theories are usually considered to be comparable precisely <u>because</u> meanings or references are shared, so it is of no help if the meaning or reference of two terms is the same but the theories in which they are embedded are not comparable by the agents involved. The fact that the causal theorists required an omniscient observer and a criterion of consubstantiality to determine reference in general should have alerted us to the fact that their account of reference was not generally useful for inquiring agents seeking to

compare two theories.

It may be claimed that the causal theorists have not been given a fair chance to give a criterion of coreference that would enable the comparison of theories. Many of them advocate two separate components of meaning or two modes of belief ascription (for example, Putnam talks about reference and "stereotype"). The notion of meaning (or reference) discussed so far concerns only the wide mode of ascription. Thus, the causal theorists might urge that narrow content be pressed into service at this point to resolve the problem of comparing scientific theories and of theory-choice. However, that would undermine the initial intention, since wide content was supposed to be the mode of ascription that held out the greatest promise of dealing with the problem of comparing theories. To say that the wide mode of ascription ensures that terms from two different scientific theories will have the same reference but that the narrow mode of ascription will be needed in order to compare those same theories, is to concede that the causal theory of reference is not the proper route to comparing theories. It is to say that some other way of comparing scientific theories should be sought.

There is another possible response to the objection raised against the causal theory. A defender of the account might say that the point is surely not that scientists themselves must always be able to compare the theories that they hold, but that the theories held should generally be comparable in retrospect. Indeed, it may be said, Paulette is not a good example of a scientific inquirer and is insufficiently informed and versed in both theories, which explains why she fails to communicate with her former self in this particular case and to effect a comparison of the two theories. As long as the causal theory gives us a general way of determining the reference of terms drawn from different scientific theories, it can be considered to have delivered on its promise of telling us what scientists operating with different theories are referring to. It may be conceded, the response continues, that some of the conditions involved are only satisfiable in principle and not in practice, but that is the kind of philosophical idealization that is necessary in giving a perfectly general account.

This response is unsatisfactory. Although the above example involves an imperfect agent, it demonstrates a broader point: the causal theory requires scientists to be in possession of information that is not generally available to them in order to compare their theories. Therefore, it does not generally enable inquiring agents to make a choice based on a direct comparison of two theories. It implies that proponents of one scientific theory might be in the position of talking about the same things as their rivals, while being incapable of isolating their agreements and disagreements with those rivals. Moreover, their lack of communication might not be due to any shortcoming on their part, since the

causal theory makes the ability to determine coreference generally contingent upon acquiring information about the intentions of other agents and the nature of the substance at the initial baptism. This information may be forever irrecoverable, and will be attendant on the results of the ultimate scientific theory. Indeed, the problem with the causal theory of reference is that it makes it in principle impossible for scientific agents in the process of inquiry to compare their theories. That is because it requires that they be in possession of the ultimate scientific theory (with its criterion of consubtantiality and so on) in order to effect a comparison at all. The causal theory requires scientists to wait upon the end of inquiry before comparing their respective theories, but presumably, they will not have reached the end of inquiry if they are comparing scientific theories in order to make a choice among them. Such a choice cannot be made, according to the causal theorists, if they are not already in possession of the ultimate scientific theory. That is the price paid by the causal theorists for detaching the reference of a scientist's terms from the content of the scientific theory from which those terms derive.

2.5. Ostension and Scientific Taxonomy

Now I will introduce another type of difficulty that emerges when one tries to use the causal theory of reference as an account of scientific terms. It involves a certain presupposition about scientific taxonomy and ostensive or perceptual identification which, I will claim, is crucial to the account given by the causal theory. Since this presupposition is not in agreement with scientific practice and since it is not clear how the causal theory can do without it, its account of the reference of scientific terms is compromised further.

Recall that according to the causal theory, an agent uses a term to refer to whatever was featured at the baptism at which that term was first introduced. If the baptism involved an individual, then the agent always refers to the same individual; if the baptism featured an exemplar of a certain substance, then the agent refers to the same substance (as this would be determined at the end of inquiry by the ultimate scientific theory). This account has a problem in determining the referents of scientific terms when the exemplar at the baptism falls under more than one scientific category, and none of them has a privileged causal relation with the baptizer. In such a case, the causal relation is no longer capable of playing the desired role of singling out the referent. Kitcher has mentioned a case of this sort, in which he imagines someone baptizing a tiger. But he notes that the organism at the baptism does not exemplify a single kind, so we cannot assume that the new name applies to the category of tigers rather than the category of quadrupeds, carnivores, mammals, or vertebrates. (1982, 341-2) The problem for the causal theorists lies in the fact that they implicitly rely on some way of being able to tell in principle what

the natural kind featured at the baptism was and on a relation that would determine, for any exemplar, whether it was of the same natural kind as the exemplar at the baptism. A problem will arise, even for the omniscient observer, when the exemplar involved is a member of two or more natural kinds or substances.[14] In such a situation, no observer, omniscient or otherwise, is capable of determining which one of them is being singled out without further clues. The simple act of ostension accompanied by the demonstrative 'this' is hardly sufficient to single out a referent in such cases. As Montaigne points out in the epigraph to this chapter, momentous quarrels have broken out over the meaning of that single syllable 'hoc'.

Kitcher's way out of the problem involves bringing in the intentions of the baptizer. He says that given the introducing intentions, the speaker would be disposed to withhold the term 'tiger' from many quadrupeds, carnivores, mammals and vertebrates. He therefore concludes: "The referent of the new term is the kind that best fits the dispositions to verbal behavior." (1982, 342) In Kitcher's proposed solution, the speaker's linguistic dispositions--presumably including the descriptions that the speaker associates with that particular term and would apply in the appropriate circumstances--are not merely needed as a reference-fixing device. In such cases, the historical chain and the baptism cannot be used to disambiguate the referent, so the linguistic dispositions or associated descriptions take on a reference-determining role. But since all the speaker's associated descriptions may be false according to the causal theory, it cannot tolerate such a role. Moreover, this requirement is not just an empty one, for it concerns the central claim of rigidity. Kitcher's solution effectively reverses the claim that the term 'tiger' is a rigid designator, for the associated descriptions or speaker's dispositions may just be mistaken when applied to another possible world. Therefore, the causal theory cannot be used to ground reference when a certain natural kind is not uniquely featured at the baptism.

I have argued elsewhere that examples of crosscutting categories are rife in science.[15] That is, two scientific categories can generally overlap without it being the case that one is wholly subsumed by the other. Rather than a bedrock of non-overlapping

---

[14] Papineau seems to have been the first to notice this problem with the causal theory of reference when applied to science. As he puts it, even if an initial baptism need not tie any specific descriptive criteria to the term that is being introduced, what does need to be aired is what type of thing is being named. (1979, 158) It is sometimes referred to as the 'qua' problem in the literature.

[15] See Khalidi (1993a) and (1998a), for a more complete defense of this position. See also Chapter 7 for further elaboration.

categories, or even a hierarchy of ever-widening categories, this implies that science involves a patchwork of partly overlapping classifications. If this view of scientific taxonomy is supported by the evidence from scientific practice, the causal theory of reference seems to make a false presupposition about the way in which scientists carve up the world. Kitcher's hypothetical example indicates that the causal theory needs a bedrock of non-overlapping categories in order for its referential apparatus to get off the ground, since the causal connection at the baptism with a single scientific exemplar or a set of exemplars is crucial in securing reference to that category. Therefore, the theory cannot be used as an account of the reference of scientific terms. More generally, I would claim that the proliferation of crosscutting scientific categories reveals a deeper reason for the inadmissibility of purely ostensive procedures to anchor the meaning of scientific terms. Disambiguation is always required in order to determine which of a plethora of superimposed categories is being singled out--and that cannot be done by ostension alone.

In addition to crosscutting categories, two other phenomena point to the inadequacy of ostension and the difficulty of grounding the reference of a scientific term in a baptismal event. There are a number of recherché examples of scientific categories that are coextensive but nevertheless distinct (and ones that are more realistic than the stock philosophical ones, 'creature with a heart' and 'creature with a kidney'). These come chiefly from the Linnaean taxonomy in which a species is often the only member of its genus, or a genus the only member of its family, and so on. In the case of the kiwi bird, the genus is the only representative of its family, which is in turn the only representative of its order: Apteryx is the only category in the family Apterygidae, which is the only category in the order Apterygiformes. A group of kiwis from different species could variously be taken to stand in for the genus or for one of the higher taxa to which they belong. Mere ostension or the proffering of an exemplary kiwi will not tell us which taxon is being singled out.[16]

Yet another type of case brings out the difficulty with the use of baptismal or ostensive events in determining the reference of scientific terms. This relates to the fact that scientific terms are sometimes introduced to stand for theoretically posited entities well before we can produce a concrete manifestation of those entities. One of the better known examples from the history of physics is the discovery of the positron. This particle was first postulated by Dirac in 1931, after it was predicted by his relativistic theory of the electron. The following year, Anderson established the existence of the anti-particle of the electron while studying photographs of cosmic rays. A similar situation obtains with

[16] This example is taken from Sklar (1964).

purported observations of black holes, and would arise in the future were gravitons to be detected by experimental methods. Nor are such phenomena confined to basic physics. A number of chemical elements were prepared experimentally only after their existence was surmised from gaps in the periodic table. Mendeleyev predicted the existence of three elements that were later identified in the laboratory and dubbed, scandium, gallium, and germanium. Similarly, Bohr predicted that the element with atomic number 72 (later named hafnium) would resemble zirconium, a prediction that was eventually confirmed when the element was actually manipulated. Furthermore, in paleontology, it is not uncommon to posit the existence of a long extinct species, perhaps to serve as a common ancestor for two extant taxa that are held by theorists to be related. Later, fossil evidence is sometimes unearthed which is then tied to the previous theoretical posit. In all such cases, there are initially no concrete exemplars to be singled out at actual introducing ceremonies. Yet, we presumably want to credit theoreticians who make such predictions with successful reference to the relevant entities even before they are manipulated more directly. Indeed, we cannot even credit theorists like Dirac, Mendeleyev, and Bohr with these predictions if we do not take them to be referring to the very same things that later surfaced in the lab.

2.6. Theory Independence

If the causal theory is unsuitable as an account of the reference of scientific terms, it would seem as if some other account is needed of the meaning or reference of these terms. But before trying to develop such an account, we should take note of the attractions of the causal theory and take care to preserve its desirable features.

The advantage of the causal theory is sometimes thought to be that it allows the beliefs associated with a scientific term to change while leaving reference constant. While this is indeed desirable in a theory of reference or meaning for scientific terms, it should be distinguished from a seemingly related feature: the ability to refer without harboring any (or at any rate very few) beliefs about the referent. The second feature can be illustrated by the following story from Boyd:

> There simply is nothing I did which was acquiring linguistic competence (or, at any rate, referential competence) with respect to the term 'black hole' except learn that the expression was in my language. (Consider: Tonight I read the headline 'Ultra-dwarf Mezars Discovered,' then I am already linguistically competent with respect to the term 'ultra-dwarf mezar' automatically, in virtue of the social and intellectual skills that I already have.) (1979, 390-1)

This is supposed to be a case in which the speaker comes to have linguistic competence

with respect to a term, or can be said to refer successfully using a term, without having any beliefs about its referent. Similar examples have been used to illustrate the causal theory's account of the reference of proper names. One derives from Gareth Evans:

> A group of people are having a conversation in a pub, about a certain Louis of whom S has never heard before. S becomes interested and asks: 'What did Louis do then?' There seems to be no question but that S denotes a particular man and asks about him. (1973, 198)

The intuition behind these stories is shared by many, but rather than play the intuition game, it may be worth asking whether such cases ever crop up in the course of scientific inquiry. These cases are characterized by the fact that the agent acquires a term with scarcely any of the usual accompanying beliefs.[17] But there are rarely instances of belief-independent or theory-free acquisition in science, for the scientists' associated beliefs are seldom non-existent. Even at the very earliest stage of inquiry a scientist will usually have some beliefs or a proto-theory associated with a new term. Although there are cases in which many beliefs turn out to be false, there are few cases where beliefs are simply absent. It is not a coincidence that the example Boyd uses involves the acquisition of a term by a layperson from a newspaper headline. Even though the term itself derives from science, that does not make the example one of scientific reference. The fact that the causal theory sanctions theory-free acquisition may be an advantage in explaining how non-experts refer or acquire linguistic competence, but it is not a plus when it comes to the scientific experts. It might also be said that the intuition behind saying that the newspaper reader refers rests partly on the fact that <u>other</u> language users have some (true) beliefs or a full-fledged theory about ultra-dwarf mezars. But the reliance of laypersons on experts or on other members of the linguistic community (what Putnam calls the "division of linguistic labor") is not the primary focus if our subject is the reference of the experts themselves.[18]

Therefore, a theory-free or belief-independent account is not an advantage in explaining the reference or meaning of scientific terms. However, any account of the reference or meaning of scientific terms would do well to retain another feature, namely

---

[17] The extent to which Evans' example illustrates the same point as Boyd's depends on just how much of the conversation the speaker has overheard. Moreover, our willingness to ascribe successful reference in such cases (and often the speaker's own willingness to use the proper name involved) seems highly dependent on social context.

[18] I will discuss the relation between expert concepts and lay concepts in section 6.4.

that some tenets in which a term features could turn out to be false, leaving reference or meaning constant. Or to put it differently, that two scientists could share reference or meaning while disagreeing considerably in their relevant beliefs. That much, at least, should have been learned from the causal theory. I would argue that an account of the reference or meaning of scientific terms that satisfies this desideratum can be developed without the causal theory of reference and within a generally descriptivist theory, so that will be the topic of most of the remainder of this book.

Chapter 3: **Interpretation**

> There is a country... whose inhabitants have ways of thinking, in many things, particularly in morals, diametrically opposite to ours. When I came among them, I found that I must submit to double pains; first to learn the meaning of the terms in their language, and then to know the import of those terms, and the praise or blame attached to them.
>
> David Hume, A Dialogue

3.1. Descriptivism with Holism

The previous chapter examined a prominent account of reference and questioned its suitability for giving the semantic value of scientific terms, and therefore for comparing scientific theories. I argued, in part, that the shortcomings of that account stem from its disregard of beliefs, that is, from its disregard of the theories from which scientific terms are taken. We would seem to have reached an impasse, since the attempts covered in Chapter 1 gave due regard to the content of the relevant scientific theories, but they too were found unsatisfactory, chiefly because of the threat of incommensurability. In this chapter, however, I will propose a different approach, one that is belief- or theory-based but locates the procedure of comparing theories in the enterprise of translating or interpreting an alien system of beliefs. That is what makes it different from the accounts described in Chapter 1.

In a short dialogue on the subject of moral relativism, Hume once described an imaginary journey to a foreign land where the native inhabitants were supposed to have a radically different system of morality which he finds difficult to interpret. Hume's predicament, described in the epigraph to this chapter, bears a striking resemblance to the thought-experiment of radical translation first described by Quine. In Word and Object, Quine posed the problem of understanding a group of alien speakers without knowing anything antecedently about the meanings of their words or the contents of their beliefs. The challenge faced by the radical translator is to solve for both variables at once. According to Quine, the aim of the translator is to emerge with a translation mapping the alien speakers' sentences on to sentences of the home language on the basis of hypotheses

about the translation of the alien terms. Davidson modified this by taking the procedure to yield a truth theory of the kind first introduced by Alfred Tarski. The modifications are well-known and will not be directly pertinent to the purpose at hand, which is to see how the Quinean-Davidsonian approach can shed light on the problem of comparing scientific theories.[1]

Despite the superficial similarity, Hume's conception of the situation seems to be at odds with an important aspect of Quine's and Davidson's. By saying that he must "first" learn the meaning of the terms and "then" the import of those terms, Hume suggests that the meanings and beliefs of the aliens can be ascertained separately. Such a view runs counter to the "inextricability thesis" regarding meaning and belief held by both Quine and Davidson.[2] That thesis states that discovering the meanings of an agent's terms is a process inseparable from discovering that agent's beliefs; one cannot do one without the other. The interpretive approach ascribes meanings on the basis of shared beliefs, making it a descriptional theory. The beliefs themselves have been assigned on the basis of shared meanings, making the process holistic. The meanings of terms associated with different scientific theories will come out either the same or different depending on the agreements and disagreements that exist among the tenets of the two theories.

To some minds, this preliminary characterization might conjure up the specter of cluster theories of meaning or reference, according to which the meaning of a scientific term is given by a cluster of beliefs or theoretical tenets. Perhaps the most explicit cluster theory for scientific concepts was elaborated in some early work by Putnam, where he

---

[1] Davidson's modifications serve to introduce a well-known distinction between "translation" and "interpretation". While Davidson's points in this regard are well taken, it will not deter me from talking often about translation rather than interpretation. That should not be taken to imply that I consider translation a purely syntactic matter of matching up terms, a process which does not require genuine understanding.

[2] The expression "inextricability thesis" was originally used by Michael Dummett (1974, 387-8) to refer to Quine's version of the thesis.

argued that the concepts of science were "law-cluster concepts".[3] Though Putnam's early theory and the interpretive approach are both descriptional, the latter represents a significant departure from such cluster theories, mainly in its being holistic. The arguments against cluster theories rightly find fault with the fact that each particular term in a theory is linked (more or less determinately) to some subset of tenets of the theory involved. By contrast, the interpretive approach does not sanction the antecedent specification of such a set of tenets to be associated with each term. The terms of one theory are matched up with, or translated into, those of another on the strength of our conjectures about the agreements and disagreements that exist among their sentences. The matches will suggest further grounds for agreement, which may then force a reconsideration of some of the translations already made. At the end of the process, certain translations have been made which enable us to say that certain meanings are shared. Since the process is holistic, there is no question of a unique cluster of theoretical tenets being isolated in advance as definitional (even tenuously) of any term. Not only does the interpretive approach eschew strict definitions, it does not even subscribe to a "weighted" cluster theory or some such modified proposal.

Putnam's law-cluster scientific concepts in science were constituted not by a bundle of properties but by a cluster of laws that determined the identity of the scientific concept, though in a more or less loose manner. He explained it thus: "In the case of a law-cluster term such as 'energy', any one law, even a law that was felt to be definitional or stipulative in character, can be abandoned, and we feel that the identity of the concept has, in a certain respect, remained." (1962, 53) Putnam did not specify the laws that were supposed to be central to the meaning or identity of the concept of kinetic energy; he said merely that the principle 'e = 1/2 mv2' was not one of them. That is clear in retrospect, now that the special theory of relativity has rejected this theoretical tenet while retaining the concept of

---

[3] This is not Putnam's current view. In fact, in more recent work, he states that he holds that "there is no criterion for sameness of meaning except actual interpretative practice--a view made famous by Quine and Davidson." (1988, xiii) A more fully developed version of the cluster theory of reference for scientific terms is elaborated in Smith (1981), especially Chapter 4.

kinetic energy. But science has shown us the difficulty of specifying what should and should not go into the cluster in advance, thus making it unlikely that a cluster theory is an adequate account of the meaning of scientific terms. Instead of attempting to supply a definitional cluster for such terms, the approach being outlined here specifies certain constraints on the process of interpretation which will yield an optimal way of matching up the terms of different theories, and therefore, an optimal way of comparing theories. There will be interpretive decisions to be made along the way which are complicated by the impossibility of supplying indefeasible definitions for scientific concepts. If no set of beliefs is definitional for any given concept, that makes the decision to ascribe a concept more difficult. But it is by no means impossible; the strategy is to find the best overall fit between an alien theory and our own, given the constraints and the evidence.

Some of the most salient features of the procedure of interpretation will be examined in this chapter in order to establish its credentials as a method of comparing, not just whole languages, but scientific theories. Later, in Chapter 5, the method will be supplemented with certain interpretive principles that will help to resolve the specific problems associated with comparing scientific theories. The overall aim can be stated indifferently as an attempt to propose a method whereby scientific theories can be compared, or as an attempt at giving a theory of meaning for scientific terms. But first, the interpretive approach will be defended as a means of defeating the incommensurability thesis, which serves as the sword of Damocles for any account of theory change in science.

3.2. Conceptual Schemes

Davidson's argument against incommensurability focuses on the claim that an alien language (or total theory of the world, or conceptual scheme) could be mostly true but not translatable into our own. Davidson confesses a temptation to "take a very short line indeed" with this claim. He suggests that it is refuted almost directly when one observes that "translatability into a familiar tongue [is] a criterion of languagehood." (1974a, 186) In other words, if translatability just is part of what it is for something to be a language (theory, scheme), then there is no room for claiming that a certain phenomenon is a language and at the same time maintaining that it is untranslatable. In effect, Davidson is asking the maker of the original claim: How could it be a language (theory, scheme) if it is

in principle untranslatable? To be understandable or translatable by another is just part of what it is for something to be a language.[4]

The defender of incommensurability will not be convinced so easily, though, and Davidson considers the following response. We do indeed have another criterion for what it is to be a language (theory, scheme): a language is something that organizes the world, or something that fits reality. That is, an untranslatable language is a scheme that organizes the world differently from our own scheme, or one that fits reality in an alternative manner. In replying to this claim, Davidson explores various ways of making these metaphors (organizing, fitting, and so on) more philosophically respectable. On closer inspection, he concludes: "Our attempt to characterize languages or conceptual schemes in terms of the notion of fitting some entity has come down, then, to the simple thought that something is an acceptable conceptual scheme or theory if it is [largely] true." (1974a, 194) He goes on to argue that our only handle on a scheme being largely true is through translation, since we cannot "divorce the notion of truth from that of translation." (1974a, 195) In other words, there is no such thing as a scheme for organizing reality which is largely true but untranslatable. Indeed, since all such schemes are inter-translatable, it makes it vacuous for us to talk about alternative schemes at all, making the very idea of a conceptual scheme questionable in itself.

Davidson concludes that if something is genuinely a language (theory, scheme), then it is translatable. Claims of incommensurability, or of a language that is untranslatable in principle, are therefore unfounded. But that is not all there is to it, for the opposition can respond by proposing a different version of incommensurability. The response might go as

---

[4] It is worth adding here that Davidson's claim that translatability is criterial for languagehood need not be taken to commit him to the existence of a definitional or conceptual truth (i.e. that a language is by definition something that is translatable). Rather, the import of the claim might be that, given our best philosophical understanding of what a language is, translatability is one of its permanent characteristics. We could change our minds in the course of our philosophical and other inquiries, of course, but the change in what we consider a permanent characteristic would need to be backed up by an argument or theory.

follows. Davidson may have a point about the impossibility of encountering a wholly and essentially untranslatable system which is, for all that, a genuine language. Perhaps this extreme scenario is indeed incoherent. But, surely, there are degrees. There might yet be a conceptual scheme that is untranslatable in part, whole sections of which resist our best efforts to render it in our own terms. It might be added here that this should surely be our main concern if we are dealing with the possibility of incommensurable scientific theories. For individual scientific theories do not correspond to entire conceptual schemes or theories of the world; at best, a scientific theory constitutes a part of a conceptual scheme.[5]

Davidson seems to have two lines of defense against this modified challenge of partial incommensurability, which I will try to explicate on his behalf. In so doing, I will adapt Davidson's argument so that it serves my purposes, without distorting his basic philosophical position. First, whenever we encounter a term in the language in question that we cannot render in our own terms, we can always resort to a neologism. Here, Davidson might lean on Tarski's claim of the "universality" of human language. As Tarski explains it:

> The common language is universal and is intended to be so. It is supposed to provide adequate facilities for expressing everything that can be expressed at all, in any language whatsoever; it is continually expanding to satisfy this requirement.[6] (1969, 67)

---

[5] To be sure, Feyerabend argues at some points--though not always--for the possibility of global incommensurability. But that claim rests on an extreme holist theory of meaning, as argued in section 1.4., and extreme holism will be countered in section 3.5.

[6] Tarski thinks that this is a feature of "colloquial language" rather than what he calls "scientific language". However, the language of science is being treated here as a part of natural language, rather than a regimented, formal language. Elsewhere, Tarski writes: "A characteristic feature of colloquial language (in contrast to various scientific languages) is its universality. It would not be in harmony with the spirit of this language if in some other language a word occurred which could not be translated into it; it could be claimed that 'if we can speak meaningfully about anything at all, we can also speak about it in colloquial

Any language worth its salt must contain a vocabulary by which it is potentially possible to describe the whole of experience, not a mere fragment of it. Of course, that is not to say that a language might not lack a term for a concept, but that situation can easily be remedied. When one language does not have a term corresponding to a term in another language, it compensates by resorting to a neologism, and every neologism must be introduced by connecting it to the old terms if it is to be of any use. The new terms will be inserted into our language in sentences that are couched in the pre-existing terms, and this will help us to see how they relate to the other terms in the theory. The fact that we supplement our vocabulary shows that neologizing is often the most efficient way to translate other languages. Still, these differences between languages (theories, schemes) should not lend themselves to overblown claims of incommensurability. The conceptual resources of all languages are the same, though their vocabularies may be different.

But what if these neologisms multiply to the point that they inhibit understanding? It depends what one means by this. Of course, there is no denying that in science, it is possible that a new theory might explore a wholly new subject matter, as occurs, say, with the introduction of a new scientific discipline. This type of innovation will inevitably carry with it a slew of new concepts which cannot be translated into our pre-existing terms. But then there is no point in trying to translate the terms of a new discipline into those of the scientific theories that preceded it sans the new discipline, because they involve a entirely different subject matter. No one would seriously expect a term-by-term translation to be possible in this case. Thus, there may well be instances in which neologisms need to be multiplied, as occurs with the introduction of a new branch of science. But in such an eventuality, there should be no serious impulse to translate the new science into the terms of the old, since we have obviously changed the subject.[7]

---

language'." (1956, 164) He relates this feature of language to the generation of the semantic paradoxes, but that is not of concern here.

[7] In section 7.2., I will say more about how to distinguish a change of subject matter from a change of theory concerning the same subject matter.

However, if one does not have in mind a subject-changing innovation of this type, then Davidson has a different kind of defense in ruling out the possibility of partial incommensurability. He would say that before even resorting to the option of coining a new term to translate a term from another language, we generally have the ability to take up the slack between two languages in terms of beliefs rather than in terms of meanings. As he puts it: "no clear line between the cases can be made out." (1974a, 197) The inextricability of meaning and belief implies that it is in principle possible to reinterpret a purported difference in meaning or concept as a mere difference in belief or theory, and thereby avoid altogether the need for introducing new terms into our vocabulary. On Davidson's interpretivist theory of meaning, the resort to a neologism is always in principle avoidable, thanks to the inextricability of meaning and belief. There is therefore no need to fear the possibility of the widespread neologizing that would lead to partial incommensurability.

These two lines of defense seem efficacious against an alleged situation of partial incommensurability among two languages (theories, schemes). Conceptual differences can always be accommodated by the introduction of neologisms. Moreover, neologisms will only be a last resort, since for the most part we rely on our own terms, and we have the option of doing so because meaning and belief are inextricable. This enables us to reinterpret any given purported difference in meaning or concept rather as a difference in belief or theory, thereby avoiding a wholesale resort to neologisms and the ascription of massive conceptual differences or changes. Hence, not only is there no threat of a wholly untranslatable language or conceptual scheme, there is not even a serious chance of a partially untranslatable language, in the sense alleged by defenders of incommensurability.

When the assumptions behind Davidson's argument are spelled out, it can be used as the basis of a method for comparing scientific theories. Since theories are embedded in languages, whenever we interpret them we are assuming a certain (homophonic) interpretation of the rest of the language to which they belong. Implicit in any attempt at translating a theory is the translation of a whole language. In comparing scientific theories, we are typically interpreting small fragments of a total language or theory of the world, but Davidson's argument applies to partial theories just as well as whole theories of the world. Nevertheless, many philosophers, especially philosophers of science, have been dissatisfied

with Davidson's argument against incommensurability. Some have found Davidson's "short line" to be too short. Philip Kitcher's reaction is fairly typical: "Davidson's assurances that we will always be able to translate any alien language fail to show what is especially difficult about reconstructing the languages used by past scientists, and how the difficulties can be overcome." (1978, 546n) As already noted, there are indeed specific problems of interpretation that are not addressed by Davidson's argument. Far from denying that they exist, I will consider the problems associated with comparing particular scientific theories in more detail in the rest of this chapter as well as in the following two chapters.

3.3. Indeterminacy and Incommensurability

There is an obvious way of attacking Davidson's argument that demands further consideration and will help to introduce more specific problems of interpretation. The objection is that his argument fails to ensure uniqueness in interpretation, since he admits, following Quine, that indeterminacy is a fact of life.[8] Moreover, it is precisely the inextricability thesis, which enables us to take up the slack in our beliefs rather than our concepts, which is the main source of indeterminacy. The thesis of indeterminacy states that the interpretation of an alien language will generally not be unique, so a reply to the claim of incomparability that posits not one, but many, methods of comparison may be thought to be no reply at all. It may be argued that the comparability of scientific theories requires that translation be unique. Fidelity in translation might be compared to fidelity in love: neither tolerates the existence of alternatives.

Hacking has written on the difference between incommensurability and indeterminacy, arguing that they have quite opposite implications. He observes that indeterminacy and incommensurability pull in opposite directions: "Indeterminacy says there are too many translations between schemes, while incommensurability says there

---

[8] But Davidson thinks that the degree of indeterminacy is more limited than Quine does. Note that I will not get into the debate over the nature of indeterminacy, for example, as to whether it is anything over and above the under-determination of semantic theory by evidence.

are none at all." (1982, 59) Hacking is correct to observe that indeterminacy and incommensurability are different in terms of their specific claims. But when it comes to questions of theory-comparison and the issues of rationality, realism, and progress (which are the larger stakes in this debate), they seem to be related. Both draw attention to the threat of irrationality in the development of science, since the claim that there are multiple ways of understanding one theory in terms of its successor is as subversive of one's rational assumptions about science as the claim that there is no way. Davidson, it might be said, has defeated incommensurability while surrendering to its irrationalist ally.[9]

This objection to Davidson's approach to disarming incommensurability derives from the fear that there will be two significantly different but empirically adequate translations of an alien language or theory. If interpretation is always modulo indeterminacy, it might be said that there is nothing to prevent a massive misunderstanding from being perpetrated in the interpretive situation. This objection will be met in two steps. The first consists in showing that indeterminacy in general is not conducive to the same irrationalist fears that the incommensurability thesis has excited. The second step, which will be taken in the next chapter, will be to show how, in practice, the interpretive approach can be used to compare different scientific theories and how alternative interpretations can be ruled out.

To answer the charge, a closer look at indeterminacy is required. It seems safe to say that the kind of indeterminacy that is most relevant in this context is the one that

---

[9] Kuhn himself has hinted at the connection between indeterminacy and incommensurability. After quoting Quine's claim that two translations may accord with all the evidence and yet disagree in truth values assigned to sentences, he has commented: "One need not go that far to recognize that reference to translation only isolates but does not resolve the problems which have led Feyerabend and me to talk of incommensurability." (1970b, 268) In addition, Kuhn sometimes suggests that different translations get things partly right, but that none is superior to the rest; this is apparent in some of the passages quoted in section 1.5.

derives from the inextricability of meaning and belief.[10] Davidson describes this species of indeterminacy by saying that there could be a truth theory for a language that satisfies all relevant empirical constraints and makes a certain sentence true and another equally acceptable theory that does not make that sentence true. (1979, 228) Since a disagreement in meaning can be reformulated in principle as a disagreement in beliefs, the possibility arises of the existence of two empirically adequate truth theories for an alien language, which differ with respect to the truth value of a certain sentence or certain sentences. Davidson illustrates it thus:

> If you see a ketch sailing by and your companion says, 'Look at that handsome yawl', you may be faced with a problem of interpretation. One natural possibility is that your friend has mistaken a ketch for a yawl, and has formed a false belief. But if his vision is good and his line of sight favorable it is even more plausible that he does not use the word 'yawl' quite as you do, and has made no mistake at all about the position of the jigger on the passing yacht. (1974a, 196)

This example shows the way in which the inextricability of meaning and belief can lead to indeterminacy by providing the interpreter with competing interpretations. But it also demonstrates that the interpreter can adjudicate between those competitors. Here we are effectively faced with a choice between attributing a difference in belief or a difference in meaning. Davidson's companion may have formed a false perceptual belief about the jigger on the yacht and thereby mistaken a ketch for a yawl. Or else, he may use the term 'yawl' to mean what we mean by 'ketch'. Davidson imagines a situation in which the latter is the most plausible hypothesis. Since this kind of indeterminacy does not leave the truth values of all sentences intact, the holistic character of belief ascription enables the interpreter to rule out certain interpretations on the basis of one or more related beliefs also attributed to

---

[10] The indeterminacy of logical form is obviously not a particular concern in the attempt to compare rival scientific theories. As for the so-called inscrutability of reference, I will be ignoring it here since my focus is on comparing meaning or concepts, not reference.

the speaker. In this case, the evidence tips the scale in favor of attributing a difference in meaning rather than one in belief.[11]

Davidson stops short of saying how such decisions are to be made, remarking merely that, "when others think differently from us, no general principle, or appeal to evidence, can force us to decide that the difference lies in our beliefs rather than in our concepts." (1974a, 197) However, while there is no general rule, these decisions cannot be entirely arbitrary, since the distribution of the aliens' true beliefs will generally be affected by the difference between meaning change and theory change. This will provide the interpreter with an opportunity to rule in favor of one interpretation and against another by gathering more evidence. As long as there is a difference in the distribution of truth-values among the sentences of the theory to be interpreted, there will be evidence that can decide between the rival interpretations.

In the above example, the interpreter must decide whether to translate 'yawl' homophonically or to translate it as 'ketch'. If it is translated as 'yawl', the speaker has uttered a false sentence and if translated as 'ketch', the sentence is true. What enables the interpreter to decide between attributing a true or false belief to the speaker is a perceptual belief already attributed to the speaker about the position of the jigger on the yacht, a belief ascribed on the basis of other evidence. When it is brought to bear on the decision between ascribing a difference in theory and one in meaning, the verdict is decisive in favor of the latter. Davidson says nothing to assure the interpreter that all such choices will be settled so easily; in the example he cites there is one belief that is deemed most relevant and it is a perceptual one at that. These facts suggest that the decisions are easier to make than they often are, but the story serves to illustrate the general plan of action and the kinds of constraints that are in place.

In making sense of a scientific theory, interpretive choices of this kind will not always be as cut-and-dried as they are in Davidson's ketch-yawl example. Nevertheless,

---

[11] But notice that this difference in meaning does not necessarily imply a conceptual deficit, since the speaker may have both concepts, ketch and yawl, but be using different terms for them.

there will be similarly relevant considerations that help to eliminate alternatives and enable a comparison of the two theories. It may be difficult to decide between two or more runners-up, but considerations favoring one interpretation over another will not simply be lacking. In fact, this is suggested by the very nature of interpretation. The translation or interpretation of a theory is a mapping from the terms of that theory to the terms of the home theory. When viewed thus, it is clear that some kind of translation between two theories is always possible provided one is willing to tolerate an indefinite number of false sentences in the translated theory. The trick is therefore to come up with a mapping that metes out truth and falsity among the sentences of the theory to be translated in such a way that accords with the evidence collected by the interpreter. That is, we want a mapping that makes it possible to explain and predict (or retrodict) the linguistic and other behavior of the agents who hold the translated theory.

When he first proposed the method of radical translation, Quine specified four constraints on the translation function. These state that translation must preserve observation sentences and truth functions, as well as stimulus-analytic and stimulus-contradictory sentences. (1960, 68) But indeterminacy is supposed to survive the application of Quine's constraints, so the question arises whether other constraints can be adduced that will eliminate the indeterminacy. Putnam has charged Quine with a form of conventionalism for specifying four constraints on the translation function and failing to consider the existence of others. Rather than produce minimalist constraints and conclude that translation is indeterminate, an anti-conventionalist is under an obligation to look for additional constraints, according to Putnam. He describes the "conventionalist ploy" as follows: "Once a set of constraints has been postulated as determining the content of the notion in question... a proof is given that the constraints in question do not determine the extension of the notion in question." (1975c, 162) Ironically, he claims that Quine is guilty of giving what amounts to a meaning postulate for the notion of translation, even while inveighing against the logical empiricists for proliferating meaning postulates elsewhere. Putnam concludes not that indeterminacy is defeated by his argument, but that, "how much indeterminacy of translation there is, if there is indeterminacy of translation, is surely an empirical question." (1975c, 185) Something similar to Putnam's argument seems to be behind David Lewis' attitude towards these cases of purported indeterminacy. Rather than

express doubts about whether two interpretations of an agent could be empirically adequate and yet have satisfied all the constraints, Lewis turns the tables by saying that the existence of two such theories would just show that we have not uncovered all the constraints. As he puts it: "Credo: if ever you prove to me that all the constraints we have yet found could permit two perfect solutions,... then you will have proved that we have not yet found all the constraints." (1974, 343)

The points made by Putnam and Lewis immediately raise the question: What are the constraints? The most central and widely debated constraint or set of constraints surely have to do with the requirement for making the person being interpreted come out to be rational. Various interpretive maxims or principles have been advanced in this connection, notably a number of formulations of the Principle of Charity. In early writings on interpretation, Davidson held that the interpreter must aim to maximize agreement with the interpretee, in other words, to maximize the number of true sentences held by the interpretee (by the interpreter's lights). But this formulation has been widely criticized, not least because it is not clear how to make sense of "maximizing" truth when every person has a potentially infinite number of beliefs. More recently, Davidson has come to talk about "optimizing" (rather than maximizing) truth or agreement: "We make maximum sense of the words and thoughts of others when we interpret in a way that optimizes agreement (this includes room, as we said, for explicable error, i.e. differences of opinion)."[12] (1974a, 197) Similar principles have been advanced by other writers, notably Richard Grandy, whose "Principle of Humanity" calls on the interpreter to make the pattern of relations between the speaker's beliefs, desires and the world as similar to the interpreter's as possible. (1973, 443)

It is not clear how exactly such principles are related, or indeed whether they amount to substantive recommendations which can be followed by real-life interpreters. When it comes to the relationship between the Principles of Charity and Humanity, some

---

[12] He has also written: "The basic methodological precept is, therefore, that a good theory of interpretation maximizes agreement. Or, given that sentences are infinite in number, and given further considerations to come, a better word might be optimize." (1975, 169)

writers have argued that there is a fundamental difference between them, since the former is supposed to be "normative" (it specifies what it would be rational to believe in a particular situation) whereas the latter is "projective" (it specifies what the interpreter would have believed in that same situation). But other writers have argued that projective and normative principles are effectively equivalent, or that they come to the same thing in the end, notably Dennett (1987, 83-101, 342-4). As for the issue of whether they amount to substantive guidelines for actual interpreters, Lewis has noted that these rules are simply "the fundamental principles of our general theory of persons," which makes it seem unlikely that they can be strictly codified. (1974, 334) More decisively, John Haugeland has cast doubt on the possibility of coming up with a strictly formulated maxim of rationality or reasonableness. As he puts it: "'making reasonable sense' under an interpretation is not defined--and I doubt that it can be." (1978, 219)[13] But he also thinks that this is not a problem in real life, since "it is seldom hard to recognize in practice," and "explicit conditions can be stated for making sense about certain problem domains or subject matters..." (1978, 219)

I have nothing original to add to these observations, with which I am in broad sympathy. Belief ascriptions that find the holder of a scientific theory to be rational, consistent, and predictable, are generally to be preferred, other things being equal. It may be difficult to say in the abstract what this comes to, but as Haugeland implies, we are rarely at a loss as to how to proceed in practice. Typically, we will have some reason to assume at the outset that some beliefs will be shared and not others. We begin by singling out certain beliefs that we expect to agree on with our interpretee. Agreement on them will suggest that the terms featured in them should be translated uniformly. That will force translations of other sentences, some of which will be found true by our lights and others false. In case these interpreted beliefs have truth values that differ from our expectations,

---

[13] Compare what Haugeland says elsewhere: "... we are driven back to the more general but also somewhat fuzzier notion of 'making sense' as our criterion for the adequacy of interpretations. There is, to my knowledge, no satisfactory philosophical account of what it is to make sense; indeed, it is questionable whether a precise, explicit definition is even possible." (1981, 28)

given the agent's other actions and utterances, and assuming rational conduct, we might consider going back and revising the purported area of agreement identified initially. All the while, we apply general standards of rationality in deciding what makes most sense.

Where I do have something new to add is in proposing a further set of constraints in the form of a number of interpretive principles or maxims, which I will argue are especially suited to the enterprise of comparing scientific theories. These will not be fully spelled out until Chapter 5, where they will be justified with reference to the case studies to be explicated in Chapter 4. The whole idea of proposing interpretive constraints, perhaps with different constraints operating in different domains of discourse, is inspired partly by the positions of Putnam and Lewis on the interpretive enterprise, and their calls for piling on more constraints. It is also apparently not inimical to Davidson's original intention, for he has written: "I believe the range of acceptable theories of truth can be reduced to the point where all acceptable theories will yield T-sentences that we can treat as giving correct interpretations, by application of further reasonable and non-question-begging constraints. But the details must be reserved for another occasion." (1974b, 152) Davidson has not outlined such constraints elsewhere. But I would go further than he does here, for I would argue that the constraints to be specified (in the guise of a set of interpretive principles or maxims), will issue in an optimal interpretation. The most acceptable theory will therefore give the correct interpretation--and will not merely be capable of being <u>treated</u> as such, as Davidson would have it.

Now it may be protested that no set of interpretive principles, which act as constraints on interpretation, could guarantee that there will be a single optimal translation of one scientific theory in terms of another. After all, it may be said, no matter how stringent, the constraints might still underdetermine the outcome. At most, the interpretive principles will serve as a heuristic, not as an algorithm that will always issue in a unique solution. In response to such a challenge, I cannot claim to possess a foolproof guarantee that there will be a single solution to the interpretive problem in each case. However, given the power and diversity of the constraints to be encountered later in this work, it is unlikely that more than one interpretation could satisfy them equally. Moreover, it should be noted that one of the principles in particular, the Principle of Conceptual Charity, is designed to act as a tie-breaker amongst rival interpretations. Briefly, it enjoins

the interpreter to rule in favor of theoretical difference rather than conceptual difference, in case considerations seem equally weighted on both sides. It therefore provides the interpreter with a powerful device to rule out commonly encountered alternatives.

In addition to interpretive rules particularly suitable for making sense of scientific theories, there is another major constraining factor in the interpretation of scientific theories. That is due to the fact that the interpretation of scientific theories is a case of not-so-radical interpretation, since theories are not whole languages, but are embedded in languages. Methodological maxims such as the Principles of Charity and Humanity, which call (roughly) on us to find our subjects mostly correct or generally rational, are meant to apply over the totality of an agent's beliefs, not merely those within a particular domain, say a specific scientific theory. In the interpretation of a single scientific theory, truth may be compromised at the expense of optimizing it elsewhere in the agent's comprehensive theory. But the fact that these maxims apply over an agent's total theory of the world is still crucial, for it means that the interpreter of a scientific theory can assume many beliefs to be shared outside the contested domain. There will always be a preponderance of sentences of the total language held constant in particular scientific disputes. In the case of any scientific controversy or change in scientific theory, we usually hold constant beliefs taken from other disciplines or those that have different subject-matters. Other demarcation lines are sometimes drawn by observational beliefs on the one hand and meta-scientific or methodological beliefs on the other and they tend to furnish much of the shared agreement. Of course, these beliefs will not <u>always</u> be agreed upon, but there will usually be enough overlap to introduce an important constraining factor. The reluctance to revise these beliefs is such that an interpretation of one scientific theory in terms of another which would involve a drastic reinterpretation of other parts of the language is likely to be rejected out of hand. Weaker versions of such proposals are not unheard of, however. The debates surrounding quantum logic can be understood as being (partly) about the reinterpretation of classical physics in terms of quantum physics, specifically about whether it would be worth abandoning agreement on some logical principles in order to come out agreeing on some of the theoretical sentences. The restrictions imposed by the relative fixity of the borders of neighboring disciplines, as well as the observational coastline and the meta-scientific hinterland, provide a further constraint that helps to

remove the indeterminacy in the interpretation of any given scientific theory. These beliefs often serve as a testing ground for the process of comparison: since they will generally be shared, they can be counted on to help dismiss outlandish interpretations.[14] Incidentally, this also shows that the approach to theory-comparison being defended here does not depend on the feasibility of radical interpretation, since in any scientific dispute there will always be an area of agreement outside the disputed domain that both sets of theorists accept.[15]

In this section, I have not offered a knock-down argument against the indeterminacy of translation or interpretation. I began by agreeing with Putnam and Lewis that, in the face of indeterminacy, we have an obligation to reduce it without limit by application of appropriate constraints. I went on to suggest that we already have some resources to chip away at indeterminacy using constraints of rationality and the like. Then, I promised in subsequent pages (especially Chapter 5) to come up with principles which will enable us to chip away at it further with additional constraints which should deliver an optimal interpretation. Finally, I pointed out that the interpretation of scientific theories is mightily constrained by the usual need to retain a homophonic interpretation of the rest of the global theory in which those scientific theories are embedded. These considerations combined will ensure that indeterminacy is not a live threat to the possibility of interpreting scientific theories.

3.4. Neologisms, Ambiguity, and Nuance

---

[14] As we saw in section 1.4., even Feyerabend sometimes acknowledges that some scientific transitions (e.g. that from Newton to Einstein) have left most other parts of the language or conceptual scheme unaffected (e.g. procedures to estimate the size of eggs at the grocery store).

[15] Chomsky has denied that linguists ever actually practice radical interpretation and has cast doubt on the relevance of the thought-experiment for the study of language. (1992, 104-8). be that as it may, it should be clear by now that neither the actual occurrence nor feasibility of radical interpretation are required for this account of theory-comparison.

No account of the meaning of scientific terms can be expected to ensure that the terms of one theory are always capable of translation into the preexisting terms of the other theory, for neologisms will sometimes be needed to translate a rival or past scientific theory. The interpreter will typically be forced to neologize at certain points, as mentioned in section 3.2. Neologisms are obviously indispensable in the event that the interpreter runs out of terms to match up with those of the interpretee and must coin new ones to complete the process. It is impossible to say in advance when neologisms are required and how much neologizing is permissible, but neologisms should only be introduced as a last resort, when the slack cannot be taken up by the interpreter's own terms. Moreover, neologizing cannot proliferate across the board. That would go without saying were it not for the fact that there is sometimes a tendency to treat every difference in belief as a difference in concept, for which an altogether new term needs to be introduced.

Neologizing should be regarded as an innocuous feature of interpretation and the conditions for introducing neologisms will be discussed in section 5.7. But it has sometimes been taken as evidence that the comparison of theories cannot be effected by translation or interpretation. Some of the considerations Kuhn advances for incommensurability focus on the problematic nature of neologizing. He starts by pointing to the difference between translating a language into one's own and learning a new language. He then suggests that the existence of a set of interrelated neologisms in the supposed translation is an indication that what is afoot is language-learning rather than translation. And since incommensurability concerns the possibility of translation, the objection continues, it cannot be foiled by showing how one might learn the language of a new scientific theory. This is what lies behind the problem of "clusters of interdefined terms" which was identified in section 1.5. Kuhn uses the example of the chemical term 'phlogiston' to illustrate this point. He says that the term cannot be translated because of its relation to a number of other terms, like 'principle' and 'element'. "Together with 'phlogiston'," Kuhn explains, "they constitute an interrelated or interdefined set that must be acquired together, as a whole, before any of them can be used, applied to natural phenomena." (1983a, 676)

To begin to rebut this charge, note first that the sheer fact of neologizing cannot be grounds for defending the claim of incommensurability. That would effectively make it the

case that no two languages or theories are ever inter-translatable. For two languages to be capable of translation, they would have to have exactly the same number of terms (assuming no redundant terms and no term deficits) and these terms must be capable of being put in a one-to-one correspondence without neologisms. Clearly, there can be no practical point in drawing the boundaries of translation so narrowly that no real translation would count as a candidate. That cannot be all that incommensurability amounts to, for it would make it a trivial thesis applicable to any two natural languages or theories. Kuhn might say that it is not the sheer fact of neologizing, but the existence of an interrelated set of neologisms from an alien theory that constitutes grounds for incommensurability. This charge cannot be examined closely without looking at specific examples that purportedly exhibit such interrelated clusters, which is a task for Chapter 4. For the moment, however, note that it does not follow that, if one problematic term requires a neologism, then all closely related terms will too. The only reason for thinking so and for suspecting that there are clusters of "interdefined" terms is the mistaken assumption that these "definitions" are central to giving the meaning of the terms in question. But if, as I have already emphasized on more than one occasion, what is definitional for one theory may not be so for another, then there is no reason to think that these clusters should resist translation wholesale. In many cases, one term will require a neologism though closely related terms do not, even ones which might appear to be linked "definitionally". In Chapter 4, I will argue, for example, that though we need to neologize for the term 'phlogiston', we do not for the term 'dephlogisticated air'.

Another problem that Kuhn finds with the translation of rival or past scientific theories is the one dubbed "conceptual disparity" in section 1.5., and it also seems to involve neologisms (indeed, it is not clear that he considers it a separate problem, though it is usefully distinguished from the above problem). Kuhn observes correctly that neologizing need not be viewed with suspicion, since even if there were no English word for rabbit, the native's 'gavagai' could be introduced as a neologism. As he puts it:

> If the description [associated with 'gavagai'] is successful, if it fits all and only creatures that elicit utterances involving 'gavagai', then 'furry, long-eared, bushy-tailed... creature' is the sought after translation, and 'gavagai' can thereafter be

> introduced into English as an abbreviation for it. Under these circumstances, no issue of incommensurability arises. (1983a, 673)

However, Kuhn goes on to claim that there are some cases of neologizing that do not conform to this pattern. It is worth quoting what he says in full:

> In learning to recognize gavagais, the interpreter may have learned to recognize distinguishing features unknown to English speakers and for which English supplies no descriptive terminology. Perhaps, that is, the natives structure the animal world differently from the way English speakers do, using different discriminations in doing so. Under those circumstances, 'gavagai' remains an irreducibly native term, not translatable into English. Though English speakers may learn to use the term, they speak the native language when they do so. These are the circumstances for which I would reserve the term 'incommensurability'. (1983a, 673)

The reason that some terms of the alien language are not capable of translation, according to Kuhn, is that the natives "structure the animal world differently" and that they use "different discriminations as they do so." But it might be asked whether these structures and discriminations could not themselves be translated into the terms of the target language, so that 'gavagai' could eventually be given a translation. There may be related terms that must be translated before the translation of the desired term can be given, but that does not constitute an insurmountable obstacle. Of course, with theories that are separated by a large gulf of sophistication, neologizing might have to be fairly widespread, but that is not to say that two such theories are incomparable, just that one of them is silent on matters about which the other has an elaborate theory. This would be a case of the kind of subject-altering change mentioned in section 3.2.[16]

At any rate, it seems clear that Kuhn is not exercised by the kind of innovation typically introduced by a whole new branch of science, since he uses an example from ordinary discourse to demonstrate what he calls "conceptual disparity". As we saw in

---

[16] In section 7.2., I will consider and rebut another account of what it is for one theory to structure the world differently from another theory, an account developed on Kuhn's behalf by Hacking.

section 1.5., Kuhn illustrates this translational problem by using the example of the French word 'doux', which he claims can mean 'sweet' when applied to honey, 'bland' when applied to soup, 'tender' when said of a memory, or 'gentle' when predicated of a slope or a wind. Recall that Kuhn rejects the suggestion that this is simply an ambiguous French term. Rather, he emphasizes that 'doux' is a unitary concept for French speakers and English speakers "possess no equivalents." (1983a, 679-80) In a footnote, he states that this objection to the Quinean approach is equivalent to the difficulty raised above, that of clusters of interdefined terms. He writes that long English paraphrases for French terms provide no substitute, partly because of their clumsiness but mostly because such terms are "items in a vocabulary certain parts of which must be learned together." (1983a, 685n.12) Kuhn allows that a Quinean translation manual can deal with cases of straightforward ambiguity, but he argues that these examples should be distinguished from standard examples of ambiguous words, such as 'bank'.

However, it is not so obvious that this is not merely a case of standard ambiguity. Kuhn notes that there can be one-many linkages in a Quinean translation manual and adds: "Where the linkages are one-many, the manual includes specifications of the contexts in which each of the various links is to be preferred." (1983a, 679) Indeed, in one of the passages quoted above, he gives one-word specifications of such contexts for the word 'doux'. As he specifies them, the contexts are too vague; the translation manual or truth theory needs to be relativized to a speaker and a time. But the general point is that Kuhn has given (in brackets) quite adequate one-word context-dependent translations that would enable an English speaker to interpret a French speaker's utterance of 'doux' on a variety of occasions. To make matters clearer, we may want to introduce subscripts for each class of utterance: 'doux1', 'doux2', and so on.

The example that Kuhn adduces can either be tackled by the introduction of a neologism for the French term, or else it may not require the use of a neologism at all and may be handled by introducing subscripts for an ambiguous term and using more than one of our terms to translate. In neither case would it render the two theories or natural languages incommensurable. It would be irresponsible to pass a verdict on which of the two remedies to adopt in this case without tackling the case as a whole. Such case studies will be treated in detail in the following chapter, although they will be drawn from science

rather than ordinary language. If it turns out to be a case of ambiguity rather than neologism, the problem of conceptual disparity seems quite distinct from that of clusters of interdefined terms. The English speaker need only distinguish different meanings of the French word 'doux' and proceed to give English equivalents for each based on the context in which it appears. Although the context specifications may be elaborate, they are in principle no different from those needed to distinguish familiar homonyms like 'cape' (sleeveless outer garment) and 'cape' (peninsula on the coast).

This might provide Kuhn with an opportunity to rephrase his position. After admitting that a translation need not replace words and phrases on a "one-for-one" basis, he adds: "Glosses and translators' prefaces are not part of the translation, and a perfect translation would have no need for them." (1983a, 672) He claims that he derives this injunction from Quine's own conception of translation. But Quine explicitly states concerning the analytical hypotheses of a translation manual:

> Certain contexts may be specified in which the word is to be translated one way and others in which the word is to be translated in another way. The equational form may be overlaid with supplementary semantical instructions ad libitum. (1960, 69-70)

In the case of scientific discourse, ambiguous terms should be treated as distinct terms in the analytical hypotheses. As I have mentioned, this fact can be signalled by the use of different subscripts, with a specification of the context in which each term features. The translation function can then be accurately described as a mapping from the alien theory to the home theory, since there will correspond exactly one element in the home theory (or target set) to every element belonging to the alien theory (or source set). Strictly speaking, that description is only appropriate when both theories are supplemented by the necessary neologisms. Since the alien theory will generally contain terms for which one must neologize, the translation function will only satisfy the condition for being a mapping if these neologisms are counted among the elements of both sets. And if ambiguity is dealt with as suggested, distinct elements of the source will always be taken to distinct elements of the target, so it will not contravene the condition for being a mapping. Indeed, it is misleading on this view to speak of 'ambiguity' at all, for that suggests that two meanings are present on each occasion of use. A better term would be 'equivocality', or to use the

more standard expression, 'polysemy'. As for the thorny problem of identifying ambiguity and distinguishing it from, say, vagueness, some proposals will be put forward in section 5.3. to help with this task.

In light of these clarifications, I will make a last stab at addressing the source of Kuhn's discomfort with the kind of method adopted by the interpretive approach. What might lie behind his example of the French term 'doux' is a sense, shared by many authors philosophical and otherwise, that translation is somehow ineffable and that a fully satisfactory translation is an unattainable ideal. It might be claimed that if these words are treated as examples of standard equivocality, by introducing subscripts and considering them as separate entries in the translation manual, what is left out is the fact that it is the same word that is being used by French speakers. This suggests that the requirements placed on a fully adequate translation function are so numerous that it becomes unlikely that they can be satisfied all at once. For example, there is reason to doubt that the shape and sound of words, their familiarity and etymological relations to other words, can generally be preserved by a single translation function that also preserves literal meaning. These and other requirements, often lumped together under the heading of connotation[17] or nuance, are slippery to be sure. My claim, however, is that these features can be ignored when translating scientific discourse, since all that is relevant is informational content and truth conditions--in short, literal meaning. The justification for this can be sought in the aims and values of science itself. In other words, the autonomous character of the constraints that are applicable when comparing scientific theories is a by-product of the autonomy of scientific values.[18] There may be different sorts of constraints that are suited

---

[17] Although not in the technical sense of the word, as when it is used in conjunction with 'denotation', for example.

[18] The claim that science has autonomous values is sometimes expressed by saying that it is "value-neutral". However, as Isaac Levi has argued, it should be taken to mean that no further values are involved in making scientific judgments than those inherent in the canons of scientific inference. He writes that "the value-neutrality thesis does not maintain that the scientist qua scientist makes no value judgments but that given his commitment to

to the translation of other types of discourse, but they need not concern us here. In scientific discourse, the fact that the same word is being used for two different concepts can be regarded as irrelevant to the literal meaning of the words. If philosophers protest that this would be to assimilate them all to the status of insignificant lexical equivocality of the 'bank'-'bank' type, it is salutary to note that the philosopher's favorite example of allegedly harmless equivocality turns out not to be so harmless after all. The word 'bank' became equivocal as a result of gradual etymological divergence. The word for a mound or ridge came to be applied to any solid raised surface, including a money-changer's table or bench--and the rest is the history of capitalism.

The above prescriptions, as well as the interpretive principles to be outlined later in this work, reveal a commitment to a certain notion of interpretation or translation particularly suitable to the comparison of scientific theories. If the truth and falsity of the sentences of the alien theory is our primary concern, that licenses us to ignore other features of the theory. This is more apparent from a contrast that might be drawn with the translation of other types of discourse. In translating poetry, it might be said that alliteration or rhyme scheme must be preserved and this requirement will place a special constraint on the admissible translations of a poem written in a foreign language. There may also be special desiderata to be satisfied by other kinds of translation. In translating political speeches, certain historical allusions might have to be substituted by others, and in rendering jokes, the equivalents for puns may have to be sought. However, such features can be ignored in choosing an optimal translation function appropriate to scientific discourse.

I think it is safe to conclude that Kuhn has not succeeded in defending incommensurability against the interpretive approach. He seemed to have three separate sources of concern (although they were run together in his argument): neologism, ambiguity, and nuance. But none of them poses insurmountable problems for the task at hand and Kuhn has not managed to specify a clear sense in which the interpretive approach

---

the canons of inference he need make no further value judgments in order to decide which hypotheses to accept and which to reject." (1960, 356)

inevitably fails to capture something important about a term from an alien scientific theory, thus making it "an irreducibly native term". The analogy proposed by both Kuhn and Feyerabend, that learning a new scientific theory is like an infant's learning a language from scratch, is seriously misleading. The terms of a new scientific theory are learned within the context of a total theory of the world. To be understood, they must be related to existing terms that are shared amongst science and other parts of our global theory. Even if some alien terms require us to resort to neologisms, they cannot be learned entirely "from scratch".

3.5. Extreme vs. Moderate Holism

The holistic character of interpretation has already been encountered. Recall that in Davidson's ketch-yawl example, the attribution of a certain belief or concept depended on the attribution of certain other beliefs and concepts. Had the other attributions differed, we might not have decided with Davidson to pair the friend's term 'yawl' with the interpreter's word 'ketch'. Rather different verdicts might have been obtained in the (unlikely) event that the agent had been ascribed mistaken perceptual beliefs, or had we chosen (more improbably) to interpret the word 'that' differently. Other terms have themselves been translated on the basis of evidence, a fact that becomes painfully clear when we try to translate them differently and find that this forces too many revisions in the truth values of the translated sentences of the alien theorists. The inextricability of meaning and belief (or theory) does not imply that meaning change cannot be extricated from theory change in practice. Despite the fact that the meaning of a term is dependent on the theory in which it resides, meanings can be shared though theories may differ.

Having said that, the version of holism espoused by the interpretive approach must be distinguished from the more extreme version associated with the network picture of linguistic meaning attributed to virtually all the authors discussed in Chapter 1. In interpreting an alien language, it is not the case that a disagreement with some of the alien's beliefs renders all the alien's concepts different from our own. The extreme version of holism, which has been widely attacked, is often visualized in terms of the "web of belief". The picture suggests that a change at any point in the system acts like an extensional displacement that is transmitted to the whole network. The genealogy of the

network picture may be traced to Quine's "Two Dogmas of Empiricism", where he introduced the metaphors of the "man-made fabric" and the "field of force". (1961, 42-4) But that is not to say that Quine is guilty of subscribing to this naive version of holism; indeed, his gloss on the metaphor suggests that the accusation would be unfair. A more explicit and concise statement of the metaphor can be found in Hempel: "the concepts of science are the knots in a network of systematic interrelationships in which laws and theoretical principles form the threads." (1966, 94) This image suggests that an unavoidable shift afflicts all the conceptual knots of the network with every twitch of the theoretical threads, in such a way that the new network cannot be superimposed onto the original one. The picture is reminiscent of Feyerabend's "contextual theory of meaning". Recall that Feyerabend contrasts his extreme holistic theory of meaning with the "Swiss cheese" theory of meaning, which allows conceptual holes to be plugged without disturbing the entire theory.

The critique of semantic holism has been given prominence in a work by Fodor and Lepore. In an argument borrowed from Dummett, Fodor and Lepore suggest that holism might have dire consequences for the possibility of intentional explanation, a science of psychology, the standard picture of language-learning and communication, and (most significantly for our purposes) scientific theory choice. The reason is that holism presumably dictates that the meaning of any term in an agent's set of beliefs is given by that whole set of beliefs. But, in general, no two agents' sets of beliefs are identical, so the meanings of two terms in the idiolects of different agents cannot be identical. Since this rules out the possibility of comparing two agents' sets of beliefs, or indeed a single agent's sets of beliefs at different times, the objectors conclude that it seems to spoil things for intentional explanation and scientific theory choice, among other things. (1991, 8-9)

If the variety of holism inherent in the interpretive approach is not to lead to the disastrous scenario just sketched, one needs to explain what exactly blocks this misalignment. On the interpretive approach, the judgment that a certain concept is shared among two theories does not convey anything <u>specific</u> about the contents of those theories. Every time a concept is said to be shared among two theories we can safely conclude that there is a certain amount of agreement among them, but we cannot thereby conclude anything about the precise nature of that agreement. A concept can be shared among two

agents even though many beliefs are not shared among them. In the ketch-yawl example, the speaker was attributed the concept of a ketch as soon as it was decided to ascribe any ketch-beliefs to him at all. Provided such ascriptions are final, there can be no further question as to whether the agent has the concept of a ketch. Otherwise, the interpreter should have ascribed different beliefs, ones containing a different concept. Before reaching the point at which a belief is ascribed, much evidence will have been gathered and interpretive choices made. But once that process is over and the agent's beliefs have been spelled out, the interpreter is committed to attributing the concepts that are featured in those beliefs. Hence, some nodes will remain in place after any change in theory by dint of the fact that some beliefs will always be shared. It is not as if the whole network will be overrun despite the interpreter's valiant efforts to resist the forces of meaning change.

This response to the critique of holism emphasizes the fact that conceptual ascriptions are grounded in the interpreter's judgments and choices. A translation or interpretation is constructed between two sets of beliefs that provides the best overall explanation of the alien's utterances and actions. The decision to translate an alien term with one of our own is not made on the grounds that the term features in all the same beliefs, or even a specific set of requisite beliefs. Of course, it will turn out that certain beliefs are shared for each particular term, but we cannot specify which ones these will be in advance since they can only emerge after the process of interpretation is complete in accordance with certain interpretive constraints. To put it in terms of Quine's famous example, after several encounters with my linguistic informant, I decide that the available evidence suggests translating the native's term 'gavagai' by my term 'rabbit'. That is not to say that I need to share all the native's beliefs about rabbits in order to make this decision. For example, it may turn out that the native regards rabbits to have religious significance and that his term 'gavagai' is often mentioned in the same breath as another term that I have already translated as 'sacred'. Still, that should not force me to attribute a different concept to the native, say the new concept schmabbit. I merely attribute the belief that rabbits are sacred. Since the relation between words and the world is not piecemeal and direct, the fact that two agents share a certain concept does not imply any particular thing about their holding of certain beliefs (as it would for the cluster theorists, for example), or their relation to the environment (as it would for the causal theorists of reference, for

example). The sharing of most concepts among interpreter and interpretee has no immediate implications for the sharing of particular beliefs. At the same time, it blocks the possibility of a radical misalignment in the concepts of the two theories.

Therefore, our practice as interpreters belies the alleged consequences of holism. Actual interpreters do not regard every difference in belief as leading to a difference in concept, but normally absorb large differences in belief within shared concepts. Since this is central to the practice of real interpreters, and since the ascription of concepts is grounded in actual interpretive practice, this enables us to say that, as a matter of contingent fact, every change in belief does not lead to a change of meaning of all relevant terms. This shows us what is wrong with Fodor and Lepore's critique. That critique presumes that the ascription of a concept is at the mercy of a surrounding network of beliefs, rather than being the result of an interpretive decision. To be sure, these decisions are sensitive to the beliefs that are held by the interpretee, but conceptual ascriptions do not vary inexorably with every variation in belief. Moreover, once all the available evidence suggests translating one of the native's sentences by one of our own, the component terms of the two sentences are matched up willy-nilly. Interpreters do not require <u>every</u> connection between terms to be preserved for the terms to be correlated, as the critics of holism seem to assume. An interpretation is constructed that best explains the utterances, beliefs, and actions of a holder of the alien theory--and that is what determines the shared and unshared meanings. Rather than an exact isomorphism between two theories, the interpreter tries to achieve an overall fit which satisfies the constraints.

In response to the perceived problems with semantic holism, some theorists of meaning may want to appeal to an atomist theory of meaning, according to which the meaning of each word is given directly by physical relations to the world itself. However, one need not resort to such a theory, for there is a more moderate version of holism about meaning that does not have the extremist consequences mentioned. One can maintain that the meaning of any word depends generally on the way that that word and other words are used, but resist the suggestion that an alteration in the use of <u>any</u> word changes the meaning of <u>every</u> related word. On a moderate holist theory of meaning, meanings coincide whenever a term from one theory is used to translate a term from another theory.

In such cases we say that the two terms have the same meaning, or equivalently, that they pick out the same concept. The decision to translate or interpret a term in a certain way depends generally on the way that that term and other terms are used, but we do not regard every difference in the way that terms are used as leading to a difference in meaning. This interpretive practice is what ensures that every difference in the way that terms are used does not lead ineluctably to a change in meaning of all our terms. Thus, there is a moderate holist alternative to extreme holism.

3.6. Interpretive Principles and Theory Choice

The basic features of the interpretive approach that make it suitable for the task of theory comparison have now been mapped out. After providing an account of Davidson's response to the alleged possibility of totally incommensurable languages, I proposed a reconstruction of his argument against partial incommensurability, in such a way as to make clear how it could be applied to the problem of interpreting scientific theories. I went on to argue that the indeterminacy that afflicts interpretivism can be constrained for the purpose of translating one scientific theory into another. Then, I addressed some of the concerns associated with the use of neologisms to translate terms from an alien scientific theory. These concerns included ones that Kuhn raised concerning clusters of supposedly interdefined scientific terms and alleged conceptual disparity among scientific theories. Finally, I distinguished the moderate holism associated with the interpretive approach from the extreme holism that was encountered in Chapter 1. Unlike extreme holism, moderate holism does not have the consequence that the terms of a scientific theory will change their meanings with every theoretical change. When these features of the interpretive approach are strengthened by some of the interpretive principles to be articulated in the following two chapters, a framework will have been provided for giving the meaning of scientific terms or for comparing scientific theories, and hence for distinguishing meaning change from theory change.

This is not to deny that there will be some difficult decisions to be made concerning the ascription of certain concepts. For example, it is sometimes difficult to say if an agent's beliefs, discriminations, and interactions with the environment have sufficient complexity to warrant ascribing a certain concept. But the examples in Chapter 4 and the principles

presented in Chapter 5 will help to spell out the considerations that enter into such decisions and lay down principles of interpretation that should drive these decisions. Some of these principles elaborate and make concrete a few of the observations made in this chapter, particularly concerning conceptual charity, neologizing, and preserving literal meaning. Other principles introduce new considerations, particularly suited to the interpretation of scientific theories. These include a principle that specifies which terms are to be taken as standing for conceptual primitives requiring separate entries in the lexicon and which to regard as complex concepts made up of simpler units. Another principle tells us when to translate a scientific term uniformly and when to translate it by different terms on different occurrences. Yet another principle recommends treating putative definitions as one would other theoretical tenets in a scientific theory and cautions against privileging them. When applied to an alien scientific theory, these constraints will together issue in a best way of interpreting that theory in terms of our own.

Notice that this way of doing things effectively rehabilitates concepts. If one argues for an optimal interpretive mapping between any two scientific theories, then one is arguing that there will be a best way of translating the terms of an alien theory into the terms of the home theory. That should lead us to conclude that the two theories share some determinate set of concepts. Thus, concepts acquire fixity and a genuine existence as by-products of this interpretive process. That is not to say that they should be thought of as bona fide objects of a spatiotemporal nature. But they do have the metaphysical status of entities (as I shall explain in Chapter 6) and can be talked about, compared, and so on. For it is clear that such entities are regularly invoked to play an explanatory role in the cognitive sciences, among other disciplines. Rather than summarily dismiss appeals to concepts and conceptual changes, or attempt to reconstrue them as disguised talk about something else, philosophers owe us an account of what concepts are (which I will try to do in Chapter 6). Despite the fact that the notion of a conceptual scheme has been repudiated by Davidson, there is no need to banish concepts altogether. The slogan associated with this position might therefore be: Concepts without conceptual schemes.

Finally, it should be made clear how the interpretive approach is suited to determinations of theory choice. From the way in which the interpretive situation has often been described, it might be thought that, by the time one theory is being translated

into another, the choice between two theories is moot. Since it has usually been assumed that the interpreter holds one theory (the target) and is comparing it to an alien theory (the source), the dice might seem to be loaded in favor of the interpreter's theory. However, the interpretive process must be carried out by some agent or another with some set of beliefs, which might as well be one of the theories involved. Interpreting from the point of view of one theory or another will not be prejudicial, for I will argue in Chapter 6 that the interpreter aims to preserve all distinctions made by the interpretee and does not make any distinctions not made by the interpretee. Therefore, there is no asymmetry in the interpretive process provided the interpreter makes sure not to impute a conceptual apparatus that is any more fine-grained or coarse-grained than the person being interpreted. It is only after the process of interpretation has been completed and the area of disagreement between the two theories has been identified that the question of choice can arise. If we are then to choose between the two theories, we must withdraw to the area of agreement and apply decision-theoretic procedures to determine which beliefs to import from among those that are in dispute.[19] One might argue that a rival theory ought to be interpreted from a neutral perspective rather than the perspective of our own theory. But we cannot locate the neutral ground between two theories unless we can locate their area

---

[19] As Levi has emphasized, any revision of an agent's beliefs is analyzable into two steps, a contraction followed by an expansion. By that he means, that one must first get rid of the theoretical tenets that one wants to give up, and then import those tenets that one wants to import. Any process of belief revision can be broken down into these two steps. He writes that "both contraction and expansion can be viewed as the fundamental types of revision subject to critical control, and all other sorts of revisions may be then understood as sequences of changes of these kinds," adding that he is "not claiming that the historical record will reveal that replacements of one theory by another always take place as the net result of an explicitly or consciously implemented sequence of contractions and expansions. However, if such replacements are defensible, they should be decomposable (for purposes of analysis) into sequences of this sort; in such a sequence, each step must be justifiable." (1980, 65) Therefore, once the area of agreement has been identified, embracing the alternative theory is two steps away, in the form of a contraction followed by an expansion.

of agreement. And we cannot locate their area of agreement unless we can interpret one theory in terms of another. So we must interpret from the perspective of our own theory first, and then locate the area of agreement (which is identical with the neutral ground) from which to make a rational choice.

Chapter 4: **Cases**

> I could not but see, for example, when Einstein set philosophers talking about relativity, that philosophers' convictions about the eternity of problems or conceptions were as baseless as a young girl's conviction that this year's hats are the only ones that could ever have been worn by a sane woman.
>
> R.G. Collingwood, An Autobiography

4.1. Reconstruction of Theory Fragments

In this chapter, four main case studies will be discussed from the history of natural science; a section will also be devoted to showing that the interpretive method has some application to the realm of political and social theory. All these cases have already been examined by philosophers who represent different approaches to the problem of theory-comparison. Sometimes their analyses will be followed (after locating them within the interpretive framework), but sometimes they will be resisted. Interpreting the theories generally involves some degree of reconstruction or recasting. In some cases, that requires tidying up or rephrasing certain crucial parts of the alien theory; in other cases, it involves attributing tenets or supplying arguments that are only implicit in the theories. These acts of reconstruction have been indicated at the relevant junctures. I would claim that the reconstructions involved do not do violence to the original theories. Some philosophers are wont to chafe at any attempt at reconstruction[1]; the attempts to follow will be justified for those who do not think that it is illegitimate in principle.

The following examples do not concern entire theories so much as small groups of problematic concepts drawn from particular theories. For example, in interpreting classical mechanics in terms of relativistic mechanics, certain concepts are singled out for

---

[1] Thus, Feyerabend remarks that "it is plausible to assume that the comparison of A and B [two ancient Greek cosmologies] as interpreted by the participants (rather than as 'reconstructed' by logically well-trained but otherwise illiterate outsiders) will raise various problems." (1975, 264)

treatment: mass, velocity, force, and so on. Not only are these the most contentious concepts of the two theories, it seems enough to focus on small clusters in order to indicate the general method. These sub-theories cannot be isolated from the larger theories in which they are embedded, and at certain times, some interpretations will be considered that have an important effect on our interpretation of the rest of the total theory. Therefore, the rest of the surrounding theory will not be ignored entirely, though the focus will be on a small cluster of concepts in each case.

Another cautionary note is in order when considering the following case studies. In all of them, the choice between two theories has already been made, so that one of the theories is just false from our present perspective. But this fact should not be considered prejudicial to the interpretation of the theory that has been rejected. As explained in the previous chapter, the interpretive process must be carried out from some perspective or another, and it cannot be carried out from a neutral perspective, since the common ground can only be scouted once the interpretive process is complete. But in all cases, it is only a short distance to retreat to neutral territory and, therefore, to a perspective from which a choice can be made (just substitute 'shared' for 'true' and 'disputed' for 'false').

Finally, it needs to be pointed out that in what follows, issues about sense and reference are being bracketed for the time being in discussing how one is to interpret alien scientific theories. It may be objected that some of the philosophers who have tackled these case studies are explicitly interested in the reference or extension of these terms, while other are interested in the meaning or concept associated with these terms. But what they all have in common is a concern with translation or interpretation. Therefore, I will focus in this chapter on the enterprise of translation and frame my judgments, criticisms, and conclusions in these terms. Later, in Chapter 6, I will try to justify further the claim that what ought to be preserved in interpretation is the content of an agent's concepts and that the extension of an agent's terms (in a simple-minded sense to be articulated) are not ignored entirely but can be incorporated into the interpretive approach.

4.2. Classical Physics and Relativistic Physics

One of the most widely-discussed case studies when it comes to the problem of meaning-change is that of the replacement of Newtonian dynamics by relativistic dynamics. It is with reference to this scientific revolution that Kuhn talked about a "displacement of the conceptual network through which scientists view the world." (1970, 102) He claimed that the differences were irreconcilable, later implying that the two theories, like other pairs, were incommensurable. Of this same scientific change, Hartry Field has claimed that it is indeterminate (in Quine's sense) whether we should identify one of the basic terms of Newtonian theory with one or the other of two terms of relativistic physics. He has argued that it is referentially indeterminate what Newton's term 'mass' referred to from the perspective of Einstein's theory, whether 'proper mass' (i.e. 'rest mass') or 'relativistic mass'.

Field starts by criticizing Kuhn for failing to establish the conclusion that Newton's term 'mass' does not denote <u>any</u> of the physical quantities in Einstein's theory. He admits that Kuhn shows that Newton had many beliefs about 'mass' that are no longer accepted, but maintains that this does not rule out the possibility that Newton was indeed referring to some quantity in relativity theory but had many false beliefs about it. However, he claims that there are other considerations that <u>do</u> serve to rule out the possibility and proceeds to present them.[2]

Field's argument can be summarized as follows. Consider the term 'mass' in Newtonian dynamics and the terms 'proper mass' and 'relativistic mass' in the special theory of relativity. Some sentences of Newton's theory come out true (i.e. equivalent to sentences of relativity theory) if one substitutes 'proper mass' for 'mass', while other

---

[2] Note that Field's comments on Kuhn help to vindicate one of the claims made in the previous chapter concerning the relationship between incommensurability and indeterminacy. It was argued there that similar implications can be read into the two theses. While Kuhn and Field concur in holding that Newton's term 'mass' does not correspond to any <u>one</u> term of Einstein's theory, Field thinks that it corresponds (in a sense to be specified shortly) to <u>more than one</u> of the relativistic terms, while Kuhn thinks it corresponds to <u>none</u>.

sentences come out true if one substitutes 'relativistic mass' for 'mass'. As an example of the former kind, Field proposes:

> (5P) For any two frames of reference, mass with respect to frame 2 = mass with respect to frame 1.

And as an example of the latter kind, he gives the sentence:

> (4R) Momentum = (mass) v

Field then proceeds to make four main claims, which can be paraphrased as follows (1973, 468-472):

(1) Neither set of sentences made true by the two substitutions is more central to Newton's theorizing and scientific practice.

(2) Neither concept proper mass nor relativistic mass is more central to special relativity: some laws of physics come out simpler in terms of proper mass, others in terms of relativistic mass.

(3) Newton was not referring to some third concept of Newtonian mass, different from either concept found in relativity theory, because Einstein showed that there is no such quantity.

(4) Newton's term 'mass' should not be construed as being denotationless, provided one assumes a plausible principle of substitutability for such terms.

From these four premises, Field concludes that:

(5) Newton's term 'mass' partially denotes both proper mass and relativistic mass; the term is said to be referentially indeterminate. (1973, 474-5)

Field goes on to define the concept of partial denotation in formal terms. First, he defines a structure as a function mapping each name or quantity term into an object or quantity, and each predicate into a set. Then, he states that a sentence is m-true (m-false), i.e. true (false) relative to a structure m, if the sentence would be true (false) if the denotations and extensions of its terms were as specified by m. A sentence is true (false) simpliciter if it is m-true (m-false) for every structure m that corresponds to it. Moreover, the sentence is truth-valueless (neither true nor false) otherwise. (1973, 477-8)

It is important to notice that "denotes" cannot be defined "in any acceptable way" in terms of "partially denotes". As Field observes, if one takes a term that partially denotes exactly one thing as fully denoting that thing and one then equates full denotation with

denotation simpliciter, one would obtain undesirable results. It would turn out that a term might not denote anything and yet not be denotationless (i.e. because it might still partially denote something). (1973, 475) The gist of Field's proposal is that any substitution that makes some sentences of a theory true relative to some given assignment should be taken as an indication of partial denotation. Moreover, he suggests that the notion of partial denotation should be taken as more primitive than denotation, and that the latter term should be completely abandoned in favour of the former.

Recall that, by making claims (1) and (2) above, Field effectively admitted that there were certain kinds of considerations that might help to decide between one of the two analytic hypotheses, identifying 'mass' either with 'proper mass' or with 'relativistic mass'. He did not deny that these sorts of considerations may play a role, but maintained that in the particular case under examination, they were equally weighted on both sides. It is this last claim that was forcefully challenged in replies to Field made by John Earman and Arthur Fine. Earman effectively undermines both (1) and (2), by claiming first that some of the tenets made true by one of the substitutions are more central to Newtonian mechanics, and, furthermore, that one of the two concepts considered as candidates for substitution is more central to relativity theory. Fine corroborates these claims by citing textual evidence that indicates that this accords with Einstein's own view.[3]

Earman begins by casting Newtonian mechanics in a four-dimensional intrinsic (i.e. coordinate-free) form:

(N1) m_N is a scalar invariant

(N2) $P_N = m_N V_N$

(N3) $F_N = m_N A_N$

where m_N, P_N, V_N, F_N, A_N are, respectively, the Newtonian mass, Newtonian four-momentum, four-velocity, four-force, and four-acceleration. He goes on to make two claims:

---

[3] Fine's appendix to Earman's argument presents conclusive evidence from Einstein's writings that he regarded the theory in the same light. In one letter, Einstein wrote: "I find it not very good to say that the mass of a body in movement is increased by the speed. It is better to use the word mass exclusively for rest mass." (1977, 538)

> This form is (i) the most perspicuous form known for either theory in terms of making the underlying mathematical and physical assumptions clear and explicit, and (ii) the most useful form known for comparing the two theories. (1977, 535)

Earman notes that there are "exact analogues" in special relativity (R1), (R2), (R3) of (N1), (N2), (N3), with proper mass in place of mass, relativistic four-momentum in place of momentum, etc. He then makes the claim that these tenets are central to both Newtonian mechanics and special relativity (effectively denying Field's claim (1) above). Therefore, he implies that a translation that makes these tenets come out true is one that is to be favored, namely the substitution of 'proper mass' for Newton's 'mass'. Then, Earman goes on to claim that 'proper mass' itself is more fundamental to special relativity (effectively denying Field's claim (2)). He writes:

> The so-called "relativistic mass" comes from a three-dimensional coordinate-dependent effect associated with inertial coordinate systems. If, therefore, 'relativistic mass' denotes a new kind of mass, then for every distinct kind of noninertial coordinate system there will be yet another new kind of mass. It seems to me, however, that instead of multiplying masses it is preferable to say that there is only one kind of mass in SRT, namely, that denoted by 'proper mass'. (1977, 537)

Finally, Earman notes that he is unable to come up with any examples from other scientific theories for which Field's argument is convincing.

The above exchange between Field and Earman was worth relating at some length since I will now argue that it provides support for the procedure for theory-comparison outlined in Chapter 3. The full argument in favour of Earman's formulation is rather involved and it would take us too far afield to pursue it in detail.[4] What is significant about it, however, is that it draws attention to certain crucial tenets of Newton's theory that come out true under the proposed formulation. Earman says that the tenets of Newton's theory, (N1)-(N3), have "exact analogues" in relativity theory, (R1)-(R3). He concludes that the concepts that occur in them should therefore be identified with each other, that is, that "mN and m0 [rest mass] have the same denotation". (1977, 535-6) This is just a way of saying

---

[4] The argument is presented in Earman and Friedman (1973).

that these theoretical tenets are shared between the two theories and that their constituent concepts should be translated uniformly when they occur in other tenets of the theory.

In the previous chapter, it was stated that interpretation often begins by taking certain beliefs of an agent or certain tenets of a theory to be shared among interpreter and interpretee. In science, these tenets may be culled from a particular axiomatization or a certain salient formulation of the theory to be interpreted. This is clearly the principle at play when Earman casts the two theories in the four-dimensional, intrinsic form (N1)-(N3) and (R1)-(R3). While Earman takes these tenets to be crucial (and therefore shared), Field clearly would not. The disagreement between Field and Earman is underlined by noting that Field's principle (5P) corresponds in part to Earman's (N1), yet while Earman takes it to be central, Field regards it as no more central than certain other tenets of the theory. This means that he does not think it matters particularly whether it comes out true under a certain interpretation of Newton's theory. He thinks it equally plausible that it should come out false (not shared) and other tenets should be the true ones. Some of the latter set of tenets are such that substituting 'relativistic mass' for 'mass' in them will issue in true tenets of special relativity. However, Earman implicitly regards them as less central in arguing against Field's view that there is nothing to decide between the rival analytic hypotheses. If Newton is interpreted as having the concept of proper mass (as Earman advocates), then he was just mistaken (from the perspective of relativistic physics) in asserting certain tenets. In fact, such tenets help to show precisely where Newton went wrong.

Another desideratum for the interpretation of one theory by another is to make certain other tenets are consistent with the theory, though perhaps false. Thus, one of the chief virtues of the form given by Earman and Friedman is that it enables one to make sense of Newton's First Law. The claim they make is that in the usual three-dimensional formulations of Newton's theory, the First Law must be taken as referring to some given inertial frame or class of frames. However, the theory neither says that inertial frames exist nor does it specify what they are, thus rendering the statement incomprehensible. Remedies have been suggested for these difficulties, for example including an explicit existence assertion in the First Law, or regarding it as a definition of inertial frame. Against these proposals, Earman and Friedman argue that the difficulties do not even arise in the

four-dimensional formulation because of the existence of Absolute Space and Time. They write that "the concept of a straight line in space-time is well-defined, and the First Law can be stated in the four-dimensional form: The world line of a particle free of impressed forces is a straight line in space-time." (1973, 338) The principle at work is one that calls for distribution of truths and falsehoods among tenets of Newton's theory in such a way as to make it as rational and consistent as possible in conformity with the evidence.

Earman's argument also relies on yet another kind of consideration to adjudicate between analytic hypotheses; that pertains to the overall appeal of the formulation he gives and the overall fit that it generates between the two theories. He claims that the formulation he gives is the "most perspicuous" and "most useful". Now this might seem to be a kind of consideration that cannot be incorporated into the interpretive approach. However, Earman's judgment of perspicuity and usefulness relies on beliefs that he thinks will be shared among holders of the two theories, though they are not themselves part of the theories of relativistic or Newtonian physics, but are derived from the realm of meta-science. As argued in Chapter 3, once disciplines are demarcated, beliefs drawn from neighboring areas can be relied on to compare theories. In other words, there will be certain meta-scientific tenets or principles that tend to favour one interpretation over another. These are spelled out in greater detail in the paper by Friedman and Earman, which generally uses meta-scientific and philosophical considerations to argue for the superiority of their formulation over others. Much of the discussion revolves around the consistency of Newtonian dynamics with the Newtonian theory of gravitation.

The satisfaction of this last desideratum helps to defuse a possible objection to the argument, namely that the theory is capable of numerous multiple substitutions that Earman does not even consider. It might be claimed that Earman's interpretation assumes that a number of other terms in Newton's theory are also interpreted in a particular way, namely, Newton's momentum, velocity, force, and acceleration. Thus, if Newton's velocity were interpreted as 'relativistic velocity', our judgments might have differed. As Field suggests, 'velocity' may be indeterminate between rate of change of distance with time, t, and rate of change of distance with proper time, T, where $dT = dt\,[1 - (v^2/c^2)]^{1/2}$. It may be said that all combinations of possible multiple substitutions must be tried before one can say that any interpretation is the most satisfactory. To deflect this objection, it can be said

that the overall structure and symmetry of the formulation is unique. As I indicated earlier in this section, this is the only formulation which casts the two theories in an intrinsic or coordinate-free form, which is to say that this formulation does not contain any explicit reference to time and space. Therefore, the formulation brings out the fact that one of the main differences between the theories lies in the structure of space and time, which is Euclidean for classical mechanics and Minkowskian for relativistic mechanics.

The issues raised by Earman's response are complex and lead deep into the territory of the philosophy of physics. Luckily, however, our present concerns do not stand or fall on the merits of the particular position he takes. All that is necessary to vindicate the approach being proposed is that the considerations that are cited by Earman be of the right kind. It is important to note that both parties to the debate concur that the interpretation of Newton's concept of mass is to be decided upon by determining which tenets of his theory it makes true and which tenets of neighboring theories come out consistent (e.g. those drawn from meta-science, gravitational theory, and so on). The crucial difference is that Field claims that two different interpretations of the concept fare equally well when examined in this light, whereas Earman argues that they do not. Of course, this raises the possible objection that there may yet be unexplored constraints that would tip the balance the other way, leading us to translate Newton's 'mass' as 'relativistic mass' rather than 'proper mass'. That is a possibility, but an argument would have to be given for it, along the lines of the detailed argument that Earman and Friedman give for the interpretation they favour. If these rival grounds are convincing, our method of comparing Newton's theory to Einstein's may have to be revised.

A different objection, and one that Earman himself mentions is that, "The Newton conjured up by (N1)-(N3) is a fictitious Newton." (1977, 536) Earman's reply consists mainly in pointing to the evolution of Einstein's own views in formulating the special theory of relativity. Thus, Einstein himself initially talked in terms of three-dimensional quantities, shifting later to covariant quantities. Earman notes: "It is hardly plausible to think that a change in the reference of 'proper mass' took place as a result of this shift." (1977, 537) But if that is the case, Earman suggests, a critic cannot maintain that Earman's formulation involves changing the reference of Newton's term. Earman's reply to the objection is wholly in keeping with the intent of the interpretive approach. The objection

seems to assume that unless there is explicit evidence that Newton used four-dimensional language, such a formulation cannot be foisted on him. However, there can be many grounds for ascribing a certain belief to an agent, only one of which is explicit avowal of such a belief.

In any case, the point of comparing Newtonian with relativistic physics is not necessarily to resurrect the historical Newton. Rather, it is to come up with a formulation of Newtonian physics that is as rational and consistent as possible, albeit false at some points. Note also that some of the rival interpretations that Earman and Friedman canvas involve attributing to the Newtonian physicist certain beliefs that we have little reason to believe were held, for example an assertion of the existence of inertial frames. That is an interpretation that is inferior to one that has the Newtonian physicist casting the theory in a coordinate-free form. There may never have been an explicit formulation of this sort by a Newtonian physicist, but it accords better with the Newtonian physicist's actions and other beliefs.

4.3. Phlogiston Theory and Post-Phlogiston Theory

Another example that is popular in writings about theory-comparison is that of the phlogiston theory of eighteenth century chemistry. The problem, as some writers see it, is that it is hard to square two claims about the theory: (a) that 'phlogiston' is a term that fails to refer, and (b) that the phlogiston theorists made some correct statements and important discoveries. Kitcher explains some of the basic tenets of the theory:

> The phlogiston theory attempted to give an account of a number of chemical reactions, and, in particular, it offered an explanation of processes of combustion. Substances which burn are rich in a "principle", phlogiston, which is imparted to the air in combustion. (1978, 529-30)

Thus, when wood is burned, phlogiston is supposedly given to the air, and when a metal is heated, phlogiston is also emitted and the "calx" of the metal remains.

Then Kitcher paraphrases the following statements of chemical reactions from the writings of the phlogiston chemists:

(P1) Metal + air -heat-> Calx of metal + phlogisticated air

(P2) Calx of mercury -heat-> Mercury + dephlogisticated air

(P3) Metal + acid --> Salt + inflammable air

(P4) Iron + steam -heat-> Calx of iron + inflammable air

There is evidence from their observational reports to suggest that the phlogistonians were describing the following reactions (in terms of modern chemical theory):

(M1) Metal + air -heat-> Metal oxide + oxygen-deficient air

(M2) Oxide of mercury -heat-> Mercury + oxygen

(M3) Metal + acid --> Salt + hydrogen

(M4) Iron + steam -heat-> Iron oxide + hydrogen

A glance at the two sets of statements suggests a number of identifications. It is uncontroversial that 'metal', 'air', 'acid', and 'salt' should be translated homophonically and it is plausible to identify the 'calx' of a metal with that metal's oxide. This leaves three problematic terms in (P1)-(P4): 'phlogisticated air', 'dephlogisticated air', and 'inflammable air'. Once the interpretation of the other terms has been fixed, it is natural to translate these terms as 'oxygen-deficient air', 'oxygen', and 'hydrogen', respectively.

However, as Kitcher sees it, there is a basic problem in making such identifications (specifically the last three), since the presupposition that something is <u>emitted</u> in combustion infects all the terminology. As he explains:

> The view that phlogiston is a substance emitted in combustion is central to the phlogiston theory, and is the doctrine from which the theory develops. Hence, it is quite natural to assume that the reference of 'phlogiston' is fixed by this view, so that 'phlogiston' refers to that which is emitted in all cases of combustion. But there is <u>nothing</u> which is emitted in all cases of combustion. So it seems that we must conclude that 'phlogiston' fails to refer. (1978, 531)

But, Kitcher goes on, if phlogiston fails to refer, then so do the terms 'dephlogisticated air' and 'phlogisticated air', which are "abbreviations" for the expressions 'the substance which results from removing phlogiston completely from the air' and 'the substance which results from adding phlogiston to the air until no more can be absorbed', respectively. He asks rhetorically: "How can there be a substance which remains when phlogiston is removed from the air if there is no such substance as phlogiston?" (1978, 532)

Kitcher appeals to (a modified version of) the causal theory of reference in order to deal with the example.[5] However, I will employ the interpretive approach instead and proceed with the solution that Kitcher dismisses. Based on their experimental reports, we decide that the sentences (M1)-(M4) are shared among phlogiston theorists and modern chemists. In other words, (P1)-(P4) should be translated by (M1)-(M4) and the following interpretations should be advanced: 'phlogisticated air' is translated as 'oxygen-deficient air'; 'dephlogisticated air' is translated as 'oxygen'; and 'inflammable air' is translated as 'hydrogen'.[6] In addition, let us agree with Kitcher that the term 'phlogiston' fails to refer and can be translated neologistically as 'phlogiston' (with this proviso in mind). This assumption will be justified at the end of this section and the consequences of dropping it explored.

But there is another set of sentences that the phlogiston theorists also held, namely:

(P5) Phlogisticated air is the substance which results from adding phlogiston to the air until no more phlogiston can be absorbed.

(P6) Dephlogisticated air is the substance which results from removing phlogiston completely from the air.

Given our translations so far, these sentences should be rendered as follows:

(M5) Oxygen-deficient air is the substance which results from adding phlogiston to the air until no more phlogiston can be absorbed.

---

[5] I have already argued that Kitcher's version is unsatisfactory in Chapter 2.

[6] These formulations are meant to be neutral between treating the terms as names of substances and treating them as predicates (abbreviations for 'is an oxygen-sample', etc.). In either case, they are being considered as simple rather than complex expressions. Thus, 'dephlogisticated air' is a simple expression (not a description with a proper occurrence of a name, i.e. 'air from which phlogiston has been removed'), or a simple predicate (not a conjunction of three predicates, i.e. 'is an air-sample from which something which is a phlogiston-sample has been removed'). This treatment will be further justified in section 5.4.

(M6) Oxygen is the substance which results from removing phlogiston completely from the air.

Now, on most views of non-referring terms, (M5) and (M6) will be considered either false or lacking truth value, since we have already assumed that 'phlogiston' is a term for which we have no equivalent. Meanwhile, (P1)-(P4) are true sentences of that same theory. Indeed, that is just what we would expect: (P5) and (P6) are sentences that serve to pinpoint where the phlogiston theorists erred. It is reasonable to say that the phlogistonians correctly identified many chemical reactions involving oxygen and other substances but they went wrong in thinking that oxygen was produced by removing phlogiston from the air. Indeed, Kitcher presents conclusive evidence that the phlogistonians identified other important properties of oxygen. Priestley noted that mice flourished in 'dephlogisticated air' and breathed it himself, finding that, "The feeling of it to my lungs was not sensibly different from that of common air; but I fancied that my breast felt peculiarly light and easy for some time afterwards." Similarly, Cavendish observed the formation of water by synthesis of 'dephlogisticated air' and 'inflammable air' (hydrogen). (1978, 533)

However, Kitcher disagrees with the above interpretation. He thinks that this case necessitates what he calls a "context-sensitive theory" of reference, one that would recommend translating tokens of the type 'dephlogisticated air' differently depending on the context. He claims that "if we treat all tokens of the same type in the same way, then we shall be led to the position defended by Kuhn and Feyerabend: there is no term of contemporary English which specifies the referent of 'dephlogisticated air', so that a term which is central to the presentation of the phlogiston theory resists translation into contemporary language." (1978, 534) What is wrong with the translation offered above, which proposed a (type-type) translation of 'dephlogisticated air' as 'oxygen'? In Kitcher's view, it would "render some of Priestley's arguments or assertions inexplicable." The example he gives is Priestley's conclusion that whatever gas remains after heating the red calx of mercury will be dephlogisticated air. In this case, he says, "we hypothesize that the referent of his token 'dephlogisticated air' is fixed as that which remains when phlogiston is removed from the air." (1978, 535) That is, in this context, 'dephlogisticated air' is to be translated not as 'oxygen', but as the description 'air from which phlogiston has been

removed'. The reason he gives is that this hypothesis enables us to render the following judgment explicable: believing that the liberation of mercury involves absorption of phlogiston, Priestley inferred that the residual air would be poor in phlogiston, that is 'dephlogisticated air'.

What seems to be bothering Kitcher is the existence of a one-step deductive inference that he thinks will cause problems for a translator who renders 'dephlogisticated air' uniformly as 'oxygen':

(P7) Liberation of mercury involves absorption of phlogiston.

∴ (P8) The residual air will be dephlogisticated air.

This argument will be rendered as:

(M7) Liberation of mercury involves absorption of phlogiston.

∴ (M8) The residual air will be oxygen.

The first inference seems to go through by virtue of the meaning of the term 'dephlogisticated air', while the latter clearly fails. Kitcher intimates that an interpreter translating the term as 'oxygen' will therefore find the phlogiston chemists to be guilty of a logical fallacy.

But closer inspection reveals this not to be the case. For the first inference to go through, one needs the additional premise:

(P6) Dephlogisticated air is the substance which results from removing phlogiston completely from the air.

Once it has been decided to translate 'dephlogisticated air' as 'oxygen', the term is being treated as a simple not a complex expression. Thus, one can no more assume that 'phlogiston' has a proper occurrence in it than one can that 'cat' has a proper occurrence in 'cattle' (to use Quine's famous example). Therefore, (P6) is needed for the inference and if this is translated by (M6), the chemist's error becomes clear: both premises of the argument are false or lacking truth value (although its conclusion happens to be true). Moreover, its formal structure is sound and it is clearly not fallacious. It can be interpreted as follows:

(M6) Oxygen is the substance which results from removing phlogiston completely from the air.

(M7) Liberation of mercury involves absorption of phlogiston.

∴ (M8) The residual air will be oxygen.[7]

Far from committing a logical blunder, Priestley has used false or truth-valueless premises to deduce a true conclusion. That is understandable, given the way we have interpreted his other beliefs. The uniform translation of 'dephlogisticated air' as 'oxygen' stands.

Perhaps Kitcher would object to the translation of (P6) as (M6) on the grounds that it is a definition for the phlogiston theorists. He might say that (P6) should be taken as defining 'dephlogisticated air' as air from which the phlogiston has been removed. Therefore, to render (P6) as (M6) is to rob it of its definitional character; it is to translate an analytic statement into a false synthetic one. That may be so, but it is quite common to construe a scientific theory such that what was once considered definitional is no longer considered so. This is just further proof that definitions are not useful in scientific inquiry. An interpretation of a particular scientific theory need make no effort to preserve purported definitions.

Before dismissing this objection entirely, it may be worth acknowledging the grain of truth in it. What is true is that in an earlier guise, the phlogiston theory would have been interpreted differently. The version focused on by Kitcher (and therefore in this discussion) is the one associated with Priestley and Cavendish. But as Kitcher points out, when Stahl first coined the term 'dephlogisticated air', the only belief associated with the term was that there was a substance that resulted from the absorption by air of phlogiston. When interpreting Stahl, it may be appropriate to construe 'dephlogisticated air' as a complex expression that contains a proper occurrence of the vacuous expression 'phlogiston' (though that judgment would have to be made after looking in greater detail at Stahl's theory). But by the time of Priestley and Cavendish, when more beliefs have come to be associated with the term, the theory has changed and the most plausible interpretation is that 'dephlogisticated air' is a simple expression and should be translated as 'oxygen'. If it were translated as a complex expression throughout, this would greatly

---

[7] Strictly speaking, an innocuous third premise is also needed: Absorption of phlogiston by a substance involves removal of phlogiston from the surrounding air.

decrease the number of shared beliefs without justification and the phlogistonians could not be said to have discovered any of the properties of oxygen. On the basis of the evidence presented, the interpretive approach would recommend translating the term (type) 'dephlogisticated air' as 'oxygen', construing sentences like (P2) as true and sentences like (P6) as false.

There is nothing paradoxical about this situation, since the interpretive approach, being descriptional, will generally not ascribe the same concepts to an agent as more beliefs are acquired by that agent. Translation is not generally constant under belief fixation. In fact, that is one of the advantages of the interpretive approach: its sensitivity to all the agents' beliefs in ascribing concepts to that agent. It is precisely this phenomenon that Kitcher found the (pure) causal theory of reference unable to deal with. Moreover, it would not do to translate 'dephlogisticated air' sometimes as 'oxygen' and other times as a description with an occurrence of a vacuous expression ('air from which phlogiston has been removed'). This can only be warranted if there is some indication that the phlogiston theorists themselves took the expression to be equivocal. Pending such an explicit sign, translation must be type-type and the term must be rendered uniformly.

I will now attempt to justify an assumption made at the beginning of this section, that 'phlogiston' is a term that fails to correspond to any of our terms. On the interpretive approach, a judgment that a term fails to correspond to one of our terms is made when all the most plausible translations of the term are tried and it is found that none of them result in enough truths among the sentences in which the term occurs, and attempts at repairing the situation by compensatory translations also fail. In practice, the vast majority of possible translations can be dismissed out of hand and the likely candidates will typically be few. It is impossible to come up with an algorithm that would avoid the vagueness in these formulations, but the general approach will be given credence by seeing how it applies to the case of 'phlogiston'.

Kitcher states at one point that, for a time, Priestley believed that 'inflammable air' was the same as phlogiston and used the term 'phlogiston' to record its properties. What would the interpretive approach make of the resultant theory? Let us take the simplest case: pretend that the term 'inflammable air' was dropped from usage and that 'phlogiston' was used in its stead in all occurrences. In our previous translation we identified

'inflammable air' with 'hydrogen'. If we translate all occurrences of 'phlogiston' as 'hydrogen' and make the other translations as before, we will obtain two sets of sentences in which the term 'hydrogen' occurs, one of which is associated with the properties of hydrogen and the other of which is associated with the properties of a purported substance that is given off in combustion. Since the first set of sentences will be true and the second false, we will conclude that the phlogistonians correctly identified many of the properties of hydrogen, but erred in thinking that hydrogen is given off in combustion and that all combustible substances are rich in hydrogen. Thus, it is reasonable to say that 'phlogiston' should be translated as 'hydrogen' in this hypothetical theory, though the chemists had some importantly mistaken beliefs about hydrogen (compare the above judgment regarding 'dephlogisticated air' and oxygen). There are other possible theories that are intermediate between this one and the previous one, which would involve retaining the term 'inflammable air' in some occurrences and replacing it with 'phlogiston' in others. Generally speaking, the very fact that the term 'inflammable air' is used in making true statements about hydrogen will militate against translating 'phlogiston' as 'hydrogen', because the use of two terms is a good indicator that the chemists thought they were two separate substances.

For the hypothetical form of the theory described above, it was plausible to translate 'phlogiston' as 'hydrogen', but for the actual version, such a translation is not warranted, because it would greatly decrease the number of true sentences of the theory without justification. The reason is that the properties of hydrogen that were correctly identified by the phlogiston chemists were mostly associated with the term 'inflammable air' in the theory, whereas none of the significant properties of hydrogen were predicated of 'phlogiston'. In the absence of other plausible candidates to serve as translations of 'phlogiston', there is nothing for it but to conclude that 'phlogiston' fails to correspond to any of our terms.

It may be objected that certain outlandish interpretations have not yet been considered. For example, someone might propose to translate 'phlogiston' as 'oxygen' and compensate by translating the two-place predicate 'is emitted to' as 'is absorbed from'. Thus, when the phlogiston theorists assert:

(P9) When wood is burned, phlogiston is emitted to the air.

we should interpret them as saying:

(M9) When wood is burned, oxygen is absorbed from the air.

Since (M9) is a true sentence, we seem to have some reason to believe that 'phlogiston' should be translated as 'oxygen', provided we make the compensatory translations suggested (we may also have to translate 'is rich in' as 'is poor in', etc.). This proposal illustrates the possibility of indeterminacy discussed in the previous chapter, for a disagreement in beliefs has been eliminated by a reconstrual of concepts. The interpreter can take 'phlogiston' not to correspond to any of our terms and can translate 'is emitted to' homophonically, thus finding (P9) false or lacking truth value. But alternatively, one might reconstrue the terms of (P9) as suggested and hand down the verdict that it is true.

However, the above proposal does not wash. The compensating translations are such that they would make a large number of other sentences of the theory false. If the two-place predicates mentioned above and similar ones were all replaced as suggested, that would render false many of the sentences in which phlogiston has no occurrence and which were formerly taken as unproblematically shared between phlogistonians and modern chemists. Moreover, 'dephlogisticated air' could no longer be translated as 'oxygen', unless there was good reason to suppose that the chemists were using two terms to refer to the same substance, and a different translation would falsify many sentences without good reason. This shows how related beliefs can be brought to bear to rule out alternative interpretations of this kind, much in the way they were in Davidson's simple ketch-yawl example in Chapter 3. Perhaps the objector will persist in this line, stating that many other revisions would have to be made in order to make the resultant translation a plausible one. Apart from the fact that this proposal would merely be a promissory note for an alternative interpretation, it should be remembered that in this case, the revisions will have to extend to parts of the language outside chemical theory. The predicates mentioned above ('is emitted by', 'is rich in', etc.) pertain to parts of the language external to chemistry. Therefore, ordinary discourse and other sciences would also be affected by the proposed translations. Wide-ranging revisions that trespass on the borders between theories will only be made as a last resort. Given the plausibility of the alternative, there is no reason to think that such drastic measures should be resorted to in this case. If someone suggests that these predicates be taken to be equivocal, having different

meanings when used in chemical theory and in other parts of discourse, it would invite charges of ad hoc-ism. Unless there is some indication, explicit or implicit, that the phlogistonians used these terms differently in chemical discourse than they did in ordinary contexts, an equivocal translation would be unmotivated.

4.4. Dalton's Theory and Avogadro's Theory

The task of comparing Dalton's atomic theory with Avogadro's (as the latter is represented by Cannizzaro) has been undertaken by Kathryn Pyne Parsons. Although she does not spell out a comprehensive theoretical framework for comparing theories, her explication of the differences can also be used to illustrate the interpretive method being advocated. In fact, she proceeds in a manner that is very congenial to the interpretive approach and I will generally agree with her analysis of the case study. Parsons' procedure is to interpret Dalton's theory in terms of Avogadro's, on the basis of analytic hypotheses that make the first theory consistent throughout and understandable where wrong.

She begins her analysis by noting that appearances are that the two theories are compatible. When one lists some of their central tenets, they do not seem to be in conflict. Altering slightly Parsons' formulation, some of Dalton's central beliefs can be paraphrased as follows:

> (a) An elementary-atom is an ultimate, indivisible particle of a simple (or elementary) substance.
>
> (b) A compound-atom is the ultimate particle of a compound substance.
>
> (c) A molecule is the ultimate, indivisible particle of a simple (or elementary) substance, or the ultimate particle of a compound substance.

Meanwhile, some of Avogadro's central beliefs (given the standard translation from the Italian) are the following:

> (d) An atom is the ultimate particle of a simple (or elementary) substance; indivisible.
>
> (e) A molecule is the ultimate particle of a substance.

Parsons elaborates on the apparent compatibility by saying that a quick glance at these two sets of beliefs suggests that (a) and (d) are consonant and that (e) is just a more general version of (c). However, the latter claim relies crucially on the assumption that 'molecule'

means the same in both theories. But she conjectures that that is not a warranted assumption, since 'molecule' in Avogadro's theory includes oxygen particles in its extension, but it does not in Dalton's. Parsons says that it seems clear from passages of Dalton's A New System of Chemical Philosophy that he sometimes uses 'atom' as Cannizzaro used 'molecule' and not as Cannizzaro used 'atom'. To test this conjecture, Parsons draws up a list of translations that map Dalton's terms onto Cannizzaro's (i.e. Avogadro's, which for these purposes coincide with those of modern chemists). The main ones can be put as follows: 'elementary-atom' is translated as 'atom'; 'compound-atom' is translated as 'molecule'; and 'molecule' is translated as 'ultimate particle' (a neologism).[8] The third hypothesis says that 'molecule' picks out whatever the ultimate particles of a substance are, whether atoms or molecules. But although Parsons does not say this, the problematic hypothesis may seem to be the first one. Someone might object to it on the grounds that the very term 'elementary atom' seems to presuppose that these particles are the ultimate particles of elements. For Dalton also happened to think that elements in their natural state consisted of collections of detached atoms. Of course, we now know that this is not true, since many elements form molecules of one or more atoms, for example oxygen (O2) and hydrogen (H2). So the objector would say that translating 'elementary-atom' as 'atom' fails to capture this aspect of Dalton's theory. But it was argued in the previous chapter that this is a misguided approach to the ascription of concepts. The position cannot be defended without insisting on a definitional approach to specifying the meanings of scientific terms. Unless one holds that the fact that elements exist in their natural state as collections of single, detached atoms is necessarily part of the meaning of Dalton's concept 'elementary-atom', then one cannot object in principle to the translation of 'elementary-atom' as 'atom'. Just as was the case for 'dephlogisticated air' in the previous section, there are two sets of beliefs in which the term 'elementary-atom' features, one of

---

[8] Parsons hyphenates both 'elementary-atom' and 'compound-atom' to emphasize the fact that they are "atomic predicates in the Dalton theory." (1975, 375) This corresponds to the assumption made in the previous section that 'dephlogisticated air' was a simple expression, rather than a complex one. Also, the right-hand side of the third translation should be thought of as a neologism, which might be indicated by hyphenation.

which comes out false when this substitution is made and the other of which comes out true. Among the former is the belief that elements are composed of atoms in their natural state.

The next step, then, is to see how the above hypotheses distribute truth values among Dalton's sentences and how they serve to explain Dalton's errors. To this end, Parsons identifies the following key sentences of Dalton's theory:

(D1) Oxygen molecules are elementary-atoms.

(D2) Elementary-atoms are indivisible.

(D3) Oxygen and hydrogen molecules unite 1-1 in making water vapor.

(D4) One volume of oxygen and two volumes of hydrogen yield two volumes of water vapor (within the limits of experimental error).

On the basis of the proposed translations of key terms, (D1)-(D4) can be translated as follows:

(A1) The ultimate particles of oxygen are atoms.

(A2) Atoms are indivisible.

(A3) The ultimate particles of oxygen and hydrogen unite 1-1 in making water vapor.

(A4) One volume of oxygen and two volumes of hydrogen yield two volumes of water vapor (within the limits of experimental error).

From the viewpoint of the later theory (and of modern chemistry), (A1) and (A3) are false, while (A2) and (A4) are true.

After translating Dalton's beliefs in this fashion and assigning them the truth-values indicated, Parsons says that we must explain why Dalton makes these assertions. She does this by looking at his other beliefs. In consonance with the interpretive approach, she holds that Dalton's true assertions as well as his false ones must be explained and they must be seen to cohere with his other beliefs. Of the true sentences, (D4) is the most straightforward and does not contain any problematic terms. Dalton's reason for holding it is just that it was an established experimental fact in his time. As for (D2), Parsons says that Dalton had two reasons for holding it. The first is that he thought that chemical compounds have constant composition and that chemical elements combine in certain ratios and form a small, discrete number of compounds. This led him to believe that there

is a particulate basis for matter. Moreover, since he thought that the particles of simples should be taken to be simple, he held that the particles of elements should be simple, that is, indivisible. The second reason Dalton held (D2) is that he believed that if particles of elements were divisible, they should be composed of two or more particles of the same substance. But in agreement with many contemporary scientists, Dalton held that like repels like.[9] On the basis of this, Dalton held that atoms were indivisible (i.e. not decomposable by chemical means).

As for Dalton's reasons for (D1), the claim follows almost immediately from his reasons for (D2) when one adds Dalton's belief that experimental evidence indicated that oxygen was an element. Since he held that the ultimate particles of elements were atoms, he deduced that the ultimate particles of oxygen must be atoms. When one comes to (D3), however, it appears to be the most problematic of the beliefs we have attributed to Dalton. Because of (D1) and because he thought a similar principle held for hydrogen, Dalton thought that oxygen and hydrogen should unite 1-1. This is based on another principle of Dalton's theory: where two elementary substances form only one compound, that compound is assumed to be binary, combining 1-1.[10] But there seems to be a tension between (D3) and (D4): If oxygen and hydrogen particles combine in a 1-1 ratio, how can Dalton also believe that their <u>volumes</u> mix in a 2-1 ratio? This is especially puzzling when one notices that (D3) was precisely one of the facts that led Avogadro to conclude that equal numbers of molecules of a substance occupy equal volumes.

However, Dalton was not contradicting himself. In fact, he avoided contradiction by postulating that atoms of different elements have different sizes. This allows (D3) to be consistent with (D4) if one also specifies that there are approximately twice as many oxygen particles in one volume as there are hydrogen molecules in one volume, that is that hydrogen molecules have twice the volume of oxygen particles. Now, one might think that this is an irrational move on Dalton's part. Why should the ratio of the volume of hydrogen

---

[9] This is confirmed by Partington in his history of chemistry. He writes: "Dalton... thought (with Newton) that atoms of the <u>same</u> element repel one another..." (1937, 170)

[10] See also Partington's exposition in (1937, 169-70).

to oxygen particles be exactly 2-1 so as to give rise to the right volume ratio? This does not seem so strange if one bears in mind that Dalton did not think that the 2-1 volume ratio held exactly. As indicated parenthetically in (D4), he thought this was an approximate measurement. Parsons quotes him as challenging the exact 2-1 ratio on the grounds that "the most exact experiments I have ever made, gave 1.97 hydrogen to 1 oxygen." Now this may have been ad hoc on Dalton's part, but it is not irrational. It indicates that he did not believe in what is known as the Gay-Lussac law. That says that when gases combine, the volume of the resulting gas is proportionate to that of one of the original gases. In other words, what Dalton is denying is the belief, widely held at the time, that the volumes of gases are indicators of the numbers of particles they contain. If one agreed with Dalton on this point, it would be reasonable to hold both (D3) and (D4).

Notwithstanding the fact that she adopts a slightly different methodology of interpretation than the one followed in the two previous sections, Parsons adheres to the same basic interpretive method outlined in Chapter 3. Rather than fasten first on a set of beliefs that Dalton and Avogadro are thought to have shared, Parsons almost immediately proposes a set of term-by-term translations for Dalton's terms. However, she goes on to corroborate these on the basis of beliefs that Dalton held uncontroversially, in other words, those beliefs not containing the problematic terms 'molecule', 'elementary atom', and so on. These other beliefs, ones we have good reason to think Dalton held, are of two kinds: (i) those he shared with most chemists of his time, such as (D4), and (ii) those beliefs he held despite the opinion of many of his contemporaries, such as the denial of the Gay-Lussac law. Of course, in making these claims, we assume that such beliefs can be interpreted in a straightforward fashion; in this case, we assume that terms such as 'volume', 'equal', 'combine', and so on, are translated homophonically. But as argued in the previous section, these are the kinds of terms that we do not have any leeway in reinterpreting since they do not pertain only to the contentious domain. Hence, any reconstrual of them would involve disastrous spillover into other domains. By adhering to the interpretive method, a coherent account of Dalton's beliefs can be given in Avogadro's terms.

4.5. Aristotle's Theory and Galileo's Theory

In his essay, "A Function for Thought Experiments," Thomas Kuhn has discussed Aristotle's concept of speed or velocity.[11] It is not entirely clear from this paper whether he considers him to have the same concept as later Galilean and Newtonian theorists or whether the concept is supposed to be different. Indeed, he never even cites this as an example of what he means by conceptual incommensurability, although this suggestion might be implicit. I will argue in this section that there is no doubt about how to interpret Aristotle and will interpret his theory in terms of that of his successors.

Kuhn begins by claiming that Aristotle's Physics and the tradition that descends from it show some evidence of two disparate criteria used in discussions of speed. The first, he explains, yields a concept that "is very like what we should call 'average speed'." (1964, 246) As illustrative of this notion, he cites the following passage: "The quicker of two things traverses a greater magnitude in an equal time, an equal magnitude in less time, and a greater magnitude in less time." (1930, 232a25-27)[12] However, Kuhn also notes that at other points "...Aristotle is grasping directly, and perhaps perceptually, an aspect of motion which we should describe as 'instantaneous velocity' and which has properties quite different from average velocity." (1964, 247) As an illustration, Kuhn cites a passage in which he seems to be attending to what Kuhn calls the "perceptual blurriness" of a body in motion: "But whereas the velocity of that which comes to a standstill seems always to increase, the velocity of that which is carried violently seems always to decrease." (1930, 230b23-25) Kuhn goes on to comment on the situation, as follows:

> ...Aristotle's concept of speed, with its two simultaneous criteria, can be applied without difficulty to most of the motions we see about us. Problems arise only for

---

[11] Although Kuhn seems to use the terms 'speed' and 'velocity' indiscriminately, I will put things in terms of velocity since it is clear that Aristotle thought that the direction of velocity mattered when velocities were being added, for example. But further justification of this claim would require a separate discussion.

[12] I am following Kuhn in using the Hardie and Gaye translation of the Physics (in the Ross edition of the collected works) and will begin by reporting Aristotle's beliefs in terms of the English translation.

> that class of motions, ...rather rare, in which the criterion of instantaneous velocity and the criterion of average velocity lead to contradictory responses in qualitative applications. (1964, 254)

Kuhn's remarks suggest the following interpretation of Aristotle's theory of motion. He was employing two criteria for assessing velocity that did not coincide: 1) the ratio of total distance travelled by a body to time elapsed, and 2) perceptual blurriness of a body in motion. Luckily for him, he did not knowingly come across motions that would have showed that the criteria diverged, so the difficulty was never brought to the fore. As Kuhn states: "In a world of that sort [i.e. one in which all motions were uniform] the Aristotelian concept of speed could never be jeopardized by an actual physical situation, for the instantaneous and average speed of any motion would always be the same." (1964, 254) He goes on to imagine a scientist embedded in such a world and says: "...given our broader experience and our correspondingly richer conceptual apparatus, we would likely say that, consciously or unconsciously, he had embodied in his concept of speed his expectation that only uniform motions would occur in his world." (1964, 255) Therefore, Kuhn is saying that Aristotle assumed that no non-uniform motions existed or that all motions were uniform.

There is trouble for Kuhn's interpretation, however, when one looks at Aristotle's Physics. One finds there that he clearly made a distinction between uniform and non-uniform motion and that he held that both kinds of motion were possible. That makes it hard to maintain that Aristotle believed that all motions were uniform, whether explicitly or implicitly. For instance, Aristotle writes: "Sometimes it [i.e. regularity or irregularity] is found neither in the place nor in the time nor in the goal but in the manner of the motion: for in some cases the motion is differentiated by quickness and slowness: thus if its velocity is uniform a motion is regular, if not it is irregular." (228b25-27) To be sure, Kuhn notes at one point that it is not strictly necessary to posit that Aristotle assumed that all motions were uniform in order to explain why his two criteria never conflicted. Rather, all that is needed is for a weaker condition to obtain: "The requisite weaker condition is that no body which is 'slower' by either criterion shall ever overtake a 'faster' body." (1964, 254) However, I will argue that there is no need to attribute either of these beliefs to Aristotle,

for neither is necessary to make sense of his theory and at least one conflicts directly with the textual evidence.

Let us recapitulate by listing some of Aristotle's most salient beliefs:

(a) The quicker of two things (i.e. the one with greater velocity) traverses a greater magnitude in an equal or less time.

(b) The quicker of two things (i.e. the one with greater velocity) traverses an equal magnitude in less time.

(c) The velocity of that which comes to a standstill seems always to increase.

(d) The velocity of that which is carried violently seems always to decrease.

Next, let us try to give an interpretation of Aristotle's theory in our terms. First, introduce the following proposal: 'velocity' is translated as 'average-velocity'.[13] This enables us to list the following tenets of Aristotle's theory in terms of later (Galilean) physical theory:

(G1) A body with greater average-velocity traverses the same distance in an equal time.

(G2) A body with greater average-velocity is more perceptually blurry.

If we take (G1) to be true (although not an analytic truth or a definition of the concept of average-velocity), (G2) will be considered false. That is simply because (G1) and (G2) together yield the following claim:

(G3) A body that traverses the same distance in less time is more perceptual blurry.

But it is not true that for all motions, a body that traverses the same distance in less time is one that is more perceptually blurry at <u>every</u> instant of its motion. Thus, assuming the truth of (G1) leads to the falsity of (G2). Since (G3) is false by our lights, we should ask ourselves why Aristotle believed it. According to Kuhn, it can be explained by the following implicit belief:

(G4) All motions are uniform.

---

[13] Note that this should not be thought of as a composite concept, decomposable into the concepts of average and of (instantaneous) velocity but a unitary concept, as I will explain below.

However, we have just seen that Aristotle held no such belief, whether implicitly or explicitly. So, if we cannot attribute (G4), we might ask about the wisdom of attributing (G3), or for that matter, (G2). The last two beliefs seem to presuppose that Aristotle took perceptual blurriness to be a measure of velocity at an instant. Since, as we shall soon see, Aristotle eschewed the idea of instantaneous motion, it would be more plausible to attribute a different belief:

> (G5) A body with greater average-velocity during an interval is more perceptually blurry during that interval.

This is a belief that we can agree is a reasonable perceptual approximation over certain limited intervals, such as the ones mentioned in (c) and (d) above, namely at the beginnings and ends of certain motions.

But Kuhn would probably object to this interpretation (which rejects (G2), (G3) and (G4), in favour of (G1) and (G5)). As I have mentioned, Kuhn's remarks reveal a certain hesitation between attributing to Aristotle the concept of average velocity and that of instantaneous velocity. Therefore, he might prefer to adopt a different translation based on the second criterion that he mentions. Since that criterion is supposed to pertain to instantaneous velocity, and since the concept of instantaneous velocity is just that of our concept of velocity tout court, the following proposal might be ventured: 'velocity' is translated as 'velocity'. On the basis of this new translation, we should modify the interpretation of the above tenets of Aristotle's theory as follows:

> (G1') A body with greater velocity traverses the same distance in less time.
>
> (G2') A body with greater velocity is more perceptually blurry.
>
> (G3') A body that traverses the same distance in less time is more perceptually blurry.

Again, we know that (G3') is false by our lights, but this time, we take (G2') as true and (G1') as false (for the same reasons discussed above: it might have attained a greater instantaneous velocity at some points in its motion, yet still have a smaller average velocity). Thus, (G1') is now false and (G2') true, while (G3') remains identical to (G3) in content and truth value, since neither contains the problematic concept.

The question now is, what makes the second interpretation less convincing than the first? I would argue that there are two important objections to this interpretation. First, if

Aristotle had the concept of instantaneous velocity that is associated with Kuhn's second criterion, we again have no way of explaining the false belief (G3') unless we posit that Aristotle believed that all motions were uniform. However, we have already seen that Aristotle did not believe this, based on textual considerations. The second objection to this interpretation also relies on evidence from the text. Commentators on the Physics have noted that Aristotle regards it as impossible that a body should be in motion at an instant of time. He seems to have thought that since motion can only occur during an interval with non-zero duration, motion (and, for that matter, rest) during an instant is an absurdity.[14] Thus, it becomes impossible to attribute to him the concept of instantaneous velocity--or even a criterion of perceptual blurriness at an instant.

There is nothing in the text to suggest that the criterion of perceptual blurriness should not be regarded as a measure of average-velocity, albeit over relatively short intervals of time. When understood thus, Aristotle need not have assumed that all motions were uniform or even that no bodies that achieved greater perceptual blurriness would ever be overtaken by ones that were less perceptually blurry. That is simply because the remarks that Kuhn takes to be about perceptual blurriness specify no precise way of assessing such a thing. We do not know what Aristotle would say about a body that was very blurry at one point in its motion and less so at another point; perhaps the blurriness would need to be averaged out over the whole motion. Clearly, Aristotle's employment of this criterion was quite restricted and he makes no move to use it quantitatively, as he does for the first criterion.

Given the available evidence, it is unlikely that Aristotle envisaged applying the second criterion in the way that Kuhn posits as a measure of instantaneous velocity. By contrast, Aristotle's first criterion for determining average-velocity (ratio of distance travelled to time elapsed) is quite adequate to measure a well-defined quantity in a consistent fashion. Therefore, ascribing a concept to Aristotle based on the criterion of

---

[14] He writes in the Physics: "We will now show that nothing can be in motion in a present. For if this is possible, there can be both quicker and slower motion in the present." (1930, 234a24-25) For a fuller discussion, see Bostock (1991) and Hussey (1991).

perceptual blurriness is not warranted by the evidence. One can speculate, of course, about how we would have interpreted Aristotle had he possessed a more precise criterion for measuring instantaneous velocity. We might have had more of a problem in deciding which interpretive hypothesis to select, but the problem would have been no different in principle from the problem that confronted us in translating Newton's term 'mass'. The choice would have depended on the same sorts of considerations; indeed, it might have turned out that neither hypothesis was adequate and we might have resorted to a neologism.

At more than one point, Kuhn discounts the possibility of ascribing a "self-contradictory" concept to Aristotle, saying that his concept of velocity is nothing like the logician's example of a self-contradictory concept, the square-circle. (1964, 253-254) However, he also makes some claims that lend credence to such an interpretation. One such judgement is the following: "Aristotle's concept of speed, in which something like the separate modern concepts of average and instantaneous speed were merged, was an integral part of his entire theory of motion and had implications for the whole of his physics." (1964, 257; emphasis added) On the basis of this claim, someone might propose a translation that would substitute for 'velocity' sometimes 'average-velocity' and sometimes 'velocity', depending on the criterion being presupposed at that particular point in the argument. This would seem to capture the presence of two criteria in Aristotle's theory. However, even if the second criterion were sharper in the text, such a solution could not be supported on the approach being advocated here, since the number of concepts ascribed should equal the number of concepts possessed by the interpretee. Thus, using two terms to translate an agent's single term is only warranted when that agent either implicitly or explicitly considers the term to be equivocal (a claim which will be defended in section 5.3.). An example of a truly equivocal term in a contemporary scientist's lexicon would be 'resistance', which is intentionally used to mean one thing in electromagnetic theory and another in mechanics, though that may not be explicitly indicated by its users. The claim being made here is that 'velocity' for Aristotle is not such a term. Similarly, Kuhn does not express things quite correctly, from the present point of view, when he says that we should not, properly speaking, say that Aristotle's concept of speed is confused, but that, "We may, of course, say that it was 'wrong' or 'false' in the same sense that we apply those terms to

outmoded laws and theories." (1964, 258) We should not even say concerning concepts that they were wrong or false. Rather than say that the concept itself was wrong or false, we should say that the theory was wrong or false at certain points.

Finally, a point that is worth stressing concerns the issue of whether we have (or perhaps Galilean physicists had) the concept of average-velocity. It may not be apparent whether we have such a concept that is not just a concatenation of the concepts of average and (instantaneous) velocity. But although it may not be clear in English, we do have the concept of average velocity over an interval, which is expressed in mathematical English by a 'v' with a bar on top of it (pronounced vee-bar), and is given by the ratio $\partial x/\partial t$, rather than the derivative dx/dt. Thus, we have the concept, but in ordinary English I have translated it as 'average-velocity' to emphasize its unitary character and to avoid suggesting that it was decomposable into two components.[15]

4.6. Concepts from Social and Political Theory

When it comes to theorizing about social and political matters as opposed to the natural realm, the task of theory-comparison is more complicated. Social and political theories are seldom as systematically arrayed as those in the natural sciences and the logical structure is usually not as explicit. Nevertheless, by drawing on examples from more than one theory, I will argue in this section that the method of comparison is in some cases the same. Fortunately, intellectual historians and historians of philosophy have been giving increasing thought to questions of the interpretation of past social theorists and philosophers. Therefore, some of their work can be relied upon for guidance in this task,

---

[15] In a discussion of Aristotle's mechanics, G.E.L. Owen notes correctly that 'over-all velocity' is preferable to 'average velocity', since the latter "is a function of instantaneous speed not available to Aristotle." (1986, 315) There is considerable debate over the issue of whether Aristotle had laws of motion or even a science of mechanics. I have tried to avoid becoming entangled in that debate by focusing on a single rather simple concept. It would have been more interesting (and difficult) to tackle Aristotle's concept dunamis, which some interpret as 'force' and others as 'power'. For more on this and related issues, see Owen (1986) and references therein.

much in the way that philosophers and historians of science were used in previous sections.

Quentin Skinner is an intellectual historian who has given close attention to these matters and is one of a few writers who have made explicit the methodology he uses for the interpretation of historical figures. Although Skinner's methodology differs from the one being advocated here, chiefly in its emphasis on speech-act theory and an apparent attachment to a definitional approach, many of his insights can be captured by the present interpretive approach. Some of the principles he sets out from are consonant with the ones advocated in this and the previous chapter. Skinner first outlines specific "requirements" for the identification and understanding of concepts featured in social and political theories (particularly evaluative or appraisive concepts). At first sight, his requirements appear not to be in conformity with the guidelines specified in the previous chapter. However, I will argue that they can be made to agree without too much modification. The first requirement is that "it is necessary to know the criteria in virtue of which the word or expression is generally employed." (1980, 121) Second, he states that one needs "to have a clear sense of the nature of the circumstances in which the word can properly be used to designate particular actions or states of affairs." (1980, 122) Finally, in the case of appraisive terms in particular, "We need in addition to know what exact range of attitudes the term can standardly be used to express." (1980, 122) He restates this as the requirement that "it is necessary to know what type of speech-acts the word can be used to perform." (1980, 122)

Some, but not all, of the above insights can be coopted by the interpretive approach. Against Skinner, assume that the interpretation of appraisive terms is in principle no different from that of non-appraisive ones. On this assumption (to be justified below), Skinner's first and third requirements are nothing but ways of identifying certain basic beliefs in which a term enters. That is evident when one considers his illustrative examples. In the case of the first requirement, he gives the example of the term 'courageous' and goes on to list some of the "criteria" in virtue of which it is employed. He states:

> ...the word can only be used in the context of voluntary actions; ...the actor involved must have faced some danger; ...he must have faced it with some consciousness of

> its nature; and he must have faced it heedfully with some sense of the probable consequences of the action involved. (1980, 121-2)

This seems to be simply a list of general beliefs in which the word is featured, thus, 'All courageous actions are voluntary ones,' 'A courageous actor must have faced some danger,' 'A courageous actor must be conscious of the nature of danger faced,' and so on.

Skinner's second requirement can also be seen to fit well into the methodology of interpretation proposed here. He states that one must know the "range of reference" of a term and explains it thus:

> Once I have acquired this understanding [of the correct use of a word], I may expect in consequence to be able to exercise the further and more mysterious skill of relating the word to the world. I may expect, for example, to be able to pick out just those actions which are properly to be called courageous, and to discuss the sort of circumstances in which we might wish to apply that particular description... (1980, 122)

This is, properly speaking, the extension of the word in question. It has already been seen how the interpretive approach attends to what an agent actually picks out using a term as part of the evidence in interpreting the agent's term. This was clear in the example of the phlogiston theory in section 4.3., where the laboratory reports of the phlogiston theorists served as a guide to interpretation.

Skinner's third criterion has to do with the appraisive force of the terms he is interested in. He illustrates it again with the help of the term 'courageous': "no one can be said to have grasped the correct application of the adjective courageous if they remain unaware that it is standardly used to commend, express approval, and especially to express (and solicit) admiration for any action it is used to describe." (1980, 122) But I would argue that this is assimilable to the first criterion and can be seen as a specification of certain key beliefs in which the term appears, thus, 'A courageous action is generally commendable,' 'A courageous actor is admirable,' and so on. Skinner would seem to differ with this reconstrual, on the grounds that a specification of illocutionary force cannot be encapsulated in such sentences. However, as usually employed, the notion of illocutionary force is considered to be a part of pragmatics rather than semantics and unless Skinner can show that the appraisive force of terms is not part of the literal meaning of such terms, he

will not have made a good case for attending to speech-acts performed using such terms. This claim will be justified furhter in section 5.8., where I will explain how I assume one should draw the line between semantics and pragmatics.

Skinner's unusual use of speech-act theory is evident in his disagreement with one of the foremost speech-act theorists, John Searle. In a footnote, Skinner disputes whether Searle has succeeded in showing that meaning and speech-acts are wholly separate: "Depending on one's view of meaning, one might still want to insist that speech-act potential is part of meaning, even if it is distinct from both sense and reference." (1980, 313) Those who consider speech-act potential to be distinct from sense and reference usually take it to pertain to pragmatics, that is, to account for some of the things that can be done with words that are not properly part of their literal meaning. For the most part, speech acts are supposed to involve those aspects of usage that do not involve truth conditions, by contrast with literal meaning.[16]

The idiosyncracy of Skinner's view also comes out in the assumption that the appraisive aspects of a word's usage are <u>always</u> to be subsumed under the heading of illocutionary force, not meaning. In fact, however, there is nothing to prevent praise, blame, commendation, and condemnation from being conveyed semantically, as part of the standard meaning of a term. Just because words are <u>sometimes</u> used to praise or blame independently of their semantic value, that does not mean that evaluative judgments <u>always</u> issue from the force of the words uttered. An example might make this clearer. If I say to someone who has just shoplifted a book:

I could never do such a thing.

and I say it in a suitably stern tone of voice, perhaps with downcast eyes and a slight shake of the head, it can be assumed that I am condemning the act of kleptomania. But if the same

---

[16] J.L. Austin's own view in <u>How to Do Things with Words</u> is somewhat more problematic. At some points, he concedes that literal assertions have a privileged position, but at other times he seems to deny the semantic-pragmatic distinction by saying that stating or asserting are just two among many illocutionary acts (see especially Lectures XI and XII of his (1975)). Thus, Skinner's view may accord with Austin's, though it is a minority position among philosophers of language.

words were uttered wistfully and with wide eyes to someone who has just tried skydiving for the first time, it is safe to say that there is an element of admiration in my attitude. In both cases, the words themselves do not contain the appraisive force of the utterance and one has to go beyond their meaning to appreciate what was said.[17] In short, the utterance is not a mere statement about the speaker's dispositions to refrain from acting in certain ways (shoplifting, skydiving). It is not clear that all, or even most, of the texts and statements that Skinner discusses must be viewed in this light. Appraisive statements that feature in a political or social theory do not necessarily require an appeal to non-literal force and the comparison of theories about the social world need not be different in principle from such comparisons in the natural sciences. The fact that the former are not value-neutral should not be an obstacle to their being interpreted in much the way proposed for scientific theories.

A more basic point of disagreement with Skinner's method will emerge by way of an example. Skinner writes:

> Consider, for example, the special use of the term religious that first emerged in the later sixteenth century as a way of commending punctual, strict, and conscientious forms of behaviour. The aim was clearly to suggest that the ordinary criteria for applying the strikingly commendatory term religious could be found in such actions, and thus that such actions themselves could be seen essentially as acts of piety and not merely as instances of administrative competence. The failure of this move was quickly reflected in the emergence of a new meaning for the term religious in the course of the seventeenth century--the meaning we still invoke when we say things like 'I attend the meetings of my Department religiously'.[18] (1980, 126-7)

---

[17] In this example, the non-literal meaning would be a conversational implicature rather than an illocutionary force, but the general point remains.

[18] Incidentally, there is no suggestion that commendation was anything but part of the standard meaning of 'religious' in this example. In late sixteenth century Europe, this seems uncontroversial and there is good reason for saying that the sentence, 'That act is a religious one,'was then a commendatory one, in almost any context. Thus, one can understand this example without appealing to speech-act theory.

If one were to explain Skinner's example on the interpretive approach one would say that those who used the term 'religious' in the way indicated proposed a new theory of the social world. According to the new theory, the concept religious was to pick out some actions that did not involve praying, reading the bible, going to church, and so on, on the assumption that the latter acts yet had something strongly in common with the 'new' kind of religious action. It was supposed to be more useful to group all these actions under a single concept. However, this is a good instance of a taxonomy that is not supported by the evidence and is eventually abandoned. In this case, the new theory was rejected but, just to complicate matters, it survived in the form of words used. That is, rather than abandon the word 'religious' in such contexts altogether, the linguistic community began using the term equivocally. As Skinner points out, this became a case of "genuine polysemy".

To put it differently, the above example is one in which a theory change is proposed but is rejected in favor of a meaning change. Rather than accept a theory which allowed all kinds of new actions to be described as 'religious', the linguistic community adopted two different concepts of 'religious', as it were, 'religious1' and 'religious2'. Skinner's diagnosis of why this happened is that most language users failed "to see that the ordinary criteria for religious (including the notion of piety) were in fact present in all the circumstances in which the term was by then beginning to be used." (1980, 127) In my terms, this means that some of the central tenets in which the term 'religious' occurred were not true of the new class of actions that the term was being proposed to cover. But it is not just, as Skinner implies, that actions involving administrative competence are not usefully dubbed 'pious'. That would beg the question: why not extend the application of the term 'pious' so that it comes to include such actions? The reason is that no explanatory value is added and many statements are rendered false when actions involving administrative competence are associated with those involving worship and piety. For example, many people in the late sixteenth century presumably did not believe that being punctual and strict would help them go to heaven. And if the proposed theory-change were made, the following generalization could no longer be made: performing religious actions will help one go to heaven.

Therefore, the main difference between Skinner's method of comparing theories and the one advocated in the present work concerns his criterial approach or apparent commitment to definitional truths. From his pronouncements on the above case, Skinner seems to think that piety is part of the meaning of the term 'religious' and that that is why it was a case of meaning-change to apply it in the way he describes. However, even though it turned out in this case that this was part of the rationale for saying that the new usage constituted a change of meaning, it would be a mistake to say that 'All religious acts are pious' is a definition or that there is a fixed criterion for the application of the term. On the interpretive approach, it is possible after the two theories have been compared to take all the sentences that contain a particular term and are agreed upon between the two theories and say that these sentences give the meaning of that term. However, as argued in Chapter 3, it would be misleading if all the agreed upon sentences are taken as the definition of the term. Even if there were some way to draw up a finite list of such sentences, this definitional approach is not satisfactory because it might be possible to imagine a third theory that shared the associated concept and yet disagreed with any one of the tenets mentioned. This should become clearer in section 6.5., which deals with the failure of transitivity in the ascription of concepts.

The above is a simplified account of the kinds of considerations that enter into a decision to bring a new class of actions under a certain concept, and also, the reasons for rejecting such a move. Once the new theory has been rejected in its original form, all that remains of it is the non-standard or novel use of a certain word. Although the new applications have been rejected, the word continues to be applied to the relevant class of actions and it becomes equivocal (not ambiguous). Other examples that Skinner uses of the same phenomenon are the terms 'literature' and 'philosophy', which were similarly used in "a crude attempt... to link the activities of commercial society with a range of 'higher' values." (1980, 127) The proponents of industrial capitalism tried to give legitimacy to some of their activities by taking these words and applying them to aspects of commercial activity.[19] Again, what actually occurred was that the terms became (and remain)

---

[19] Skinner derives these examples partly from Raymond Williams' Keywords. Under the entry for "philosophy", Williams writes that "the increasing use of philosophy in

equivocal, as when one talks about the "philosophy of General Motors" or the "promotional literature of American Express". But that is not because there is a unique set of core beliefs that is attached to these terms and constitutes a meaning-giving criterion. It may be said that there is no set of criteria that would characterize the writings of Democritus, Averroes, and Russell, but we should not say that the meaning of the term 'philosophy' is equivocal when applied to all three, in the way that it is when we apply it to the corporate principles of G.M.

There might seem to be a sense in which these cases are not assimilable to those that occur in science. It may be said that what is at stake is not just the making of true and false statements about social reality, but the transformation of that reality. Those who proposed the new theory of religiosity did not merely have a new theory, they intended to make people behave in certain ways and to change the character of their society. But ultimately such changes must be based on reasons if they are not to be rejected, as Skinner says all these theories were. In the case of the term 'religious', justifying the theory would have involved showing that punctual and other actions should indeed be regarded in the same light as prayer, church-going, and so on. This might have been done by pointing to passages in the Bible that lay stress on punctuality, or it might have been shown that individuals traditionally considered to be models of religious piety (saints, for example) exhibited the requisite qualities of administrative competence. It is presumably the failure of such a justificatory effort that defeated the new theory for most language-users. Moreover, in its broadest outlines, this effort is akin to that which goes into justifying a scientific theory.

These examples also show that sharing a concept is not simply a matter of using the same term. The abuse of terms can be spotted and exposed as illegitimate, and to the

---

managerial and bureaucratic talk, where philosophy can mean general policy but as often simply the internal assumptions or even the internal procedures of a business or institution... can be traced back to Ure's Philosophy of Manufactures (1835) but in [the mid-twentieth century] it became very much more widespread, as a dignified name for a local line." (1976, 198) So widespread that, today, an envelope containing an airline ticket reads: "This contains more than your ticket. It contains our philosophy."

extent that the term-using practice survives, it succeeds only in making a term equivocal or polysemous. Even in the realm of social explanation, such 'coercive' uses of terms do not take hold because they do not have the same explanatory utility, cannot be used to make appropriate generalizations, and do not exhibit the same connections to other concepts. In the case of political and social theory, this provides us with a reason for interpreting them differently or, what comes to the same thing, ruling that a change of meaning has transpired.

The task of this chapter has been to illustrate and amplify the claims of the previous chapter. It has also tried to fill in some of the detail that is necessary before the interpretive approach can serve as the basis of a method for comparing theories. Implicit in this effort has been a set of interpretive principles, which are the crucial elements that some philosophers of science have found lacking in the interpretive approach and that seemed to disqualify it from being used in the cause of theory choice. These interpretive principles or rules will be made explicit in the next chapter, where some of the more controversial claims made in this chapter will be explained and defended further.

Chapter 5: **Principles**

> [Descartes' rules of method are] like the precepts of some chemist; take what you need and do what you should, and you will get what you want.
>
> G.W. Leibniz

5.1. Reflective Equilibrium

There may appear to be something paradoxical about the order of presentation being followed in this work. In the previous chapter, certain interpretive case studies were tackled without the benefit of explicit principles of interpretation, which principles are only to be explicitly presented in this chapter. It might seem more astute, if not more honest, to proceed in the opposite fashion: outline the principles first, then apply them to specific cases. But things have not proceeded quite as I just characterized them. The general interpretive framework being applied was outlined before the case studies, in Chapter 3, and some of the principles to be used were already prefigured there. The specific principles were not given in full because it is difficult to justify them without adverting to actual cases and that cannot be done until such cases have been discussed at some length. At any rate, it seems reasonable to proceed in this kind of philosophical inquiry according to a version of John Rawls' "reflective equilibrium".[1] One operates with some tentative principles, one tries them out on some cases, then one goes back to tinker with the principles, and so on. The order of presentation here does not quite conform to reflective equilibrium because this is obviously a reconstruction after things have already been worked out backstage, but something of the flavor of reflective equilibrium can be preserved in the presentation.

Another preliminary point: What is the status of these principles? Are they part of the very concept of interpretation? Are they normative ideals that state how agents should interpret one another? Or are they empirical hypotheses about the way that agents

---

[1] See Rawls (1971, 20-21).

actually interpret one another? They certainly do not conform to the first description, which implies that the principles are definitionally built in to the very concept of interpretation. That would imply that there are such things as conceptually necessary truths about meaning or interpretation, which is not the claim that I am making (especially since they are particularly suited to the interpretation of a particular kind of discourse, scientific discourse). Rather, the claim is that these principles of interpretation are something in between descriptive and normative claims. They cannot be too far removed from the actual practice of interpreters, be they scientists, historians of science, intellectual historians, decision theorists, cognitive scientists, folk psychologists, and others. On the other hand, they will also carry with them certain reforms in our practice. These reforms will be justified by the contention that doing things differently enables us to do better what we want to do, namely to understand the actions and utterances of agents who hold different theories, to interpret theories in historical perspective, to preserve informational content, and to choose between scientific theories. If the following principles sometimes seem to be engineered to fit certain of the examples discussed and if some interpretive practices that do not conform to the principles seem to be dismissed instead of being incorporated, that is because this is not a purely empirical enterprise. The appearance of "curve-fitting" is bound to afflict a partly normative and partly descriptive enterprise of this sort.

It should also be borne in mind that interpretation is not a precise art. I mentioned in section 3.3. that some philosophers have cast doubt on the possibility of coming up with a strictly formulated maxim of rationality or reasonableness. For example, Haugeland has said that the notion of making reasonable sense under an interpretation may not be definable with precision. Likewise, it is difficult to formulate a set of interpretive rules with the exactness of a chemist or a navigator. Some of the following principles are particularly applicable in interpreting scientific discourse; others have wider scope and may be thought of as a tentative and simplified foundation for a more general theory of interpretation. In either case, we should not expect them to be capable of foolproof phrasing and ironclad articulation.

5.2. Principle of Conceptual Charity

In Chapter 3, Davidson's well-known Principle of Charity was encountered and briefly explicated. That principle asks generally (and rather vaguely) for the interpreter to attribute, whenever possible, what it would be rational to believe in a given situation. It therefore cautions against multiplying implausible hypotheses about what the interpretee believes, which in turn amounts to warning the interpreter not to impute a wide range of false beliefs (by the interpreter's lights). In early formulations of the principle, this advice was summarized by saying that the interpreter's job is to maximize agreement with the interpretee, or equivalently, to maximize truth (again, from the interpreter's point of view). If this principle is followed, it will end up being the case that the interpreter and interpretee will come out agreeing on most things at the end of the day. If this seems like a counter-intuitive result, Davidson reminds us that disagreement can only occur against a backdrop of agreement. Hence, for every apparently serious difference of opinion among two rational agents, there will always be widespread agreeement in the background. That is because every disagreement at least presupposes a common subject matter about which the disagreement takes place. What is perhaps even more surprising at first is that this consequence of the Principle of Charity also leads to the conclusion that most of every agent's beliefs are true (from any point of view). This follows because, if any two parties will come out mostly agreeing after the process of interpretation is complete, then imagine a special case in which one of the two parties is the omniscient interpreter, all of whose beliefs are true. If <u>any</u> agent would mostly agree with the omniscient intepreter, then most of every agent's beliefs are true.

In section 3.2., I mentioned that the advice given by the Principle of Charity ("maximize agreement") seems sound enough but rather vague. By contrast, the result that is supposed to follow from the adoption of the Principle ("any two agents will share most beliefs") is not so much vague as objection-prone. The objections to this interpretive result can be resolved into two main ones. First, since an agent's beliefs are potentially infinite and we have no uncontroversial way of counting beliefs, it is not clear what it would mean for the two parties to share <u>most</u> beliefs. Second, there seems to be something suspect about this claim, in that it appears to derive a result about the extent of agreement between actual agents from a general principle about the nature of interpretation.

But there is a principle which is a slight modification of Davidson's principle which is less vague and whose consequences are free from both objections. The Principle of Conceptual Charity calls on the interpreter to maximize agreement in concepts rather than beliefs. Accordingly, it has the consequence that the interpreter and interpretee will share most concepts. This principle is obviously in keeping with the general import of Davidson's principle, which is based on the idea that disagreement can only occur against a background of agreement. Given the inextricability of meaning and belief (or equivalently, concept and theory), as well as the rehabilitation of concepts (as explained in section 3.6.), this principle follows directly from Davidson's Principle of Charity. But the modified principle escapes both objections to Davidson's principle. The first objection is avoided because concepts, unlike beliefs, are always finite in number if a language is to be learnable. Moreover, there is an uncontroversial way of counting concepts, since they are equal to the number of terms one has, provided one has made certain allowances, such as accounting for equivocal terms, eliminating redundant terms, and counting some multi-term expressions as "simple expressions" standing for single concepts (as I will explain in section 5.4.). As for the second objection to Davidson's principle, it is not really applicable. If one finds that two agents share most concepts, nothing follows about the extent or nature of disagreement that may exist between them. Although the attribution of concepts is inextricable from the attribution of beliefs, an indefinite amount of disagreement can still obtain even if most concepts are shared (indeed, even if all concepts are shared).

The Principle of Conceptual Charity says that differences between two agents or theories are to be construed as theoretical differences rather than conceptual ones whenever the evidence seems equally weighted in favor of the two options. When we are faced with a choice between using an existing term in our language or theory and coining a new one, and when the existing term allows us to make sense of an agent's utterances and other behavior, we make the relevant translation rather than resort to a neologism and rule that there is conceptual difference. This advice does not entail a radical reform of interpretive practice. Indeed, it consecrates one aspect of that practice, for interpreters do not generally regard every difference in belief as leading to a difference in concept, but normally absorb large differences in belief within their concepts. This is not a claim about the conceptually necessary conditions for something to be a translation or interpretation; it

is rather a claim about the practice of real interpreters, who generally follow this precept. If one looks closely at actual interpretive practice, it becomes apparent that we routinely exercise a large measure of charity in the ascription of concepts, since there is always a multitude of possible concepts that are compatible with the evidence in any particular case. For example, in any given interpretive situation, we are in principle free to ascribe, instead of the concept electron, such concepts as electron-on-Tuesday, electron-or-beetle, electron-in-cloud-chamber, and so on, coming up with a neologism to stand for any one of these novel concepts. Even if each of these conceptual ascriptions is ruled out by the accumulation of more evidence, there will always be others available given the (always) limited body of evidence. However, we do not ascribe such new gerrymandered concepts in practice, but rather rely on our own concepts, by and large. This shows that in normal ascriptive practice, we prefer to use our own concepts to interpret another rather than coin new concepts at will. In other words, we do not avail ourselves of these hypothetical positions in conceptual space in ascribing concepts and beliefs to an interpretee.

But in advocating conceptual charity, I am not slavishly following actual practice; I am guided also by the overall aim of interpretation or translation. The aim of interpretation is obviously to render another agent's set of beliefs in our own terms. The object of the exercise is to use the terms that we already have in order to comprehend an alien thinker or understand an alien theory, which is what actual interpreters attempt to do. To be sure, new terms may sometimes need to be introduced. But the introduction of new terms will always be a supplement to the main task, which is to use our own, pre-existing terms. A translation that used neologisms across the board would be no translation at all. This is what justifies being charitable in the ascription of concepts: the goal of the interpretive enterprise itself, which aims generally at expressing the thoughts of another, as far as possible, in the terms that we already possess. Thus, we lose nothing and gain considerably if we adopt a version of Occam's razor for concepts: do not multiply concepts beyond necessity in interpretation.

One might still say that we have indeed lost something by following the dictates of this principle, namely an important area of disagreement. The principle distorts things by making serious conceptual differences come out sounding like less important theoretical ones, it might be said. However, a disagreement is not diminished by construing it as a

disagreement in beliefs rather than a disagreement in concepts. If meaning and belief are inextricably linked, the slack between us can, in principle, be taken up in terms of beliefs rather than concepts. It is this feature of an interpretive theory of meaning that prompts Gilbert Harman to say that we can always choose whether to construe a disagreement as one in meaning or one of theory. Or, as he puts it, no distinction can be made between "a person's internal dictionary and the entries in his internal encyclopedia." (1973, 97) But according to the view I am advocating, though this may be correct inosfar as it is a restatement of the inextricability of meaning and belief, it is incorrect as an actual interpretive principle, because in each particular case, there will be good reasons for taking up the slack one way rather than another, of deciding to interpret a difference as one in the dictionary or one in the encyclopedia. And if the reasons seem to be equally weighted, one should exercise conceptual charity and interpret it as a difference in theory.

To illustrate these points, one can return to two examples used in Chapter 3. Davidson's example of the man who mistook a ketch for a yawl is a good illustration of some powerful considerations for surmising a difference in dictionary rather than a difference in encyclopedia. In normal perceptual conditions and when one's companion's eyesight is good, an utterance of 'yawl' in the presence of a ketch drives us to conclude that our interpretee's term should not be translated homophonically. That is how we were able to conclude in section 3.3. that his 'yawl' should be translated as our 'ketch' (and perhaps vice versa). But a second example illustrates the pointlessness of ascribing a difference in dictionary when there is no good reason for doing so. I imagined in section 3.5., in a variation on Quine's most famous example, that the native informer utters the term 'gavagai' (which in Quine's thought experiment is translated 'rabbit') often in conjunction with a term already translated as 'sacred'. Should we conclude that Quine's native has the concept schmabbit, where schmabbits are just like rabbits except that they are sacred, or that the native shares our concept rabbit but has the belief that rabbits are sacred? If one gains nothing in understanding, but one loses considerably in terms of the overall enterprise of interpretation, then there is nothing in this version of the example that recommends the first course of action. As in any inquiry, we should adopt the methodology that makes our task easier as long as it does not skew our results. Therefore, though meaning and belief are inextricable in principle, we can extricate conceptual change from

theoretical change in practice partly by adopting the Principle of Conceptual Charity. This maxim is justified by a naturalist attitude to meaning and interpretation, which takes seriously our interpretive practices and recommends reforming those practices only where that is needed in order to bring them in line with our broader aims.

At this point, it may be objected that this principle is not always exercised by interpreters, since it is not unknown in interpretive practice for charity not to be exercised regarding concepts. It must be admitted that such a principle would not be endorsed by some intellectual historians, historians of science, or historians of philosophy. For instance, R.G. Collingwood excoriates "realists" who would say that "Plato's State is different from Hobbes', but they are both States; so the theories are theories of the State." (1939, 61) He thinks that this is "only a piece of logical bluff" or "logic-chopping". Instead, Collingwood notes, if you "called for definitions of the 'State' as Plato conceived it and as Hobbes conceived it, you would find that the differences between them were not superficial but went down to essentials." He concludes: "You can call the two things the same if you insist; but if you do, you must admit that the thing has got diablement changé en route, so that the 'nature of the State' in Plato's time was genuinely different from the 'nature of the State' in Hobbes'..." (1939, 61)

The first point to be made against Collingwood is that one cannot always go by what agents themselves regard as defining their concepts, but that these definitions should be considered on a par with other tenets of the whole theory. On the view being reiterated throughout this book, the definitions by themselves cannot be decisive--though they will be counted among the theoretical tenets and will therefore be part of the evidence on which the interpretation is based. Secondly, Collingwood is less than adamant about not rendering 'polis' as 'State', since he allows the "realist" to call the two things the same--on the condition that the realist admit that it has changed considerably en route. But this is something that the interpretive approach could allow, simply by saying that the two theories of the state are substantially different.

Of course, Collingwood may yet be right about this particular case, since Plato's 'polis' may not best be rendered as Hobbes' 'state', but that is not evident for all that he has said here. At any rate, his opposition to the "realist" seems to derive from a deep-seated opposition to any view that perceives certain abiding concerns in intellectual history and

posits a measure of conceptual continuity across theoretical change. This is clear from a send-up of the realist that he evokes a little further on:

> It was like having a nightmare about a man who had got it into his head that trieres was the Greek for 'steamer', and when it was pointed out to him that descriptions of triremes in Greek writers were at any rate not very good descriptions of steamers, replied triumphantly, 'That is just what I say. These Greek philosophers... were terribly muddle-headed, and their theory of steamers is all wrong.' If you tried to explain that trieres does not mean steamer at all but something different, he would reply, 'Then what does it mean?' and in ten minutes he would show you that you didn't know; you couldn't draw a trireme, or make a model of one, or even describe exactly how it worked. And having annihilated you, he would go on for the rest of his life translating trieres 'steamer'. (1939, 64)

At best, Collingwood's remarks here might be taken as a cautionary note about erring too much in the direction of conceptual charity. If this concern is shared by others, recall that it is not the only interpretive principle that is being followed; indeed, some of the additional principles to be adumbrated below will help serve as a corrective.

Somewhat different criticisms may be made of a discussion of Alasdair MacIntyre's, who would seem to be another violator of conceptual charity. He makes the following (famous) conjecture: "Suppose that during the seventeenth and eighteenth centuries the meaning and implication of the key terms used in moral utterance had changed their character..." (1981, 55) He goes on to notice that this would render invalid previously valid deductive arguments:

> [I]t could then turn out to be the case that what had once been valid inferences from or to some particular moral premise or conclusion would no longer be valid inferences from or to what seemed to be the same factual premise or moral conclusion. For what in some sense were the same expressions, the same sentences would now bear a different meaning. (1981, 55)

He then adduces evidence for this supposition of meaning change, arguing that man went from being a functional concept in ancient and medieval Europe to being a non-functional concept in the modern world. Finally, he adds that other key moral terms must also have partially changed their meaning. (1981, 56)

Such claims are not sufficient evidence of a genuine conceptual change rather than a mere theoretical change, and there are features of MacIntyre's analysis that belie these meaning-change claims. For he does not mention any new concepts introduced by the Enlightenment in place of the concept man, and he does not say that the ancient concepts require us to introduce neologisms to make sense of them. He wants to claim that the concept of man has changed, yet maintain that both sets of theorists and practitioners were talking, in some sense, about man. This is a course of interpretation that I am arguing against. If one wants to say that the two traditions are talking about the same thing, then it is no longer possible to insist there was a conceptual change; only a theoretical change can be involved in such cases, though possibly a wide-ranging and important one.

There is a more damaging problem with MacIntyre's analysis. If arguments that were once valid before the Enlightenment later ceased to be so, that suggests--at best--that the crucial term involved became equivocal in different occurrences, not that it changed uniformly in meaning. For if all occurrences of a term changed meaning at once, and no other change was in play, then the deductive arguments in which it figured would remain valid after all, though their conclusions would differ in meaning. It is more plausible to suggest that a change of theory blocked moral theorists from making certain inferences that were once regarded to follow from the putative definition of 'man'. Deductive arguments sometimes include as an implicit premise, a "definition" of one of the key concepts; if this "defining" premise is abandoned, the argument ceases to be valid. If that were so in this case, it would be one of those theoretical changes in which a "definition" has been revoked without changing the concept. But the revocation of what was once considered a definitional belief is not sufficient for conceptual change to take place.

A more basic objection to the Principle of Conceptual Charity might be raised here. Someone might say that the adoption of this principle indicates that the whole interpretive approach that I am advocating is instrumentalist about concepts. In a fundamental sense, the objector might say, there is no fact of the matter whether a disagreement is one in concepts or in beliefs according to this approach, and since the latter is generally more convenient, it is not surprising that a principle has been adopted that recommends taking a disagreement as a theoretical one wherever possible. By my lights, this objection makes the mistake of reifying concepts or of treating them as discrete things that are subject to

the same individuative standards as ordinary spatio-temporal objects. For it assumes that there is some deep metaphysical fact of the matter whether concepts are shared that has nothing to do with the interpretation of one agent by another. By contrast, the interpretive approach subordinates the question of whether a concept is shared among two theories or agents to the larger concern of making overall sense of that theory or agent. The interpretive approach regards concepts as part of the theoretical apparatus involved in the interpretation of agents and their ascription is subsumed under the general task of interpretation. Further justification of this view of concepts will be given in the following chapter (see especially sections 6.5. to 6.9.).

Finally, a note on what this principle is not. It should be distinguished from the metaphysical realist assumption involved in certain referential views, which has the effect of disregarding agents' beliefs altogether in determining the reference of their terms. This assumption is involved in the examples cited in the conclusion to Chapter 2, namely Boyd's about managing to refer to an astronomical entity after reading a newspaper headline and Evans' about referring to a person after overhearing a snippet of a conversation in a pub. These cases involve something more than charity; they suggest that successful reference has little to do with belief at all. Perhaps Kitcher puts this attitude most starkly when he writes: "[E]ven if I were to believe that tigers were herbivorous, spotted canines (producing some such erroneous description when asked to identify tigers), it is still possible that I should use 'tiger' to refer to the set of tigers." (1982, 341) That is clearly denied by the interpretive approach, for Kitcher's hypothetical self seems to lack <u>any</u> true beliefs about tigers on the basis of which he could be ascribed the concept or be said to refer successfully. By contrast, charity implies that <u>all</u> beliefs need not coincide for concepts to coincide--not that <u>none</u> need coincide.

5.3. <u>Principle of Uniformity</u>

Another principle used in the interpretation and comparison of scientific theories is the Principle of Uniformity. This principle says that translation should generally be type-type, or that the same term from the source theory should be substituted for a given term from the target theory on each occurrence. This is in opposition, for example, to the course of action recommended by Kitcher in translating scientific terms. As seen in the previous

chapter, Kitcher advocates what he calls a "context-sensitive" theory of reference and interprets different occurrences of the term "dephlogisticated air" differently, despite the fact that he does not think it was equivocal for the phlogiston theorists. On my view, by contrast, once we have hit upon the translation of a particular term in a certain theory, that translation should be adopted wherever the term appears in the theory. The obvious exception to this principle is the case of equivocality for a term in the source theory. Therefore, a uniform translation should be adopted unless a term is equivocal in the theory being translated. In scientific theories, a term should not be interpreted to be equivocal unless there is clear evidence for such an interpretation, whether explicit or implicit. Thus, in the phlogiston theory, 'dephlogisticated air' should not be translated sometimes as 'oxygen' and sometimes as failing to correspond to any of our terms.

Interpreters generally have ways of determining whether a certain expression is genuinely equivocal or not. But it is important to bear in mind that the conclusion that a term is equivocal should only be reached when there is a sign that this is the case from the theorists being interpreted, whether explicit or implicit. It should be said, however, that it is not always a trivial matter to distinguish in practice between equivocality and vagueness, or as some linguists would say, ambiguity and lack of specificity. At first, it may seem as if it is always open to the advocate of a rival interpretation to see only vagueness where we have identified equivocality. But luckily, linguists have devised a number of tests, some of which can be used in this connection. One obvious test is to see whether the term can be used to generate fallacies of equivocation in deductive arguments. If one of Aristotle's terms is genuinely equivocal as between 'instantaneous velocity' and 'average speed', some arguments that turn on both types of occurrence should be blocked, whereas we normally expect them to go through if they are merely vague. If arguments that rely on both uses are never deployed by Aristotle, or if they would lead to conclusions inconsistent with other parts of his theory, this gives us some grounds for regarding the term as equivocal.[2]

---

[2] Other criteria for ambiguity (as opposed to vagueness) have also been employed by linguists. Some of these have been surveyed by Zwicky and Sadock (1975), but most of their tests are not easily applicable to the interpretation of scientific theories, since ambiguous scientific terms tend to be syntactically of the same category and semantically

The Principle of Uniformity can be justified by noting that it is necessary for preserving the inferential structure of the theory we are interpreting. It has been suggested by Hartry Field, for example, that Newton's term 'mass' should sometimes be translated as 'rest mass' and sometimes as 'relativistic mass'. As seen in the previous chapter, each translation makes some of Newton's beliefs agree with Einstein's, so this strategy is sometimes advocated for supposed reasons of charity. But the charity here is misplaced and the mistake behind this move can be exposed simply by noting that such a translation would obscure the inferential structure of Newton's theory. Suppose that two occurrences of Newton's term that appear in a single deductive argument were translated differently. Then, an argument that was formally valid would generally cease to be so. When a Newtonian physicist argues as follows:

Mass is an invariant quantity for a given physical object.

Momentum is the product of mass and velocity.

∴ For a given physical object, momentum varies only with velocity.

we would interpret the argument as follows:

Rest mass is an invariant quantity for a given physical object.

Momentum is the product of relativistic mass and velocity.

∴ For a given physical object, momentum varies only with velocity.

While the first argument is deductively valid, the second clearly is not, for the premises do not contain any common terms, so the Newtonian physicist comes out committing a logical error. So long as the inferential structure of the theory, and hence literal meaning, is our primary interest, I would argue that this interpretive principle should be followed.

Historians of science and intellectual historians may protest that this is a principle that they are often forced to violate. One sometimes comes across interpreters such as Field and Kitcher who say quite explicitly that a certain key term in a text or author under discussion cannot be translated uniformly throughout, for it has a different meaning or

---

proximate. The most promising for our purposes are what they call "identity tests", which take advantage of certain transformational rules that work only for ambiguity and not for vagueness. These tests assume that vagueness is not encoded in syntactic structure but ambiguity is.

reference in different contexts.[3] I would say that, barring equivocality, insofar as they violate this principle, such interpreters have something other than literal meaning in their sights. To give some credence to this claim, it can be said that the Principle of Uniformity is expressly adopted by at least some intellectual historians. To illustrate, I will present two instances in which it is explicitly advocated by translators of philosophical discourse in interpreting two notoriously difficult philosophers: Hegel and Foucault.

In an introductory work on Hegel, Peter Singer discusses the difficulty of translating Hegel's term Geist and proposes that a translator has three options: to use 'mind' throughout, to use 'spirit' throughout, or to use whichever seems most appropriate in the context. But Singer rejects the third option, "because it is obviously important to Hegel that what he calls Geist is one and the same thing, notwithstanding the different aspects of it that emerge in his various writings." (1983, 45) Singer goes on to say that he intended to use 'spirit', but decided to neologize instead, since the term 'spirit' in English features in (as I would say) a certain religious or mystical theory, which one does not associate with a clear scientific view of the world. Then he writes that at some point "we might have to say that his philosophy is based on this somewhat superstitious view of the world and his concept of Geist is intended to refer to just such a ghostly, disembodied being." (1983, 46) But Singer cautions against assuming this from the start. I would agree entirely with these remarks, since they indicate that Singer thinks that it is only on the basis of its occurrences in Hegel's theory that we could conclude that Hegel's concept is superstitious or mystical.

A similar stance is taken by Foucault's translator in a prefatory note to The Birth of the Clinic, where he explains that it is important to translate certain key terms uniformly.

---

[3] Recall that Kitcher favors a "context-sensitive theory" of reference, which would recommend translating expression tokens of the same type differently depending on the context in which they appear. Concerning the expression 'dephlogisticated air', he writes: "[I]f we treat all tokens of the same type in the same way, then we shall be led to the position defended by Kuhn and Feyerabend: there is no term of contemporary English that specifies the referent of 'dephlogisticated air', so that a term that is central to the presentation of the phlogiston theory resists translation into contemporary language." (1978, 534)

After stating that Foucault has a tendency to use certain key words repeatedly, he says that some of these words have no equivalent. In such cases, he claims, it is preferable to use a single unusual word rather than a number of familiar ones: "When Foucault speaks of la clinique, he is thinking of both clinical medicine and the teaching hospital. So if one wishes to retain the unity of the concept, one is obliged to use the rather odd-sounding 'clinic'." He goes on to say that he has similarly deployed the unusual term 'gaze' to translate the common Foucauldian term 'regard'. (1973, vii) Here, both 'clinic' and 'gaze' are best viewed as neologisms in the context of Foucault's theory about the rise of clinical medicine in France. Although both words exist in English, they are being introduced in a special technical sense peculiar to Foucault's theory, just as Foucault himself had invested the French originals with new meanings, introducing new unitary concepts.

In the following section, I will try to elucidate further the connection that exists between terms and concepts, that is, the relation that obtains between having a term for something and possessing the concept of that same thing. To anticipate somewhat, it is safe to assume that we seek in rational discourse to have our terms conform to our concepts, an aim which is no less true of scientific discourse. That is why one term does not generally stand for two concepts, and two terms do not generally pick out a single concept. Exceptions are of course well known, but they are very much deviations from the general rule. This observation helps to undergird the Principle of Uniformity and it helps to explain why that principle also operates in reverse: one should not use a single term to translate two different ones unless it is clear that they are being used interchangeably. If two different syntactic items are being used interchangeably (whether due to historical accident, stylistic variation, or some other reason), it should be ascertainable from the practice of the scientists involved. An example of the workings of this principle can be drawn from the phlogiston theory case study. There, the decision to translate the term 'inflammable air' as 'hydrogen' in section 4.3. was taken as sufficient evidence that the term 'phlogiston' should not be translated as 'hydrogen'. Therefore, we were implicitly assuming that it should not be translated differently on different occurrences.

5.4. Principle of Simplicity

To elucidate the next principle, it is necessary to broach an issue that has been postponed for long enough. I need to give some account of the connection between terms and concepts. Sometimes in earlier chapters and earlier sections of this chapter, the two things were treated almost on a par and I have not sufficiently distinguished having a term for something and having the concept of something. The reason is that the two things are very closely linked in my view, given two crucial simplifications. The first thing to notice is that according to the approach outlined in Chapter 3, there is no concept that we do not have, at least potentially. This is just a version of Tarski's claim of the "universal character" of language, according to which natural language provides "adequate facilities for expressing everything that can be expressed at all, in any language whatsoever..." (1969, 67) Tarski adds that "it is continually expanding to satisfy this requirement." (1969, 67) By this he means, presumably, that we can always introduce a new term to stand for a new concept, so that all natural languages potentially have the same expressive resources. Having said that, one can distinguish between the concepts that a language actually has and the concepts that it has only potentially. The latter set might be open-ended, but the former is fixed at any given time. It is the set of actual concepts that corresponds to the set of terms a language or theory has, subject to the following important qualifications.

One cannot just assume that concepts and terms can be put in a simple one-to-one correspondence. There are of course such things as equivocal terms and redundant terms, but they will always be exceptions to the general rule. The rule is that we aim to deploy only as many terms as we have concepts and to supply new terms for new concepts. In the previous section, I discussed how equivocal terms and redundant terms can be spotted and how to deal with them in the context of interpretation by using subscripts (equivocal) or dispensing with them altogether (redundant), thereby equalizing terms and concepts. Still, it might be said that the number of terms we have will not for all that correspond to the number of concepts we have, because we have many concepts for which we have no terms, or at any rate, no unitary terms. After all, it is often an accident which of our concepts are honored with their own terms and which make do with composite expressions to stand in for them. There is something to this complaint, and it explains why, in the previous chapter, I made a distinction between "simple" and "complex" expressions. The point of that distinction was to identify those expressions that, although they did not consist of a

single morphological or syntactic unit, were being used as though they were single terms (i.e. "simple" expressions). The prominent examples in the case studies were the expressions 'dephlogisticated air' and 'elementary atom'. Because of their frequent use in framing scientific generalizations and explanations, I argued that they had effectively turned into simple expressions. Thus, although they are not single morphemes, they can be treated as single lexical items and can therefore be considered to stand in for concepts. When this provision is added to the qualification about equivocal and redundant terms, the one-to-one correspondence between terms and concepts can be upheld. It is therefore important to distinguish between simple and complex terms in interpreting a scientific theory.

Although the distinction I have made between simple and complex terms may appear vague, it can be grounded in a distinction that linguists have introduced between compound lexemes and syntactic compounds. John Lyons characterizes both types of expression as lexemes "whose stem is formed by combining two or more stems (with or without morphological modification)." (1977, 534-5) Examples of such expressions would include: 'screwdriver', 'window box', 'bread-knife', and 'public school'. The difference between the multimorphemic expressions that I have been calling simple and complex corresponds to the distinction between compound lexemes and syntactic compounds, respectively. While the meaning of syntactic compounds can be accounted for in terms of the productive rules of the language, compound lexemes often start out as syntactic compounds and "having become institutionalized, acquire a more or less specialized meaning." (1977, 535) To illustrate the phenomenon of compound lexemes, Lyons gives the example of the term 'country house' which is used in British English to denote a much smaller class of dwellings than the expression 'house in the country'. Some compound lexemes are even more semantically irregular, since their meanings depend even less on the meaning of their constituents. For example, in British English, a 'public school' is not public but private, and a 'public house' is not a house but a bar. In such cases, Lyons makes clear, the compound lexeme needs to be given a separate entry in the lexicon. This vindicates the proposal to treat certain multimorphemic expressions such as 'dephlogisticated air' and 'elementary atom' as "simple terms", according them their own analytic hypotheses rather than considering them as derived in a regular fashion from their

constituents (in these cases: 'de-', 'phlogiston', '-icate', '-ed', 'air', and 'element', '-ary', 'atom', respectively).[4] In Lyons' terms, 'dephlogisticated air' and 'elementary atom' are semantically irregular compound lexemes.

But how are we to know which terms are which without presupposing a particular interpretation of them? Lyons' discussion of how such compound lexemes acquire semantic properties not wholly derivable from the productive rules of the language helps to support the discussion in the previous chapter of the term 'dephlogisticated air'. There, I agreed with Kitcher that when Stahl first coined the term 'dephlogisticated air', the only belief associated with it was that there was a substance that resulted from the absorption by air of phlogiston, but I went on to argue that by the time of Priestley and Cavendish, the most plausible interpretation was that 'dephlogisticated air' had turned into a simple expression that could be translated uniformly as 'oxygen'. Lyons' account helps to justify such a course of action by describing the process whereby syntactic compounds are "institutionalized" as compound lexemes. This process has been called "petrification": "As soon as any regularly constructed expression is employed on some particular occasion of utterance, it is available for use again by the same person or by others as a ready-made unit that can be incorporated in further utterances; and the more frequently it is used, the more likely it is to solidify as a fixed expression, which native speakers will presumably store in memory, rather than construct afresh on each occasion." (1977, 536) Still, this does not yet tell us how to decide when syntactic compounds have been transformed into compound lexemes, thereby requiring separate lexical entries. In answer to this question, Lyons says: "It is easy enough to formulate the general criteria for inclusion in the lexicon: a lexical entry is required for compound lexemes... if and only if they are phonologically, morphologically, syntactically, or semantically idiosyncratic." (1977, 536) But this is precisely what is at issue in interpreting an alien theory: Is the expression semantically

---

[4] In fact, strictly speaking, both these terms also illustrate another linguistic phenomenon, that of complex lexemes, which involves attaching to a stem a derivational affix or systematically modifying it, e.g. the derivation of 'friendly' from 'friend'. Lyons makes it clear that many complex lexemes also need separate entries in the lexicon (which is what makes them "simple" in my terms).

irregular or idiosyncratic according to the productive rules of the language? Note that semantic idiosyncracy can sometimes be accompanied by regularity on the other counts (phonological, morphological, and syntactic), so we cannot use these other criteria as evidence for it. For example, Lyons says that 'country house' "is completely regular as far as its phonological, morphological, and syntactic properties are concerned; and its status as a lexeme depends solely upon its idiosyncratic and unpredictable semantic specialization." (1977, 540)

I propose to distinguish simple from complex expressions by comparing the beliefs associated with them with the beliefs associated with their parts. The distinction can be made without begging the question by examining whether the new beliefs or theoretical tenets that are associated with such an expression follow simply from the semantic properties of its parts. Even if some of these beliefs do not actually serve to contradict the semantic properties of its parts (e.g. 'A public school is not public'), the fact that they do not simply result from them is an indication that the term has become petrified and that it corresponds to a genuinely new concept. For instance, the belief that a bread-knife is a tool follows directly from the belief that a knife is a tool together with the belief that a bread-knife is a kind of knife. It cannot therefore lend support to the claim that 'bread-knife' has become petrified (though it may have, depending on other beliefs in which it is featured). But the belief that a public house serves alcohol does not follow from our beliefs about houses or our beliefs about public places, and is evidence for the claim that 'public house' has become petrified. Similarly, the belief, say, that 'Dephlogisticated air is good to breathe,' does not follow from the beliefs that Priestley held about 'phlogiston' and 'air' when they are combined in the requisite way with the morphemes 'de-', '-ate', and '-ed'. This is what justifies treating it as a compound lexeme (simple) rather than a syntactic compound (complex).

Therefore, the translational principle that calls for distinguishing between simple and complex expressions and for ruling that only the former correspond to new concepts, can be justified using a bona fide linguistic distinction between compound lexemes and syntactic compounds. No simple criterion serves to distinguish the two kinds of expression, but there are certain distinctive features of the process by which an expression evolves from being a syntactic compound to a compound lexeme (in my terms, from being

complex to being simple). I would claim that one of the hallmarks of this process of petrification is the association of a number of new beliefs with the expression, beliefs that do not follow in regular ways from the beliefs associated with its components. In the context of scientific inquiry and in the course of accumulating new beliefs about the world, we should expect this process to occur frequently and should therefore always consider the option of regarding multimorphemic expressions to be compound lexemes as opposed to syntactic compounds in interpreting a scientific theory. The process of petrification is clearly in evidence in the career of the term 'dephlogisticated air', as it evolved from a one-criterion expression introduced by Stahl to an expression that came to be associated with a slew of new beliefs by Priestley.

5.5. Principle of Warranty

The next interpretive principle can be introduced by way of an objection to the Principle of Conceptual Charity. Someone might say that if we take conceptual charity perfectly seriously, then it makes possible a highly counterintuitive manoeuver in the latter stages of interpreting a theory. Imagine that we have nearly completed the interpretation of the source theory in terms of the target theory and suppose, for the sake of simplicity, that all we have left is a single unmatched term in each theory. Since conceptual charity recommends that a disagreement should be construed wherever possible as theoretical rather than conceptual, it might be said that there can be no obstacle to matching up the two terms.

To make the problem vivid, consider the following hypothetical example. Suppose that we are interpreting the phlogiston theory and that the only term we have not managed to correlate to any of ours is none other than 'phlogiston'. Suppose further (implausibly) that the only term of ours that we have not found a translation for in the terms of the phlogiston theory is 'hydrogen'. Even if there were no shared beliefs that would emerge if this translation were made, conceptual charity seems to recommend taking this course of action rather than neologizing for the term 'phlogiston' and ruling that it does not correspond to any of our terms.

This is where the exception to conceptual charity enters into the picture. The exception can be summed up by the Principle of Warranty for ascribing a concept or

making a particular correlation between terms. This principle enjoins us not to ascribe a concept without sufficient justification, even when such an ascription would involve ascribing few or no false beliefs. The warrant will be given by the existence of a modicum of true beliefs in which the concept features, and the warrant can be undermined when there is an alternative substitution that will do equally well. Of course, the problem only really arises with leftover terms that are left unmatched in both theories, for it is only in these terminal cases that stronger candidates do not emerge for translating the relevant term, since any such candidates will already be accounted for by this stage.

It may be protested that this verges on vacuity. It seems quite uninformative to say that a concept should only be ascribed when we have adequate warrant and should not be attributed when we have insufficient reason for doing so. However, the principle, though vague, is not entirely vacuous provided one specifies the type of warrant that is required (as well as the type of evidence that does not constitute warrant), and provided one shows how it fits together with other principles. The warrant for ascribing a concept is supplied by the existence of a certain number of beliefs that come out shared if the ascription in question is made. The requisite number of such beliefs cannot be specified in advance (and there will be an indefinitely large number of potential beliefs), but in scientific contexts the decision to ascribe a concept is made in much the same way as the decision to credit a certain intellectual discovery. We can say that a concept will be ascribed if the beliefs associated with it helped launch a research program that made it possible to uncover many other true beliefs associated with that concept. This was part of the reason that it was legitimate to ascribe the concept oxygen to phlogiston theorists like Priestley. In other cases, an inclusive concept, which we may or may not have to neologize for, should be ascribed when there is not sufficient warrant for ascribing a more specific one. That course was followed in interpreting Dalton's term 'molecule', which was rendered by the inclusive concept ultimate-particle.

In many contemporary discussions, the issue of sufficient warrant often comes up in deciding whether causal contact with the right kind of entity or property is sufficient for ascribing the corresponding concept to an agent or a community. The Principle of Warranty states that causal contact is neither necessary nor sufficient. It is not sufficient, for there are cases in which causal contact obtains, but the absence of requisite beliefs tells

against ascribing a concept. The arguments for this position were already given in Chapter 2, where the causal theory of reference was criticized precisely for divorcing successful reference (let alone possessing the relevant concept) from the having of shared beliefs. If one understands causal contact broadly along the lines of the causal theorists of reference, then it is clear that some of the examples encountered earlier in this book show the pitfalls of regarding it as decisive in this regard. In some cases, it is clear that causal contact has obtained and yet the scientists involved do not achieve the associated concept. An illustration is provided by the example cited in section 2.3. of early uses of the term 'electron' (by Stoney and others). Though these scientists clearly had causal contact with electrons, they used the term to stand for the unit of negative charge and not for the particle itself. Hence causal contact is not sufficient grounds for ascribing a concept. Causal contact is not necessary either. As we saw in section 2.5., there are many cases in science in which the existence of entities and their properties are predicted successfully despite a lack of causal contact. Otherwise, we could not say that Dirac successfully predicted the existence of positrons, nor that Bohr correctly anticipated the properties of hafnium. Neither of them had had any causal contact with these types of entity. This principle does not deny that causal contact with the correct (from the interpreter's point of view) type of entity may be one consideration among others in deciding whether to ascribe a concept. But mere contact is not enough; in deciding whether a scientist can be ascribed a concept, we may also ask ourselves whether the contact was of the right quality and quantity. Brute causal contact should neither be regarded as sufficient warrant, nor even necessary for the ascription of a certain scientific concept.

Still, it might be said, there are cases in which the surrounding beliefs are incapable of singling out one of two different kinds of entity, and causal contact provides us with some reason to favor ascribing one concept over another. This arises particularly in dealing with less sophisticated inquirers who make fewer distinctions than we do. It may turn out that two or more of our concepts could be substituted for theirs equally well, though they have had a determinate causal connection with only one. In such a case, causal contact (though not of a brute variety) may indeed serve as a tie-breaker in applying the Principle of Warranty. Therefore, causal contact is not sufficient on its own, but it may tip the balance when a modicum of beliefs are found to be in place. In interpretive practice,

the entities with which our interpretees have actually had some causal connection may be privileged in deciding which concept to ascribe to them. This can be justified by saying that initiating causal contact of a certain type with a scientific entity makes it more likely that a research program has been launched that will isolate further properties of the relevant type of entity. The causal links function here as a promissory note that render it likely that these inquirers will eventually converge on one of our concepts rather than the other. However, if we cannot say that their causal links to one type of entity are any stronger or more determinate than they are with the other, we may resort to ascribing an inclusive concept that straddles both of our concepts.

The utility of inclusive concepts is worth dwelling upon further in this connection, since it does not seem to be widely acknowledged. An intuitive (but hypothetical) example can easily be given. Imagine a group of scientists working in the early part of this century who do not have enough true beliefs about neutrons or protons to enable us to ascribe either concept to them. We might decide to ascribe the concept <u>hadron</u>, which includes neutrons and protons. This can be compared to translating Dalton's term 'molecule' by the neologism 'ultimate-particle'. Note that this is importantly different from ascribing the disjunctive concept <u>neutron or proton,</u> or <u>atom or molecule</u> in Dalton's case, because these concepts have an internal semantic structure that the scientists' concept lacks. But it should be clear from the tenets of their theory that their concept includes what we would pick out as neutrons and protons. Therefore, inclusive concepts can be attributed when an alien theory does not make fine-grained distinctions, but narrower concepts should be ascribed where possible. That consideration was partly operative in the decision to render Newton's term 'mass' by 'rest mass', rather than neologize and ascribe an inclusive concept that has some of the properties of rest mass and some of relativistic mass.

This translational principle obviously also provides guidelines for deciding that a certain term of the translated theory fails to correspond to a term of our own theory. Just as we ascribe a concept only when there is sufficient warrant for doing so in the manner explained, we should resist passing this judgement when there is no sufficient warrant for doing so: that is when some other term of ours would do equally well (which is to say badly) as a translation of the contested term, or if the inquirers in question cannot be said to have initiated a research program that led to the identification of the entity in question.

If there are two or more of our terms that can serve indifferently as translations of the disputed term, then neither of those translations should be adopted. The conclusion of insufficient warrant is reached when all the plausible interpretive hypotheses serve to decrease the area of agreement between the two theories and when attempts at repairing this by compensatory translations of other terms fail to improve the situation.

5.6. Principle of Undefinability

One principle that emerged explicitly in the case studies discussed in the previous chapter, though it was prefigured as early as the Introduction, is the undefinability of scientific terms. That is to say that it is impossible to specify definitions for scientific terms which are indefeasible in the context of inquiry. Just because holders of a theory take a certain sentence to give the meaning of a certain term or to provide a theoretical definition, that should not preclude the possibility that that sentence is not shared between the two theories, though the concept in question is. Therefore, when deciding whether a concept is shared or not, we should not rely on what the agents themselves take as the definitions of those concepts; agreement in definitions does not imply agreement in concepts and divergence in definitions does not imply divergence in concepts, even in cases where definitions are explicitly provided.[5]

This principle can be illustrated by the discussion of the phlogiston theory, where it was argued that the following sentence should not be taken as definitional of the term 'dephlogisticated air':

> (P6) Dephlogisticated air is the substance which results from removing phlogiston completely from the air.

The possibility should not be ruled out that this sentence might just be one of the false (non-definitional) sentences of the phlogiston theory, and that the term that it purportedly defines has a counterpart in post-phlogiston chemistry: 'oxygen'. This interpretive decision

---

[5] This attitude to definitions is sometimes taken as tantamount to a rejection of analyticity or a denial of the analytic-synthetic distinction. But since those terms are laden with philosophical baggage, I will refrain from phrasing things in terms of the analytic-synthetic distinction, as I already indicated in the Introduction.

relies partly of course on the treatment of 'dephlogisticated air' as a simple expression (or in terms of section 5.4., as a compound lexeme). For that is what enables us to treat (P6) as a definition rather than a logical tautology. Tautologies should be preserved on pain of imputing illogicality, but definitions need not be.

To insist on preserving what an alien or rival theory considers definitional is to limit unnecessarily the area of agreement between that theory and ours. The practice of ignoring definitions is not only justified by the history of science, which shows that definitions are often revisable; it is also supported by the fact that there is sometimes disagreement among holders of a single theory as to what should be considered definitional. Indeed, different presentations or axiomatizations of a given theory often differ over what they consider to be the definitions, yet such differences ought to be considered different formulations of the same theory rather than formulations of different theories.

To avoid any misunderstanding and because it is a controversial topic, I should explain further what I mean by the undefinability of scientific terms. I take it that this amounts to denying that we can decide in advance which beliefs to hold constant and which to revise in the course of future theory changes. We cannot choose which beliefs to take as definitions and which to take as informative statements, if this means that the definitions are indefeasible or unrevisable. Corrigibilism means admitting that it is not generally within our power to hold certain beliefs fast, come what may. The objection to the existence of scientific definitions does not amount to a denial of the existence of 'meaning postulates' in Carnap's sense, if such postulates can be shown to play a useful role in a theory or to be particularly perspicuous for certain formulations of a theory, and so on. It is only an objection to the assumption that meaning postulates are unrevisable. In his later work Carnap freely admits that any and all theoretical tenets are revisable, including those of logic and mathematics.[6]

---

[6] In a reply to Quine, Carnap writes that he is in agreement with Quine's view that "no statement [in a total system of science] is immune to revision, not even the statements of logic and of mathematics." (1963, 921)

One might respond to the claim about the revisability of definitions by wondering why we cannot just dig our heels in and decide to use words in one way rather than another, thus preserving whatever definitions we happen to prefer. Suppose that we are particularly attached to a certain definition, then it might seem within our power to retain it against all odds as long as we are willing to make other, possibly extensive, revisions in our theory. The problem with this approach, however, is that such things are not generally a matter of choice; we can use our words as we please, but the same does not go for our explanatory concepts. Such a policy would lead to the retention of hollow concepts that play no explanatory role. Consequently, despite the way that terms are used, a comparison of the old theory with the new one would match up the terms that are doing the explanatory work even though the adopted definitions militate against it.

To illustrate, consider a classical physicist who was particularly attached to the theoretical tenet that momentum is the product of mass and velocity and decided to retain it no matter what theoretical winds blew. Moreover, imagine that this physicist holds the following as a definition of 'mass': mass is momentum divided by velocity. Now, as the theory changes, and we make the transition to relativity theory, the adjustment that is needed is clear; the physicist reserves the term 'mass' for what I have called 'relativistic mass', which is a quantity relative to the frame of reference, and insists on this usage, but still needs a concept to do the work of rest mass, call it 'schmass'. Has this physicist managed to vindicate the existence of scientific definitions? Not really, for I have argued that there is a unique way of comparing classical and relativistic physics, and this involves making a correspondence between 'mass' and 'rest mass' (or what this imaginary physicist calls 'schmass'). No insistence on using the terms differently can change this fact. Another way of putting this same point is by saying that there are better and worse ways of carving up the world and the concepts we use follow the ways of carving up the world rather than the whims of our usage. Although we are free to apply our words as we wish, we are surely not free to decide which concepts turn out to be central to our theories and efficacious for explaining the operations of nature. And though we might reserve a term for one concept, another concept may turn out to be the one that is required for explanatory purposes, and the first concept may turn out to be dispensable.

This denial of definitions in science suggests a certain exception in other domains. In cases where words are being used honorifically rather than for explanatory purposes, the denial of definability would not seem to hold. That is why a term such as 'bachelor' has a definition if any word does, because it is not a theoretical term that serves an explanatory purpose. But if we imagined that it came to have an important explanatory role in, say, a sociological theory about political orientation or voting patterns in U.S. elections, this might change. Someone might propose that the political behavior of bachelors exhibits certain characteristic patterns and that one can predict with some confidence how these bachelors will vote in elections given a few other pieces of information. Suppose that on subsequent modifications of the theory, as the profile of these voters were made more precise, it turned out that some married adult males also conformed to this behavioral pattern, since those males who were married but separated from their spouses for long periods exhibited the same political profile. We might want to conclude that the latter group were also 'bachelors'. Once 'bachelor' becomes an explanatory term, it raises the possibility of revising the belief that all bachelors are unmarried adult males. These points and related ones will be explored in section 7.3., where I will say more about explanatory concepts and carving nature at the joints.

Though it may seem an obvious point, it should be mentioned that the denial of definability does not mean denying that a correspondence can be made between terms from different theories. Translation functions, which are the outcome of the interpretive process, clearly require certain equivalences to be made between two sets of terms. Why is this not a commitment to the idea of definability? Just because this is an equivalence between different theories, rather than a commitment to the unrevisability of certain tenets of a particular theory in the face of any theoretical revisions. Quine calls the entries in his translation manual "analytic hypotheses", but these hypotheses are only valid for one interpretation at a time; different hypotheses may be needed after a theoretical revision in one of the two theories.

5.7. Principle of Neologization

I have said repeatedly that terms of one theory that are found not to have counterparts in another should be replaced by neologisms. These terms can be implicitly

"defined" by their relations to other terms that we have translated, but that is not to make a commitment to the definability of scientific terms. There can be nothing more to explicating them than indicating the (potentially infinite) totality of the beliefs in which they figure. But new terms should only be introduced when there are no terms among those of the target theory that can be considered adequate translations. Whenever possible, the Principle of Conceptual Charity dictates that an existing term should be used before considering coining an entirely new term. After all such options are exhausted, any term for which we have no plausible translation will receive a neologism.

Sometimes we may find it useful to import a concept even in the case of a previously rejected scientific theory. A possible example can be drawn from the case study of the phlogiston theory. A careful reader of section 4.3. will have noticed that at least one problematic term of the phlogiston theory was not adequately dealt with, namely 'phlogisticated air'. Although it was translated as 'oxygen-deficient air', this was not fully justified. There are three alternatives in translating such a term. We may well say that it is a vacuous term, since there is no such scientific kind as oxygen-deficient air in our theory, while retaining the term as a neologism. But it is also possible to neologize and retain it as a term for a substance for which we have no single expression, so that the analytic hypothesis would read: 'phlogisticated air' is translated as 'oxygen-deficient air'. In that case we would have added a new concept to our own theory. Finally, it can be said that air that is poor in oxygen is mostly nitrogen, so it should be translated as 'nitrogen'. This decision can only be made after a closer look at the theory and will be determined according to whether the phlogiston theorists had enough true beliefs about nitrogen to warrant attributing the concept to them, along the lines of what was done for oxygen. The evidence presented in the previous chapter was not adequate to decide between the three alternatives since the emphasis was on other concepts. The decision between them is attendant on more evidence. Nevertheless, we can now say that the course adopted in the previous chapter, the second option just mentioned, was not strictly speaking legitimate. That is because it assumes that we have already introduced a new concept to our own theory, that of oxygen-deficient air. But this will not yet have transpired in the initial stages of translation, so the consistent course of action would have been to adopt the first option and declare that 'phlogisticated air' is vacuous. If one goes back and effects this change, the

rest of the analysis survives intact. This is a true instance of reflective equilibrium at work; it is a case in which the judgement made in the previous chapter was not quite warranted, but the reasons have only emerged more clearly in this chapter. For expository purposes, it was easier to proceed in the more intuitive way and justify the theoretically consistent step later.

Here, a question could be raised that pertains to the defense of Conceptual Charity made earlier, in section 5.2. I claimed there that the whole point of translation was to render another theory in terms of our already existing concepts and that new concepts should be introduced only as a last resort. The justification appealed partly to the idea that no translation could use neologisms across the board. If I affirm both the general necessity of neologization and the absurdity of rampant neologization, how much neologizing is considered acceptable? As with a number of similar questions raised in this chapter, the answer must be vague. It is fair to say that there would be something wrong with a translation that required us to neologize for the preponderance of the concepts that were implicated in it, at least if the target language did not have many fewer terms than the source language. But little else can be said in the abstract. The more neologisms are required, the more a theory might be said to be conceptually distant from ours. Typically, theories widely separated in terms of time and sophistication will require more neologisms (or, at least one of them will, since the relationship is not symmetric). But for most practical purposes, the extent of neologization is usually small when compared to the <u>total</u> number of terms involved (including those shared with different fields or sub-disciplines). Rival scientific theories are expected to be at a comparable stage of sophistication, and where historical interpretations are involved, the target theory is usually more developed and complex than the source. That is because we are usually interested in interpreting "backwards" to a less complex theory rather than "forwards" to a more complex one. (Why would we want to interpret quantum physics in terms of Aristotelian physics, for example?) In practice, therefore, rampant neologization is not likely to prove to be a problem.

At this point, a worry should be considered that casts doubt on the very idea of using neologisms in interpretation. Someone might make a distinction between a translation and a semantic theory. While allowing that the use of neologisms is permissible

in the latter, it may be claimed that it is not in the former. A simple example from ordinary discourse that can be used to make the point is that of the two German words 'essen' and 'fressen', both of which mean 'eat', but while the former is reserved for people, the latter is used only of animals. One cannot provide a proper word-for-word translation of German into English because of this feature. This position is strongly reminiscent of one of the objections attributed to Kuhn in section 3.4. (especially the example of the French word 'doux'). This example, like Kuhn's, tends to undermine the original point, since by these stringent standards, it is unlikely that any two languages have ever been used to translate each other. They would have to have exactly the same number of concepts and those concepts must be placeable in a one-to-one correspondence without introducing neologisms. As I argued in section 3.4., drawing the boundaries of ideal translation so narrowly that no real translation would qualify is something of a sterile philosophical exercise. Therefore, neologizing should be seen as an inevitable supplement to any interpretive process, subject to the qualifications aired above.

However, there is a crucial distinction relevant to the practice of neologizing which needs to be made and which attaches an additional qualification to this whole discussion of neologisms. Neologisms raise a larger issue about translation. At least some of the terms for which we need to neologize will feature in beliefs that are strictly inconsistent with our theory. If we were to import them along with some of their theoretical tenets, we would be importing beliefs that are inconsistent with our theory. An obvious example is provided by the term 'phlogiston', whch figures in such theoretical tenets as, 'When a metal is heated in air, phlogiston is released.' Since we believe that no substance is released as a result of this chemical reaction, a contradiction would result if we were to import this tenet into our theoretical corpus. First we judge that a term from an alien theory fails to correspond to one of our terms and is to be replaced by a neologism; at this stage we do not need to import that term or the beliefs in which it appears. That is because the judgment that a neologism is needed is sufficient demonstration that those beliefs are not ones we share. The very fact that we resorted to a neologism to translate a sentence of theirs implies that that sentence is not counted among those to which we subscribe. For theory comparison, we merely register the fact that a neologism is needed, and proceed to declare all associated beliefs unshared. It is a separate question whether we should adopt any or all of

those tenets. For purposes of theory choice, neologisms will be ignored when we retreat to neutral ground, since they will not feature in any shared beliefs. They will be considered again when deciding which new theoretical tenets to import, if any. However, for the purpose of understanding the neologism, we will need to see how it relates to other terms and how it is situated in the rest of their theory. To this end, we can import this term along with some of its theoretical tenets, which may involve our bracketing some of our conflicting beliefs. We do not adopt their theory for good, but the hermeneutic process of understanding consists in a pretense of adopting part of their theory to apprehend the significance of some of their terms. This reveals a subtle distinction between the practice of the history of science and the problem of theory choice.

These considerations show that in any realistic example, translation entails introducing neologisms. As long as we are only interested in theory choice, these neologisms can be ignored, for we will be concerned to isolate the area of agreement, which will be devoid of neologisms. But if we are interested in understanding the rival theory and appreciating its claims, these neologisms must be understood in context, which may involve temporarily bracketing some of our beliefs or at least supplementing our theory with some of their theoretical tenets. In such a case, it does not seem to be a substantive issue whether one wants to say that one theory has been translated into another or that both have been translated into a third, as long as direct comparison and understanding can take place.

5.8. Principle of Literality

A very general interpretive precept has to do with the exclusive concentration on literal meaning in the interpretation of scientific theories. Even if one accepts that they are simple expressions or compound lexemes (see section 4.), there may still be a temptation to say that terms like 'dephlogisticated air' in the phlogiston theory or 'elementary atom' in Dalton's theory are not correctly rendered by 'oxygen' and 'atom' respectively, because that translation leaves something out. In the first case, it leaves out the origins of the term and its association with the concept of 'phlogiston', and in the second, it leaves out the connection of the term to the concept of 'element'. However, facts about the origins of the terms do not necessarily contribute to the literal meaning of the sentences in which these

terms feature. Rather, they relate to the aspects of usage labelled "connotation" or "nuance" in section 3.4., in the course of discussing Kuhn's views (especially his example of the French term doux). Such aspects can be ignored for our purposes, given the interest in truth conditions and explanatory value.

The relationship between literal meaning and inferential connections has already been mentioned, particularly with reference to the importance of preserving the deductive structure of a theory. I have not offered a full account of the difference between literal and non-literal meaning, mainly because there is no general consensus among philosophers and linguists about how to draw the semantic-pragmatic distinction. But I agree with the view that takes the distinction to be bound up with the notion of truth conditions (as opposed, say, to context-independence or conventionality). There is a growing consensus in the literature on semantics and pragmatics that takes the province of a semantic theory to be truth-conditional and regards the domain of pragmatics to lie beyond what is truth-conditional.[7] Given the close connection between truth conditions and inferential role, the insistence on preserving the deductive (or, more generally, inferential) structure of the theory being translated is part and parcel of the exclusive focus on literal meaning. At the very least, the principle of ignoring non-literal aspects of meaning is consistent with some of the other principles being advocated here.

But how plausible is it that force can be skimmed off the surface of meaning, leaving literal meaning undisturbed? This is an assumption that has been denied by some intellectual historians, notably Quentin Skinner, as we saw in Chapter 4. Skinner and others maintain that central to the whole business of interpretation is the determination of the speech act performed by the agent being interpreted. But for Skinner, at least, this view is accompanied by a denial of something like the semantic-pragmatic distinction. This position seems to be inspired by the stand that Austin ends up by taking at the end of How to Do Things with Words. After beginning by making a clear distinction between

---

[7] Two notable works that fall into this category are Gazdar (1979) and Levinson (1983). In the former work, Gazdar posits that non-literal force is what is left over from Grice's non-natural meaning (the speaker's reflexive intentions) after subtracting truth conditions.

constatives and performatives, Austin ends his work by casting doubt on the feasibility of drawing a line between the two, and declares that asserting is just one more species of performative. But this is by no means the consensus among philosophers of language and linguists, and if this claim were accepted, it would make any systematic attempt to theorize about meaning problematic. That is why the distinction is being assumed in this work, although no full-blown attempt to ground it will be made.

Someone might still ask how to go about supplementing a literal interpretation with certain non-literal aspects of meaning. That is, suppose that a literal interpretation has been given, how would one go about adding on such things as implicature, metaphor, illocutionary force, irony, and other rhetorical and non-literal effects? The answer lies in providing what Kuhn calls "glosses and translator's prefaces". (1983a, 672) I would agree with his claim that such glosses are no part of the translation proper, but would add that they are required in supplying non-literal aspects of an alien theory or text. If the conjecture made in the Introduction about the greater literality of scientific discourse is correct, then examples from science are probably rarer than those from social and political theory. A possible example from the history of political philosophy, a subject that would seem prone to examples of non-literal force, can be taken from Hegel's writings. One of the most influential components of Hegel's political theory is the distinction between 'positive freedom' and 'negative freedom'. Many interpreters concur that 'negative freedom' just corresponds to the ordinary notion of freedom found in the works of classical liberal authors such as Hobbes and (later) Mill. If this is the case, then 'positive freedom' would seem to be a separate political concept and it has been rendered by some as 'self-realization', or something of the sort (alternatively, it may be translated as 'positive freedom', as long as this is considered a simple expression rather than a complex one consisting of the concepts, positive and freedom). But, someone might object, this translation misses something important, for Hegel's point was surely that 'positive freedom' was the "true" concept of freedom. Although this was part of Hegel's point, the response to this is that it is part of the non-literal force associated with his theory rather than its literal content. What Hegel is doing at the literal level is to propose a new ideal in the political realm to rival the notion of freedom as found in liberal political philosophy. The added suggestion that this ideal should be graced with the highly prestigious term

"freedom" is part of the non-literal content of the theory rather than part of its literal content. This can be indicated in a Kuhnian gloss or translator's preface, since it may not be obvious from the content of Hegel's theory itself. The proper way to indicate forces associated with theories of this kind is by giving such supplementary information. Kuhn may be correct to say that this is not part of the translation proper, but I am suggesting that the translation proper can only be expected to capture literal meaning.

5.9. Hazards and Pitfalls

I have acknowledged that the principles of interpretation outlined here are not always followed by real interpreters. Nevertheless, I have tried to justify them so that they do not appear to be instances of philosophical legislation that is oblivious of actual practice. Although I hope that the justifications will be convincing, I think it is also worth dwelling briefly on some of the reasons behind the violation of these principles. Since I think that these reasons are not good ones, this is a diagnostic exercise and can be regarded as an attempt to specify some of the pitfalls involved in interpretation.

There are many instances of the violation of the Principle of Conceptual Charity; instances, that is, of the multiplication of concepts beyond necessity. Interpreters of past theories or introducers of new theories will often take them to involve completely different concepts that do not match those of other, more familiar theories. They intimate that there is conceptual change afoot when all that has transpired is theoretical change. This is the case with some presentations of the special theory of relativity that state that two new quantities are introduced by the theory, 'rest mass' and 'relativistic mass', neither of which corresponds exactly to the familiar term 'mass' in Newtonian physics. However, if the analysis in section 4.2. is correct, the proper thing to say is that 'rest mass' in special relativity corresponds to 'mass' in classical physics and that this is a case of theoretical change. Unnecessary neologization, or the misidentification of theoretical change as conceptual change is also sometimes encountered in certain interpretations of other cultures and eras. The reason for the tendency to see conceptual change where only a theoretical change has transpired can also be less strictly philosophical. It is sometimes motivated by an attempt to exoticize or alienate some alternative theory or culture. When one uses a neologism rather than an existing term from one's own conceptual repertoire,

one conveys a sense of abstruseness or esotericism, or else one conveys an essential foreignness that makes it easier to be dismissive of the alternative viewpoint. This can be a result of hostility to the culture or era in question, but it can also be a well-meaning attempt to underline uniqueness in order to valorize that cultural group or historical period. In either case, it is often misguided and can be avoided by deploying conceptual charity.

Here it might be protested that ascribing new concepts cannot be more dismissive than ascribing false beliefs, which is what I have been advocating doing when possible. The ascription of new concepts may be thought not to involve attributing falsehood to our interpretees, thereby doing less damage to their views. Indeed, it may be said that this is the properly charitable course of action. But notice that the ascription of new concepts, on the view that I am advocating, also involves attributing false or truth-valueless beliefs. Any new concept that we need to introduce is one that we do not already have and any sentences in which it features will be false or lacking in truth value, as I already mentioned in Chapter 4. Thus, ascribing new concepts always involves attributing unshared beliefs.

Lest it be thought that interpreters err only in the direction of conceptual addition rather than conceptual subtraction, an example should be given of the other sort of error. There are cases that are touted as mere theoretical changes, but are actually conceptual. A good example of this opposite tendency can be derived from Williams' work on certain key concepts in modern European thought. As seen in section 4.6., the two terms 'literature' and 'philosophy' were used by nineteenth-century capitalists to associate the activities of their companies with those of the cultural elite. In these cases, the publications and corporate strategies of rising industrial concerns were given airs by intimating that they were on a par with higher intellectual pursuits. For a variety of explanatory purposes, this attempt failed. To put it succinctly, there is just not enough in common between the promotional publications of corporations and the works of poets and dramatists; they cannot be subsumed under the same explanatory theories. Certain associations were not widely accepted by the linguistic community and this is why, the terms 'literature' and 'philosophy' became equivocal. As in other cases, we can tell that more than one concept is involved by applying one of the equivocation tests mentioned above. For example, by running a putative deductive argument that contains instances of the single term in its two

different guises, it should become clear whether or not the argument commits a fallacy of equivocation.

To sum up, both the confusion of conceptual change with theoretical change and the reverse confusion of theoretical with conceptual change can be exposed and corrected. The former is often perpetrated by innovators who are eager to pretend that they have more in common with past theory and practice than is found to be the case on closer inspection, as with the nineteenth century capitalists. The latter confusion, the tendency to see conceptual change where only a theoretical change has transpired, is often committed for the opposite reason: to intimate considerable difference or major innovation where less is actually in play. Sometimes what lurks behind it is an attempt to mystify a theory or exoticize a culture on the part of its interpreters. This may even be done by proponents of the theory or members of the culture themselves, who may have a stake in seeing considerable breaks or ruptures between their intellectual or cultural traditions and those of others.

Chapter 6: Concepts

> An Expedient was therefore offered, that since Words are only Names for Things, it would be more convenient for all Men to carry about them, such Things as were necessary to express the particular Business they are to discourse on... [M]any of the most Learned and Wise adhere to the new Scheme of expressing themselves by Things; which hath only this inconvenience attending it; that if a Man's Business be very great, and of various Kinds, he must be obliged in Proportion to carry a greater Bundle of things upon his Back, unless he can afford one or two strong Servants to attend him.
>
> Jonathan Swift, Gulliver's Travels

6.1. Concepts and Extensions

In this chapter, I will attempt to justify further the claim that the way to compare theories is by focusing on their concepts rather than adverting to the reference of their terms. In order to make this case, I will begin by noting that reference has been variously explicated by philosophers. In particular, I will distinguish a metaphysical realist notion of reference from a more simple-minded notion of extension. While the latter is unobjectionable and makes an appearance within the interpretive approach, it does not by itself provide us with a way of comparing scientific theories, for reasons to be explained. As for the former, I will advance some arguments to put it into doubt, at least as a route for comparing scientific theories. To be sure, doubts about a certain conception of reference have already been articulated in Chapter 2, but those doubts will be generalized and underlined in this chapter. I will digress to show that some of the examples commonly taken to motivate a metaphysical realist account of reference (the Twin Earth examples) can be recast in such a way that they no longer do so. It will be crucial to my argument that there is often a difference between expert concepts and lay concepts, and that my concern is exclusively with the former and not the latter.

Having said that theories are to be compared by way of their concepts, it should also be mentioned that concepts themselves are in need of further explication and that the concept-sharing relation is not without its problems. Therefore, I will also devote some

space in this chapter to enunciating a satisfactory account of concepts. One problem that arises for this particular theory of concepts is the breakdown of transitivity in the ascription of concepts; that is if term a from one theory picks out the same concept as term b from another theory, and term b picks out the same concept as term c from a third theory, it does not always follow that a and c pick out the same concept. In light of the breakdown of transitivity, I will defend the view that concepts are entities not objects. And though they should be thought of realistically, they ought not to be reified. Finally, I will relate this account of concepts to some that have been recently proposed in the cognitive sciences. There is an affinity between the view of concepts employed in this work and an influential psychological theory of concepts. Moreover, there are other cognitivist theories of concepts that can be said to be operating at different levels of description and isolating different entities. Thus, although they might appear at first to be incompatible, there is considerable agreement between this philosophical attitude to concepts and some of those based on empirical or computational research.

6.2. Reference vs. Extension

There is a logician's fantasy according to which individuals and sets are labelled with names and predicates like so many jars on a shelf. The fond hope is perhaps to dispense with words entirely and traffic only in things, like the professors at the Academy of Lagado in Swift's satire, which provides the epigraph to this chapter. Whatever its heuristic value in the pages of logic texts, this simple-minded referential picture has proved to be misleading in many philosophical contexts. One score on which it appears lacking is the task of comparing scientific theories. Even before the introduction of the causal theory of reference, some philosophers hoped that scientific theories might be compared simply by comparing the extensions of their terms. They sometimes talked as if one could line up the terms of the two theories, locate their extensions and proceed to compare them. But the unfeasibility of this course of action should be apparent, not least because of the problems with ostension which were aired in section 2.5. One cannot just pick out the extension of a scientific term by pointing to a set of entities.

Extensions are not ignored in this account, but the notion of extension that is being endorsed here is simpler and more straightforward than the one often encountered in the

philosophical literature. One determines the extension of a scientist's term in much the same way as one determines the extension of any predicate: by taking note of when the scientist is willing to apply the term and when he or she withholds it. In other words, the extension of a certain term as used by a particular agent is to be understood in a straightforward manner, as in the following account from Carnap:

> It is generally agreed that, on the basis of spontaneous or elicited utterances of a person, the linguist can ascertain whether or not the person is willing to apply a given predicate to a given thing, in other words, whether the predicate denotes the given thing for the person. By collecting results of this kind, the linguist can determine... the extension of the predicate 'Hund' within a given region for Karl, that is the class of things to which Karl is willing to apply the predicate... (1955, 235)

When it comes to scientific theories and their proponents, how can one determine the class of things to which one is willing to apply the predicate? As Peter Smith has pointed out, in the case of past scientists one often has access to both linguistic evidence (published writings about the theory, written communications with others, laboratory notebooks, and so on), and non-linguistic evidence (scientific instruments, prepared samples, and so on), to adduce the extension of their predicates. (1981, 63-65) Knowledge of the extension of a term used by a scientist can serve as one clue to the meaning of that term and the content of the theory; and the content of the theory can in turn be used to pick out the extension of its terms.

This means that the extension of a scientific term cannot be determined independently of the beliefs of the scientist or the content of the relevant scientific theory. This distinguishes extension from reference as it is understood by the causal theory of reference, which conceives of it as being theory-independent (see section 2.6.). Since extensions are theory-dependent on my understanding, they cannot be discussed without bringing in questions of meaning. This claim is brought out vividly in some observations of Hempel's, who seems to conceive of extensions in a similar manner. After stating that the notion of sameness of meaning is intrinsically unclear, Hempel maintains that matters are not improved if one brings in extensions, since "it is not clear how the extension of a term as used in a given theory is to be characterized in a nontrivial way." (1969, 260) He then elaborates on the problems involved in comparing extensions, insisting that they are akin

to those involved in comparing the meanings of theoretical terms. Consider the term 'electron' as used in Bohr's first theory of the hydrogen atom and as used in contemporary physics. If one thought that the two terms had the same extension, one would need a specification of their respective extensions. But then one would need to know which sentences of the two theories count as determining those extensions. If one ruled that all sentences count, the terms will not be coextensive since the two theories are incompatible. (1969, 264) Hempel proceeds to mention other problems in determining extensions, which are reminiscent of those associated with comparing the meanings of terms drawn from different theories. That is because neither meaning nor extension is determined independently of the content of the theory in question.

When they are construed in this manner, extensions are one source of evidence for the ascription of meaning. That is, they help the interpreter in making the decision to map a term from one theory onto a term from another (in accordance with the Principle of Warranty from section 5.5.). But, in keeping with holism, the content of an agent's beliefs and the meanings of an agent's terms also provides us with clues as to the extensions of those terms. The extensions of a scientist's terms can be ascertained by way of the agent's actions and beliefs, not least the indexical beliefs among them. They may also be got at by observing scientific experiments or by reading lab reports. But when identifying these extensions, it is important that they are not discriminated any more finely than the agent does and that any discriminations made by the agent are preserved.[1] The interpreter pays attention to the extensions of agents' terms provided distinctions are not made that are more fine-grained or more coarse-grained than those made by the agent, as revealed by the agent's other relevant beliefs and actions. When interpreting a scientific theory, we adopt the perspective of our own theory, but in deciding whether their terms correspond to terms of ours, we are careful to preserve all the distinctions made in their theory and to

---

[1] Compare Bilgrami: "When fixing an externally determined concept of an agent, one must do so by looking to indexically formulated utterances of the agent which express indexical contents containing that concept and then picking that external determinant for the concept which is in consonance with other contents that have been fixed for the agent." (1992, 5)

make none that they do not make. The interpretation picks out both mental contents and extensions in a way that captures the agent's conception of things ("narrowly", to use the philosophical jargon), rather than with reference to the external causes of contentful states ("widely").

The account of extension that I have just sketched is clearly distinct from the account of reference provided by the causal theory of reference. As I pointed out, the crucial difference is that on my account, extensions are not picked out independently of the scientist's surrounding theory. The extension that one hits upon is mediated by the agent's beliefs, in the sense that all the distinctions made by the scientist are preserved and none of the distinctions not made by the scientist are ascribed. But to say that extensions are theory-dependent is not to deny that a wedge can be driven between the extension and the content of the theory. The extension and the theory or concept may be out of step, since one might over- or under-extend a concept, or misapply it on occasion. In some cases, getting the extension right provides partial grounds for ascribing the relevant concept, and conversely, picking out the wrong extension is a partial consideration for withholding the concept (for example, Dalton's term 'molecule' was not translated homophonically in section 4.4. partly because he picked out the wrong extension using the term). But in other cases, we tolerate a certain degree of extensional divergence and ascribe a particular concept in any case because of agreement on certain crucial beliefs (for example, the concept atom was ascribed to Dalton despite the fact that he thought the basic particles of gaseous oxygen were atoms rather than molecules, so he got the extension partly wrong).

This shows that one can legitimately talk about the extensions of scientific terms within the scope of the interpretive approach, provided such extensions are not thought of as theory-independent. In particular, they should not be picked out in such a way as to make distinctions not made by the agents themselves or in such a way as to ignore distinctions made by them. At the same time, some allowances can be made for under-extending and over-extending a particular concept, and the relevant concept may be ascribed even though the extension of that concept is not quite correct by our lights. Such conduct may occasionally be recommended by the principles of interpretation outlined in Chapter 5.

6.3. Metaphysical Realist Reference

In the philosophical literature, the straightforward notion of extension discussed in the previous section is insufficiently distinguished from the metaphysical realist notion of reference. As I have already suggested, the metaphysical realist conceives of reference as a belief- or theory-independent relation between words and the world. Since it is a matter of a brute link between language and reality, this referential relation is also supposed to be suitable for thinking about counterfactual examples, for it enables one to determine what an agent would (counterfactually) have referred to in another possible world--irrespective of what is believed by the agent in that world (or this one) and what happens to be true in that particular world. That is what allows terms to be rigid designators, picking out the very same thing or things in all possible worlds.

The point of this section is to demonstrate further the problems with a metaphysical realist view of reference. In section 2.4., I argued (using the example of Paulette) that the causal theory of reference not only failed to ascribe beliefs with psychological efficacy, it also failed as an account of what it was heralded as, namely an account enabling the comparison of agents or theories. This seems to be a general problem for metaphysical realist views of reference, and to help strengthen the claim, I will show in this section how similar problems arise for two other metaphysical realist views of reference, those of Hartry Field and David Lewis. In the next section, I will also address two of the more influential counterfactual examples to see how they might be handled without appealing to such a notion of reference.

The characteristic feature of these two approaches to the problem of theory comparison is that they have affinities to the interpretive approach but are wedded to a strong referential picture. These views do not suffer from the same specific problems encountered in discussing the causal theory of reference, but they fail for related reasons having to do with the notion of reference at play. The criticisms of the two theories will help to underline the distinctive aspects of the approach being advocated in this work and to vindicate further the claim that one compares theories by focusing on their concepts or the meanings of their terms, rather than the reference of those terms.

The first contrast to be drawn is with Field's theory of "partial denotation", which was encountered in Chapter 4. Recall that Field proposes to define a function, which he

calls a "structure", for each scientific theory. The function would map every name onto an object, every quantity term onto a quantity, and every predicate onto a set. Such a structure, m, enables one to say that sentences of the interpreted theory are m-true (m-false) depending on whether the sentence would be true (would be false) if the denotations and extensions of its terms were as specified by m. As we have already seen, he claims that mapping 'mass' in Newton's theory onto the quantity of proper mass gives rise to some true sentences of Einstein's theory, and concludes that that is evidence that 'mass' partially denotes proper mass (the same applies when Newton's term 'mass' is mapped onto the quantity, relativistic mass). Field's approach shares some aspects of the present approach. But the crucial difference is that the interpretive approach privileges that substitution which leads to the optimal mapping, as spelled out in previous chapters. Since I have claimed that there will be a single optimal solution, there is no need for partial denotation. Field is saying that any mapping that makes any number of sentences true is a successful one in some sense, since it provides evidence of partial denotation. In other words, since there are generally a number of mappings that are successful in yielding some true sentences, they should all be taken as being partially denotational. But then, would he allow even those that yield only a paltry number of truths, for example the mapping that took classical 'momentum' to relativistic 'energy' (after all, both quantities are conserved in a certain class of interactions)? In most actual cases, partial denotations are likely to be too plentiful to have explanatory value.

This suggests a deeper contrast with the interpretive approach: Field does not say that there is an implicit background theory to which we are comparing the theory being interpreted. On his view, assignments of truth value are not made relative to some theory or another; all we need are objects and extensions. His commitment to partial denotation seems to derive in part from the idea that reference is an unmediated relation between words and the world, and that the truth of certain sentences serves as evidence for the existence of this relation, which has independent metaphysical reality. When this assumption is dropped, all that remains is a function mapping terms of a theory T2 onto terms of another theory T1. But in that case, it is misguided to attribute "partial denotation" to any assignment that yields some true sentences.

Another useful contrast can be made with Lewis' proposal for defining theoretical terms. Lewis draws on Ramsey's method, but in a twist on the theoretical-observational distinction, he postulates that the theory being interpreted contains two sets of terms, T-terms and O-terms, characterized as follows. A T-term is "a theoretical term introduced by a given theory T at a given stage in the history of science," and an O-term is, by elimination, "any other term, one of our original terms, an old term we already understood before the new theory T with its new T-terms was proposed." (1970, 428) Clearly, this assumes that a homophonic translation is in place for all the terms that belong to both theories (for that is just what "old" terms are), an assumption that is unwarranted. One cannot merely go by the shape or sound of the terms of a new theory in deciding whether they should be understood as they were in the old theory. And if one relies on the intentions of the theorists who propose the new theory, these cannot be ascertained without a proper interpretation of the theory in question.

But Lewis' method has another important defect that brings out a contrast with the present approach and demonstrates his commitment to a metaphysical realist view of reference. First, Lewis makes use of Carnap's proposal to write a theory as a conditional of the theory's Ramsey sentence and the theory itself (written in a way that exhibits the occurrences of the T-terms; see section 1.3. for details). Then he states that there are three possibilities: 1) the theory is uniquely realized, 2) the theory is not realized, and 3) the theory is multiply realized. Now if there are n T-terms, Lewis claims that each can be defined as the entity that, along with n-1 other entities, comprises an n-tuple identical with all and only n-tuples that realize T. That is, each can be defined as the ith component of the unique realization of T. Immediately, this raises the question of what is to be said when the theory is not uniquely realized (i.e. not realized at all or multiply realized). Lewis is quite clear on this score, for he states that, "the theoretical terms of unrealized theories do not name anything." (1970, 432) Similarly, he writes: "[T]he theoretical terms of multiply realized theories are denotationless." (1970, 433)

The first of the above consequences of Lewis' theory is unsatisfactory.[2] It says basically that all the T-terms of unrealized theories fail to refer. As Lewis explains:

> The T-terms were introduced on the assumption that T was realized, in order to name components of a realization of T. There is no realization of T. Therefore they should not name anything. 'Phlogiston' presumably is a theoretical term of an unrealized theory; we say without hesitation that there is no such thing as phlogiston. (1970, 432)

However, this leaves out the treatment of more problematic terms such as 'dephlogisticated air' which, although they belong to "unrealized" theories, may well be interpreted as referring all the same. The problem is especially acute with theories that Lewis calls "near-realizations". In such cases, he writes: "We might want to say that the theoretical terms name the components of whichever n-tuple comes nearest to realizing the theory, if it comes near enough." (1970, 432) This solution is messy and detracts from the apparent simplicity of Lewis' view. But the main point is that it suffers from the assumption, avoided by the interpretive approach, that two theories can only share a concept if they agree on all sentences in which the appropriate term features. Although Lewis' proposal is intended as a method of defining theoretical terms, it does not seem suitable as a device for comparing theories. That impression is confirmed from what he goes on to say.

---

[2] The other consequence, that the T-terms of theories with multiple realizations fail to refer, is less problematic. To make it more palatable, Lewis suggests that multiple realization is rendered unlikely by the fact that the interpretation of the O-terms must remain fixed for all the multiple realizations, noting that "this O-vocabulary may be as miscellaneous as you please..." (1970, 433) This way of putting things is unsatisfactory because of the nature of Lewis' distinction between O-terms and T-terms. As already mentioned, it is implausible to assume always that all the terms of one theory that also occur in another are to be matched up homophonically. However, if a large set of non-problematic terms has been matched up according to the method being suggested in this work, it does seem safe to say that it is unlikely that multiple realizations of a theory might be equally adequate and that nothing could be found to choose between them.

A similar objection to the above was made by Putnam against Carnap's method of partial interpretation. As Lewis mentions, Putnam objected that theories with false observational consequences are "wrong, not senseless". But Lewis' rejoinder is that the theoretical terms of such a theory are not senseless, just denotationless. Their sense is given by their denotation in possible worlds in which the theory is uniquely realized and does not have false consequences. (1970, 435) For Lewis, the sense of some property term Φ_1 is a function that assigns to any world w the property named by Φ_1 in world w. However, he is not forthcoming when it comes to specifying how to identify the very same property in another possible world. Unless one specifies how this is done, Putnam's objection retains its force. For example, the problem of comparing classical mechanics to special relativity theory can be stated in these terms: In a possible world in which Newton's laws are obeyed, which physical property is to be identified with proper mass (as it occurs in the special theory of relativity)? Not only does Lewis state that all T-terms of an unrealized or nearly-realized theory fail to refer, he does not tell us how to go about determining their sense (by way of their reference in other possible worlds). He simply assumes that we have an independent handle on the property in question in those possible worlds in which the theory is realized. Therefore, he has not shown how a non-realized or nearly-realized theory can be compared to a theory that is realized.

There is a common failing to these metaphysical realist views of reference. Both have trouble accounting for the fact that terms from non-realized theories (i.e. theories that differ from the true theories) nevertheless succeed in referring and can be compared with terms from the realized theory. Field's remedy is to abandon reference for partial denotation, with the consequence that a single term will come out partially denoting different things and that different terms from the same theory will come out partially denoting a single thing. Lewis' route is to acknowledge that terms from false theories do not refer, but to say that such terms have another kind of meaning, sense. But he is unilluminating when it comes to saying something specific about their sense. Both analyses lose sight of the philosophical woods for the technical trees. The main point of "defining" or specifying the referents or extensions of theoretical terms in science is surely the ability to compare scientific theories. However, in trying to anchor terms to objects or

properties "out there" both Field and Lewis forego the ability to carry out a comparison of scientific theories.

Despite these criticisms of two metaphysical realist referential views of scientific terms, one should not think of the interpretive approach as being anti-referential.[3] I indicated in the previous section how this approach to comparing scientific theories incorporates extensions in its methodology. A proviso was added to the effect that the extension cannot be disengaged from the content of the theory to which it pertains and that extensions should not be construed in such a way as to make distinctions that are finer or coarser grained than the ones made by the scientists themselves. This section has further argued that when reference is understood in terms of a metaphysical relation between words and the world, it ceases to be of use in comparing theories or systems of beliefs and should therefore be disregarded for these purposes. But one can still integrate a straightforward notion of extension into the interpretive approach.

6.4. Twin Earth and Other Fables

No inquiry into meaning is complete these days without a discussion of interplanetary travel. I have rejected a metaphysical realist notion of reference, but can the interpretive approach deal with the famous Twin Earth cases? These stories, made popular by Putnam, concern a planet that is similar to earth in all respects, except for the fact that all the H2O is replaced with a substance with a very different chemical composition, say XYZ. The interesting thing is that XYZ looks to the untutored eye just like H2O; indeed, it

---

[3] A willingness to talk about extensions or referents may seem to distinguish my approach from other interpretivist ones. After all, one of Davidson's papers is entitled "Reality without Reference." But the message of that paper is that "reference cannot be explained or analysed in terms more primitive or behavioral," not that one should cease to talk about reference altogether. (1977a, 215) He writes: "If the name 'Kilimanjaro' refers to Kilimanjaro, then no doubt there is some relation between English (or Swahili) speakers, the word, and the mountain. But it is inconceivable that one should be able to explain this relation without first explaining the role of the word in sentences, and if this is so, there is no chance of explaining reference directly in non-linguistic terms." (1977a, 220)

has all the same superficial properties and can only be told apart by sophisticated chemical analysis. For the layperson, however, the difference does not show up, and the question is: To what do our lay doppelgängers on Twin Earth refer when they use the term 'water'? The correct answer is supposed to be that they refer, not to H2O as we do, but to XYZ. Since their psychological states are the same as ours, the example purportedly shows that we need a referential component of meaning which is independent of the psychological or intensional one, a claim encapsulated in the notorious slogan: "meanings ain't in the head." The example can be used to motivate the causal theory of reference, with its insistence that reference is a belief-independent relation, determined by a "wide" causal-historical connection between the agent and the world rather than by the agent's "narrow" mental state.

How does the interpretive approach answer the question about the reference of the Twin Earthian terms? By saying that when they use the term 'water', our doppelgängers refer to water. They refer neither to H2O nor to XYZ, and moreover, when we use the term 'water' we refer to water also. In fact, we have <u>two</u> distinct concepts: <u>water</u> and <u>H2O</u>. The former concept is a commonsensical one derived from a certain naive theory of many ordinary substances and is pertinent to everyday interests of ours revolving around food, hygiene, shelter, and so on. The latter is a (composite) concept that derives from a sophisticated molecular theory of chemistry that populates the world with elements and compounds, and explains their reactions and properties. To help see that the two concepts are distinct, notice that the extension of what the layperson calls 'water' is not the same as the extension of what the chemist would call 'H2O'. Many samples of water are not the chemist's H2O, but an aqueous solution of one kind or another. For instance, sea water is a solution of sodium chloride and other compounds in H2O, as is mineral water. In fact, when samples of water approach chemical purity the common folk sometimes use the term 'distilled water' to refer to them. Moreover, water is not just an impure form of H2O, for certain impurities are tolerated by our commonsense theory, while others are not. Many mineral impurities, such as that of sodium chloride, are not considered to affect the applicability of the term, whereas others, such as ground coffee beans or tea leaves, are. In addition, there are many samples of chemically pure H2O that we would not call 'water', notably samples of ice and steam, and samples of other substances that we would, notably

heavy water (with deuterium in the place of hydrogen). Moreover, it is not just that the two concepts have different extensions, but that they pertain to different theories. It may be thought that the chemist has one theory and the layperson has another, but it would be more accurate to say that the chemist holds both (compatible) theories at once, one qua chemist and the other qua layperson. The commonsense concept of water did not drop out with the introduction of the concept of H2O. And although there are certain connections between the two concepts, they are by no means identical.

Similar points regarding water and H2O have been made by Chomsky. He notes that if a cup contains pure H2O into which a tea bag has been dipped, it is tea, not water, though it could have a higher concentration of H2O molecules than what comes from the tap or is drawn from a river. (1995, 22-23) In addition, some empirical psychological evidence gathered by Barbara Malt can be used to bolster the claim of a conceptual difference between water and H2O. In a paper provocatively entitled "Water Is Not H2O," Malt (1994) found that when it came to determining which liquids were normally called 'water' by lay subjects, neither their beliefs about the simple presence or absence of H2O nor their beliefs about the proportion of H2O in a variety of liquids accounted well for the application of the term. This not only supports the claim that the two concepts are different, it also casts doubt on psychological essentialism, the view that psychological subjects operate by and large with an essentialist theory of the world. Therefore, I am denying the alleged truism that 'Water is H2O', which has been hailed as the preeminent example of a scientific essentialist truth. Very roughly, we can say that scientists have determined that water consists predominantly of liquid H2O when purified in certain ways. Still, the concepts of water and H2O are quite distinct and coexist comfortably in our total theory of the world. They belong to crosscutting taxonomies that are not rivals.[4]

The relationship between commonsense and scientific taxonomy has been explored by Scott Atran, in a work that argues that many folk taxonomies survive and exist comfortably alongside scientific taxonomies. This fact is often obscured because we

[4] For more examples of crosscutting taxonomies and further explanation, see Khalidi (1993a), (1998a), and Chapter 7.

assume that the mere appearance of a scientific classification of a particular domain leads automatically to the displacement of the common-sense classification. By focusing on biology, Atran argues persuasively that this is not always true historically, not even in the modern (supposedly scientific) era, since even today common-sense meaning is not directly tied to scientific meaning. He states that, "If laypeople accept modification of a folk taxon, it is because the scientific taxon proves compatible with everyday common-sense realism; if not, the scientific concept can usually be set aside, and the lay notion persists as a 'natural kind' regardless." (1990, 6-7) Atran also indicates that science and common-sense often coexist amicably side by side without clashing. Thus, the transition from natural history to biology "involved not so much a radical rupture with common sense, as maintaining a continuing access through its reevaluation." (1990, 13)

But what happens when we do not have the luxury of a folk concept? Surely there are Twin-Earth cases in which it would be a distortion to say that the folk concept is different from the scientific one. Such a case is provided by another famous example that has been taken by some to show that we need two components of meaning. Tyler Burge's character Bert is not only afflicted with a disease of the muscles of his thigh, he also has the misfortune of misdiagnosing it as arthritis.[5] As we all know (presumably), arthritis is a disease of the joints not the muscles, so he cannot really have arthritis in his thigh. However, according to Burge, in reporting Bert's belief, we would naturally say: Bert believes he has arthritis in his thigh. That is, we ascribe to Bert our concept <u>arthritis</u> and say that he has a false belief involving it. By contrast, an identical twin of Bert's, who believed the very same thing in a community where there was an arthritis-like disease that did indeed strike the thigh and was called 'tharthritis' by the experts there, would have been ascribed the concept <u>tharthritis</u>. Different ascriptions are made although "what is in the head" is the same.

In this case it is not open to us to say that there is a folk concept of an <u>arthritis-like disease</u> that can be ascribed to both Bert and his twin. It is implausible to suggest that we have two concepts in our general theory of the world in this example, in the way that we

---

[5] For details, see Burge (1979).

have the distinct concepts of water and H2O. Most naive concepts of disease seem to give way to sophisticated ones, rather than continue to exist alongside them. There may be some exceptions, such as 'jaundice' and 'diarrhoea', but arthritis does not seem to be one of them.[6] Rather, the more plausible analysis treats both agents as being parasitic on their respective communities. If their only beliefs are the ones reported, we are clearly only ascribing the respective concepts by courtesy, as it were. This is the phenomenon that Putnam once dubbed "the linguistic division of labor", though in recognizing it, one need not adopt his account of the phenomenon, which involves the causal theory of reference. As I mentioned earlier, this inquiry does not address the linguistic division of labor, since it is concerned with the terms of the experts themselves, rather than those of neophytes who rely on experts to make discriminations between elms and beeches, stoats and weasels, aluminum and molybdenum, and other superficially similar natural kinds. As these cases are usually described, Bert-like laypersons do not have a rudimentary theory that warrants our ascribing to them the relevant concept, yet referential success and even conceptual competence might be awarded by courtesy. The basic idea is that laypeople have the requisite theory potentially , since they are able to ask the experts for a fuller theoretical explication of the term and to defer to them if required to supply some associated beliefs. I will not attempt to spell out in detail how this occurs, since the full account surely belongs to the domain of sociolinguistics, but there is no reason to think that the linguistic division

---

[6] It is not obvious why diseases should be less resilient to scientific advances than other folk categories and why commonsense disease categories generally give way to scientific ones. Atran suggests briefly that concern with taxonomic nosologies (classification of diseases) is a specialized affair "by and large restricted to doctors and naturalists of the seventeenth and eighteenth centuries, and to twentieth-century ethnolinguistics and ethnomedicine. Most folk have no need or use for it." (1990, 311-312) The fact that the treatment of disease in many cultures is left largely to experts may help to explain this lack of interest and the corresponding ease with which many traditional folk diseases have dropped out and given way to scientific categories. Notice, moreover, that the purposes and interests of the experts are by and large the same as the folk, namely the treatment of diseases. For more on the role of interests in individuating a scientific domain, see section 7.2.

of labor cannot be accommodated within a descriptivist or belief-based account of meaning; it does not seem to necessitate a causal or directly referential theory.[7]

If developed scientific theories and other systematic corpora of beliefs are our primary concerns, degenerate cases and ones of minimal beliefs will not even arise. In comparing scientific theories, we are always dealing with theories held by experts. One should resist the temptation to appeal to a metaphysical realist theory of reference to explain the ability of laypersons to refer using terms borrowed from the experts. The linguistic division of labor is adequate to do the job and does not by itself seem to imply a non-descriptivist theory of reference or meaning. But what if the linguistic labor is not divided and the experts themselves are in such a situation? Such are the cases that the Principle of Warranty is designed to deal with (see section 5.5.). If the scientists being interpreted have minimal beliefs associated with a particular term, then we ascribe the corresponding concept to them only when there is sufficient warrant for doing so, as explained in the previous chapter.

6.5. Failure of Transitivity

Now that the interpretive approach to meaning has been further defended and some of the competing referential pictures opposed, the time has come to explicate in more detail the notion of meaning or concept being used here. There is a feature of the interpretive approach, not yet expanded upon, that can be used to shed further light on this. That is

---

[7] A descriptivist construal of the phenomenon of the division of linguistic labor can be found in the work of a host of philosophers. Papineau writes: "But this illuminating thesis of the division of linguistic labour should not be considered, as it often is, as any argument for the causal theory of reference. For it is perfectly consistent with any account of what makes a term as used by experts refer to what it does." (1979, 168) Compare Mellor: "It need not be my beliefs that fix the reference or extension of terms which I can use quite well in my limited way. So I defer to experts, whose job it is to say what such a term really applies to. The reference or extension in any possible world of the term as we use it may nevertheless still be some Fregean function of our experts' beliefs." (1977, 304) See also Dummett (1973, 138-9), Smith (1981, 75-6), Bilgrami (1992, passim), and Chomsky (1995, passim).

what I have called the failure of transitivity in the ascription of meanings or concepts. When theories are compared according to the method advocated in the last three chapters, the following scenario is possible. Imagine that theory T1 is compared with theory T2 and that their terms are matched up in a particular way, and then T2 is compared with another theory, T3. In general, after such multiple comparisons, transitivity in the matching of terms may not be preserved. Let a, a term from the first theory, be matched up with b when that theory is compared with a second theory, and suppose that in comparing the second theory with a third theory, b is matched up with some term c. There is nothing in this method to guarantee that a comparison of the first and third theories would lead to matching a with c. That is what the failure of transitivity amounts to.[8]

This is not just a hypothetical possibility; it can be illustrated with reference to one of the case studies discussed in previous chapters. Consider the phlogiston theory: let T1 be its most primitive form as introduced by Stahl, let T2 be the more sophisticated version of the theory found in the work of Priestley and others, and let T3 be post-phlogiston chemical theory. Now take the relevant terms from these three theories to be 'dephlogisticated air', 'dephlogisticated air', and 'oxygen', respectively. First, it seems plausible (though this was not argued for in Chapter 4) that in a comparison of Priestley's theory with Stahl's, a homophonic translation of 'dephlogisticated air' would be in order. Second, recall the claim that Priestley's term 'dephlogisticated air' should be mapped onto the post-phlogiston term 'oxygen'. These two assertions might lead one to expect (by transitivity) that Stahl's term 'dephlogisticated air' should be translated as 'oxygen'. However, it was explicitly argued that that was not the case, since at the time of its introduction, it had an occurrence in a single belief. This example can be used to illustrate a breakdown in transitivity in the translation of terms, and hence, in the ascription of concepts. One immediate consequence of this breakdown is that, strictly speaking, one

---

[8] It is worth pointing out that this feature is not unique to the interpretivist account of concepts. It seems also to be a feature of virtually any cluster theory of concepts, though the fact does not seem to have widely recognized or addressed. Moreover, it would also seem to apply to the Prototype Theory of concepts, which has been widely influential in psychology and cognitive science, and will be discussed in the following section.

cannot speak of the identity of scientific concepts on the view being presented here, since transitivity is part of the definition of the relation of identity.[9]

The breakdown of transitivity in translation may bring on the charge of anti-realism about concepts. Briefly, the objection might go as follows. Failure of transitivity for concepts implies an inability to formulate strict identity conditions for concepts. That is just because, according to the rules of logic, the logical relation of identity is one that is symmetric, reflexive, and transitive. Moreover, if as Quine insists, there are no entities without identity, then there are no such entities as concepts, at least not as the interpretivists characterize them. But if the interpretivist view leads to the conclusion that concepts are not real entities, the objector can say that this shows that there is something wrong with this account of concepts. Schematically, the argument looks like this:

Transitivity breaks down for concepts. (according to the interpretive approach)
There is no identity without transitivity. (according to the rules of logic)
∴ There is no identity for concepts.
There are no entities without identity. (according to Quine's dictum)
∴ Concepts are not entities.

There are two responses that can be made to this objection; the first exploits an analogy, while the second is more direct. An analogue to the relation of sharing a concept is the relation of membership in a biological species. On the dominant "definition" of species, two populations belong to the same species if and only if their members can interbreed to produce fertile offspring. In general, it turns out that there can be three populations of organisms, A, B, and C, such that A interbreeds with B, and B interbreeds with C, but A and C fail to interbreed. In such cases, biologists say that A and B belong to the same species, and B and C belong to the same species, but that A and C do not. As Ernst Mayr puts it: "Widespread species may have terminal populations that behave toward each other as distinct species even though they are connected by a chain of interbreeding populations." (1963, 536) This means that belonging to the same species is not a transitive relation, and yet this is not seen to raise foundational problems for biology or for the notion of a species.

---

[9] This point was first made to me by Saleh Agha.

By the same token, if we find that the translation or concept-sharing relation is not transitive, this should not raise foundational problems for psychology or the theory of meaning or concepts.

A similar pluralistic view of ontology has been explicitly advocated by E.J. Lowe, who has defended the practice of admitting things in our ontology that do not have determinate identity conditions. Rather, his view is that "whether objects of a given kind should be thought actually to exist should, in general, turn on considerations of whether an inclusion of such objects in one's ontology has explanatory value." (1995, 513) Accordingly, Lowe makes a metaphysical distinction between objects and entities: the former have determinate identity conditions, while the latter do not. He goes on to propose examples of things that do no have determinate identity conditions which we happily include in our ontology. Among them are sub-atomic particles, since "identity statements concerning them can genuinely be indeterminate." (1995, 512) An orthogonal distinction to the object-entity distinction can also be made between concrete and abstract things: the former are spatiotemporal in nature while the latter are not. Employing these two metaphysical distinctions, I conclude that concepts can be admitted into our ontology as abstract entities.[10]

But rather than rest with this analogy, it is also worth considering whether failure of transitivity is a particular problem for the practice of ascribing concepts in the interpretive framework. That is, we should see whether failure of transitivity is a problem for the disciplines that utilize meanings and concepts. First, note that ascriptions of meaning are always made for a subject relative to an interpreter. Transitivity is not called on to do explanatory work in ascriptions of content, since even when the actions of two agents towards each other are explained, we are implicitly explaining the actions of each relative

---

[10] Lowe's pluralistic attitude towards ontology has affinities with what Dupré has dubbed "promiscuous realism", in his (1981) and (1993). Ontological pluralism also seems to be in the general spirit of a famous dictum of Quine's: to be is to be the value of a bound variable. Notice, however, that the pluralism implicit in this dictum is at odds with the stringency of Quine's other dictum cited above (no entity without identity). But an investigation of the seeming tension between these two Quinean theses is a topic for another discussion.

to the interpreter. In order to frame folk-psychological generalizations, we do not need concept-sharing to be a transitive relation, because transitivity is not what grounds generalizations. All we need are pairwise translation relations between the interpreter and a number of interpretees. For example, if we interpret scientist A to believe that all electrons have negative charge, and also interpret scientist B to believe that all electrons have negative charge, we can still say that A and B believe the same thing and explain why they act in certain ways. That is because content ascriptions are made from our point of view as interpreters. Something about A's utterances or actions has led us, after collecting the evidence, to make this ascription to A, which involves attributing our concept electron. The same is true for B's mental life. Suppose we observe that a number of our interpretees believe that electrons have negative charge and that they also believe that protons have positive charge. We can still emerge with the prediction that those agents who believe that electrons have negative charge and protons have positive charge will also (by and large) believe that electrons and protons repel each other. If that is not borne out, there may be something wrong with our initial interpretations. Notice that the issue of how those agents would interpret each other, or whether they would ascribe the concepts electron and proton to one another, is a different one. If we need to know how A would interpret B, that can only be determined from A's perspective, not from our own perspective. To find out, we can proceed to adopt A's perspective and then go on to interpret B from that perspective. This would require us first to interpret A, then adopt A's perspective, and finally to interpret B from that standpoint.

In view of the breakdown of transitivity, one cannot strictly speaking talk about concepts being identical, but one can still speak of the sharing of concepts. Sharing a concept turns out to be a complex and derivative relation between psychological agents, more like sharing a hobby (where there is no single thing held in common), than sharing a house (where there is a single thing) or sharing a first name (where there are different tokens of the same type of thing). To say that a certain agent has a particular concept is to summarize a great deal of information concerning that subject's utterances and actions. It is not to say that there is a single thing or a single type of thing that both agents possess. That is part of the reason that concepts should not be reified and we should not expect them to correspond to determinate structures in the agent's brain. To say that concepts

ought not to be reified is not to say that talk about concepts ought not to be construed realistically. I argued in section 3.6. that the notion of a concept can be made respectable within the interpretive approach, despite Davidson's arguments concerning the vacuity of the idea of a conceptual scheme and notwithstanding the inextricability of meaning and belief. But it does not follow, just because we can talk safely and realistically about concepts, that we should think of them as concrete things. In line with the distinction between objects and entities, we should say that although concepts can be thought of realistically as abstract entities, they should not be reified as concrete objects. As for the suspicion that this attitude towards concepts will not sit well with cognitive psychologists and other researchers who regularly treat concepts as concrete objects rather than abstract entities, I will attempt to counter it in the following three sections.

6.6. Concepts in Cognitive Psychology

Psychologists tend to consider concepts to be something like mental conceptions or representations in the mind or brain of the cognizing subject. One influential psychological theory posits them to be "prototypes", weighted collections of features that serve to explain typicality effects in cognition. A more recent psychological view takes them to be less self-contained, something more like entities embedded in explanatory theories. Yet another view in the psychological literature construes them as "mental models", mental analogs of actual states of affairs. Meanwhile, in artificial intelligence, the symbolic approach has sometimes identified them with "scripts" or "frames", whereas the connectionist literature takes concepts to be patterns of activation in a neural net. This section will not attempt to make sense of the recent flurry of competing cognitivist views, since it is beyond the scope of this work to examine them all. Rather, I will further defend the account of concepts being proposed in this treatment of conceptual change by arguing that there is some convergence between this account and a recent psychological account of concepts. Then, in the following section, I will question the psychological claim of local incommensurability between children and adult concepts.

As used throughout this work, the term 'concept' is interchangeable with the term 'meaning', allowing for some syntactic and stylistic infelicities. If an alien term means the same as one of our home terms, then the two theories share a concept. That should not be

such a controversial assumption to non-philosophers, since that is the way the term 'concept' is often used in the literature from psychology and cognitive science.[11] But concepts are sometimes reified in those disciplines, even treated as physical objects manifested in the agent's brain. While there is no consensus on what concepts actually are, some cognitive scientists seem to treat them more concretely than I have been, for example, as compact lists of features or collections of exemplars (the Prototype Theory). However, recent work by cognitive psychologists (the Theory Theory) suggests a less encapsulated picture of concepts, one more akin to the interpretive account . These two theories will be discussed in turn.

On the Prototype Theory, concepts are supposed to be self-contained, relatively independent clusters of features, some of which are weighted more heavily than others. This internal structure is supposed to explain the fact that some instances of a concept are more quickly recognized and more liable to be named than others. The more prototypical instances are the ones that have more of the features, or more of the heavily weighted features, associated with the relevant concept. In some experimental tasks, for example, robins are more easily identified as birds than penguins. They are therefore ruled to be more prototypical instances of the concept bird than penguins. This is explained by saying that penguins only possess such lightly-weighted features as 'have wings' and 'have beaks', whereas robins also possess heavily-weighted features such as 'sings' and 'flies'. In matching features against items in the world, the cognizer reaches the "critical sum" for the concept bird more rapidly when presented with a robin than a penguin.

---

[11] The identification of lexical concepts with meanings is by no means unique to the interpretive approach. For example, Carey writes: "... I will use 'concept x' and 'meaning of the term "x"' interchangeably." She goes on to say that in previous work, "In every case that I found a difference in meaning of a term 'x' between the child's lexicon and the adult's, there was a corresponding difference in the concept x, as revealed by patterns of inductive projection, sorting tasks, and other tasks not requiring the use of the term." (1988, 167n) Similarly Gleitman, Armstrong, and Gleitman state: "... for present purposes we make no fine distinction between theories of word meaning and theories of concept structure." (1983, 88)

More recently, a number of cognitive psychologists have begun to argue that psychological concepts are more enmeshed in relevant theories and couched in explanatory beliefs. This shift has been driven by experimental tasks that are not explainable by treating concepts as collections of features, even probabilistic collections of features. Two main cognitive effects do not comport well with such models. The first is that the kinds of features that subjects associate with certain concepts varies widely and almost without limit when one varies the experimental context in which they are tested. Rather than accessing a fixed set of features in conjunction with each concept, experimenters have found that subjects access different chunks of a global theory. There is apparently no limit to the features that even a single subject associates with a certain concept depending on the context in question. A second difficulty for the Prototype Theory is the ability of subjects to make cross-conceptual links and to relate their beliefs involving different concepts in informative ways, which abilities are not easily explained on a model of concepts as bounded, self-contained feature lists. Categorization is not a simple matter of matching features among a concept and its instances, but is determined by inferential processes driven by surrounding explanatory theories.[12]

These experimental results tally better with a picture according to which concepts are embedded in a total framework of explanatory beliefs (or theories) that one draws upon in part in performing a particular cognitive task, with different parts of the entire corpus invoked in different tasks, even ones involving a single concept. Although a full-blown Theory Theory has yet to emerge, there is dissatisfaction with a view of concepts as self-contained psychological structures, relatively isolated from one another and from pertinent background beliefs. Many theorists now hold that concepts are embedded in

[12] Some of the seminal sources for this account of concepts are Murphy and Medin (1985), Keil (1986) and (1989), Carey (1986) and (1989), and references therein.

larger theoretical networks with a dense pattern of correlations linking one concept to another.[13]

Elsewhere, I have argued that these two theories of concepts, the Prototype Theory and the Theory Theory, are operating at different levels of description and dealing with different entities.[14] Typicality effects emerge most clearly under time pressure and in tasks involving routine categorization decisions and identification of instances, when subjects are not questioned as to the reasons behind their decisions. In the psychological experiments that support the Theory Theory, by contrast, subjects are typically presented with full-blown narratives or accounts of natural processes and then asked various questions about them. The data in these cases consist of what psychologists call "protocol analyses": verbatim transcripts of subjects' responses and their attempts to justify those responses under the scrutiny of an experimenter. Neither the categorization tasks nor the subsequent justifications are subject to time constraints, and the categorizations are seldom as routine as those that occur in the experiments just described.

Despite the fact that these theories are usually considered rivals, the Prototype Theory appears to view the mind from what Dennett has called the "design stance", while the Theory Theory adopts the "intentional stance" towards the mind. On the design stance, the organism is treated as a device or artifact that behaves as it is designed to behave under different circumstances. The Prototype Theory holds that concepts are constituted from a bundle of features, and it thinks of them as being manifested in the organism when those features are detected in the world. The organism is designed in such a way that whenever a certain number of those features is detected and a critical sum is attained, the corresponding concept is tokened. Moreover, the features may largely be perceptual ones. On the intentional stance, by contrast, the organism is regarded as an agent that has formed rational beliefs about the environment and reasons about the world in conformity with

---

[13] The recent doubts cast on the Prototype Theory render attempts by philosophers of science to use it to analyze scientific theorizing somewhat regressive. For an attempt of this kind, see Giere (1994).

[14] For details, see Khalidi (1995).

those beliefs. Given the conclusions of the Theory Theory, it is more natural to think of concepts not as concrete physical objects, but as theoretical posits that facilitate the ascription of beliefs.[15] Concepts, from this perspective, are simply components of fully-fledged beliefs that have been ascribed to subjects according to our usual interpretive practices. Though these psychologists do not say so, there is less temptation to reify concepts on this way of thinking about them. Therefore, I claim that there is an obvious convergence between the emerging Theory Theory of concepts and the intentional stance towards the mind, from which stance concepts are thought of as theoretical posits, or abstract entities rather than concrete objects.

6.7. Local Incommensurability: Children and Adults

The previous section argued that the Theory Theory of concepts in cognitive psychology shows some affinity to the interpretive account of concepts. A certain problem arises, however, since at least some psychologists have used the Theory Theory to argue that there is incommensurability between the concepts of adults and those of children. Carey (1985, 1988) has argued that the conceptual schemes of children and adults (or of children at different ages) are incommensurable with one another. She has followed Kuhn in holding that pairs of incommensurable theories are ones that contain clusters of interdefined terms that resist translation from one into the other. She has found some evidence for the claim in the theories attributed to young children, discovering in them whole clusters of concepts all of which resist one-word translations into the adult vocabulary. Carey has made this claim for the preschool child's concepts alive, dead, living thing, animal, plant, baby, and others.[16] Kuhn's idea that clusters of interdefined terms

---

[15] For the distinction between the design stance and intentional stance, see Dennett (1987, 16-17). For a closely related claim that psychologists sometimes adopt one stance and sometimes the other, see Flanagan (1984, 178-80).

[16] Strictly speaking, since the child's concepts are different from the adult's, it is misleading to talk about a single concept baby. Rather, such locutions should be understood as pertaining to two different concepts associated with a single word, 'baby'.

result in incommensurability was first encountered in section 1.5.; later, in section 3.4., some principled objections were made to this claim. Carey's work provides us with an opportunity to take another look at a purported instance of this phenomenon, albeit not one derived from science, but from childhood.

One of Carey's striking examples concerns the preschooler's concepts of animal and baby. One source of evidence for her claim that these concepts are not shared by adult and child is that four-year-olds don't usually realize that all animals have babies but think that only some do. As she explains it, children think of babies as small, helpless versions of bigger creatures who, because of their behavioral limitations, require bigger animals to take care of them. Carey describes the case of a typical four-year-old boy who confirms that dogs have baby dogs and that cows have baby cows, while strenuously denying that worms have baby worms. The boy's reasoning, according to Carey, is that worms are so behaviorally bankrupt that there is no way for the small ones to have a smaller behavioral repertoire than the big ones. The four-year-old insists that there are short worms, not baby worms. (1988, 167-8)

Carey admits that this and similar cases merely raise the possibility of local incommensurability, since it is still possible that the child holds different beliefs from the adult, beliefs that are formulated over the same conceptual base. In this case, for example, we might say that the child believes that worms don't have babies. However, she says that the only way to tell is to analyze the whole set of concepts and beliefs that underlie them, and that when one does so the possibility of local incommensurability is made more probable, though she stops short of endorsing it unequivocally. (1988, 174-175) Not only is the concept baby not shared, according to Carey, the related concepts animal, life, death, living thing, and body are also different for adult and child. Recall that it is the existence of an interrelated chain of concepts, none of which can be translated into our terms that renders a corpus of beliefs incommensurable with our own, according to Kuhn. For example, Carey claims that the preschool child's concept death is nonbiological. According to the child's understanding, the dead live on in altered circumstances and death is avoidable and reversible, as a special type of sleep. Since children do not make a distinction between dead and unreal, nonexistent, and inanimate, their single nondifferentiated concept dead does not correspond to any unitary adult concept. This

concept includes both not alive (as applied to a deceased grandfather) and inanimate (as applied to a table) and plays no role in the adult conceptual system. (1988, 176-7) To illustrate, Carey reports the following exchange with a 3-year-old child. "Isn't it funny," the child asks, "statues aren't alive but you can still see them?" Carey replies, "What's funny about that?" The child replies, "Grandpa's dead and you can't see him." (1988, 178) Similar things apply to their concept alive. When preschool children are deciding whether the sun is alive or not, they are not answering the question whether the sun is animate or inanimate because they cannot even entertain that question, not having differentiated inanimate from dead. Rather, they are deciding whether the sun is active, real, existent, present or whether it is dead, imaginary, nonexistent, or a mere representation. (1988, 177-8; cf. 1985, 25-26) Their concepts of life and death do not make any of these distinctions, which are made in the adult conceptual scheme. In the course of the emergence of an intuitive biological theory in the years before age 10, all of these concepts are "simultaneously adjusted", and none are identical with the adult concepts. (1988, 180; cf. 1985, 39-40)

Notice that Carey has just told us a fair amount about the preschooler's conceptual scheme using our very own adult terms. She explains that this exegesis of the children's concepts, baby, alive, dead, and so on, is what she calls, following Kuhn, a "translator's gloss". She holds that the child's beliefs cannot be expressed in the adult language without such a gloss. (1988, 180) What makes this a gloss on the translation rather than part of the translation itself is presumably the fact that it is expressed partly in meta-linguistic terms. The gloss involves taking a step back from the theory to point out that unitary concepts for children can be interpreted in terms of more than one of our concepts. We need to point out that the children's concept can be unpacked in terms of more than one of our concepts, a move that involves mentioning their concepts rather than using them, to use Quine's well-known distinction. It is as though we were to introduce their terms in quotation marks, rather than employ them directly.

However, there seem to be ways of reinterpreting Carey's evidence that enable us to avoid local incommensurability. For some concepts, we might say that the child's concept is the same as the adult's, but that the child has some false beliefs associated with that concept. For example, we can say that preschoolers share our concept baby, but add that

they believe that worms don't have babies and think that babies are behaviorally bankrupt versions of adults. In this case, we have used our own concept (baby) to convey their beliefs. We need not resort to a translator's gloss, as Carey does, to explain how their mental life differs from ours. This treatment is perhaps most plausible for the concepts baby and animal. For other concepts, we can say that the child's concept is equivalent to some concept of ours that is picked out by another word. This seems a reasonable course of action with the concept that the child associates with the term 'alive', which may correspond with the adult concept active. Thus, the child's beliefs about the sun being alive might be interpreted as being about the sun being active. For yet other concepts, we may be forced to conclude that the child's concept is not equivalent to any of ours and that we must neologize by coining a new term that would serve to stand in for the child's concept. This may be the aptest treatment for the concept associated with the word 'dead', which does not seem to correspond neatly to any adult concept. Now, the third option is perhaps the one taken to raise problems of incommensurability. It is surely not tantamount to incommensurability on its own, but when neologizing becomes rampant, there might seem to be some justification for allowing that there is a certain slippage between the two theories. An interpretation that neologizes for a whole cluster of closely-related concepts may give us some grounds for claiming incommensurability.

Can we ensure that this degree of neologizing does not occur? There does not seem to be a knock-down argument against this eventuality, but the above analysis already renders neologizing implausible for some of the other concepts Carey mentions, namely baby, animal, and alive. Even if we must neologize for some of the concepts she examines, for example dead, we need not resort to neologisms for the others. The reason that neologizing is more plausible in the last case is that an argument can be made that there is no saying whether the child uses 'dead' to mean dead (no longer living) or inanimate (never lived). Since neither of the two substitutions makes better sense of the children's beliefs, we cannot ascribe one rather than the other. This judgment relies on the Principle of Warranty expounded in section 5.5., which enjoins the interpreter not to ascribe a given concept unless there is sufficient warrant to do so. However, no parallel case can be made for the child's term 'baby', for example. It is not necessary to ascribe an entirely new concept simply because the extension of this term is different for adult and child. As

explained in section 6.2., concepts may be shared even though extensions do not coincide. Moreover, the fact that terms for which we require neologisms are closely related to other terms, does not imply that those other terms require neologisms too. If the child's term 'dead' has no equivalent in the adult lexicon, and if that term is closely related to the child's terms 'baby' and 'animal', it does not follow that the latter terms also have no equivalent in the adult lexicon. As I argued in section 3.4., the fact that terms for which we require neologisms are closely related to other terms, does not imply that those other terms require neologisms too. That is because there are generally no constant, unchanging definitions in the context of changing beliefs. Here, I am relying on the Principle of Undefinability, introduced in section 5.6., which rejects a definitional approach to fixing the meanings of scientific terms. Thus, a close association between these terms in the child's lexicon does not support the claim that, if one changes, then the rest inevitably follow suit.

In childhood, as in science, a new concept (e.g. dead) is usually introduced against the background of the old ones (e.g. baby and animal). This seems to be a typical characteristic of the evolution or development of belief systems, whether in childhood or in science: limited conceptual innovation is accompanied by considerable conceptual continuity. That is indeed what we would expect if we were to imagine the transition from one scheme to another from the point of view of the agents undergoing the transformation, and it seems to provide a more plausible explanation of Carey's evidence. In this section, I have proposed that at least some of the children concepts Carey discusses correspond to adult concepts, but that children associate some different beliefs with those concepts. By contrast, Carey speculates that there is no single sense of different concept to be defended and holds that there is a "continuum of degrees of conceptual differences", at the extreme end of which are concepts embedded in incommensurable conceptual systems. (1988, 168) The possibility of limited neologizing (for the children's term 'dead' but not for the others) is not adequately discussed by Carey and seems on reflection to be an attractive interpretation. If this interpretation is accepted, the threat of incommensurability (albeit of a local variety) will have been avoided once again. In proposing this interpretation, I have

helped myself to at least two of the intepretive principles discussed in Chapter 5: the Principle of Warranty and the Principle of Undefinability.[17]

6.8. Connectionist Concepts

Another recently influential approach to the study of the mind is the connectionist paradigm in artificial intelligence, which is inspired in obvious ways by the architecture of the biological brain (connectionist structures are also known as "neural networks"). Connectionist models of the mind assume that the processing of information takes place through the interactions of a large number of processing elements or units, each sending excitatory and inhibitory signals to other units. A network of units is given certain inputs and "trained" to give certain outputs, in such a way as to model a simple cognitive process. One of the interesting features of this kind of "parallel distributed processing" is the fact that information is distributed over the whole network rather than being localized in specific locales in the network. Though there is no clear consensus on the identification of conceptual structures within a connectionist system, the most prominent candidate is the pattern of activations across the units of the network. If each activation across a number of units, n, is thought of as a vector with n components, each activation in a network can be represented graphically as a point in the n-dimensional activation vector space. Accordingly, a pattern of activation can be represented as a partition in that space. Thus, for example, Churchland holds that concepts are partitions in the activation vector space of a connectionist network--though he does not always sufficiently distinguish concepts from theories, conceptual frameworks, or prototypes.[18]

---

[17] For further discussion and a response to another psychological attempt to find incommensurability among the theories of adults and children, see Khalidi (1998b). This sectioin draws heavily on that paper.

[18] Incidentally, it is curious that an eliminativist about mentalistic or folk psychological notions such as Churchland should be so concerned to identify counterparts of these notions in connectionist networks. For details, see Churchland (1989, 232-4). His view seems fairly standard in the connectionist literature. According to Bechtel and Abrahamsen: "In a distributed network, each concept is represented by a pattern of

In a contribution to a volume on the cognitivist approach to scientific theories, Churchland illustrates some features of the connectionist paradigm by using the example of a neural network that can be trained to distinguish sonar echoes of underwater mines from sonar echoes of submerged rocks. (1992, 343-353) But such cognitive functions are a far cry from those involved in formulating and manipulating full-blown scientific theories, a fact which renders questionable Churchland's claim that certain features of connectionist models of scientific theories vindicate some of Feyerabend's theses about science, including the thesis of incommensurability. Even if connectionist networks prove to be capable of performing higher cognitive tasks and can be used to model processes similar to the devising and evaluating of scientific theories, one need not adopt Churchland's attitude to these computational devices. Rather than identify certain aspects of connectionist networks with mentalistic entities like concepts, theories, and beliefs, the connectionist and mentalistic processes may be thought of as existing at a different level of description than the folk psychological one, as suggested for prototypes in section 6.6. That would make the interpretive account of concepts compatible with the connectionist approach to these issues, simply because it is not a competitor for the same turf. At least one proponent of connectionism has advocated a rapprochement between connectionists and folk psychologists. According to Andy Clark, those who pose what he calls the "syntactic challenge" hold that if mental states are real and cause behavior, there must be neat syntactic analogues in the head to the semantic expressions that appear in the sentences describing those mental states. But Clark denies this and presents the following alternative picture in its place:

> Instead, I see belief and desire talk to be a holistic net thrown across a body of the behavior of an embodied being acting in the world. The net makes sense of the behavior by giving beliefs and desires as causes of actions. But this in no way depends on there being computational brain operations targeted on syntactic items

---

activation across an ensemble or set of units; by design, no single unit can convey that concept on its own." (1991, 51) The locus classicus for the connectionist research program is Rumelhart and McClelland (1986), upon which this paragraph draws.

> having the semantics of the words used in the sentences ascribing the beliefs. (1989, 5)

Because they operate at different levels of description, proponents of the two theories might find a modus vivendi. An interpretivist account of concepts may even find some comfort in the connectionist research program since it relieves the pressure to find physical manifestations of mentalistic entities and shows how different levels of the mind-brain could be organized very differently.

There is yet a third connectionist attitude to the mental, and one that has been specifically applied to the analysis of scientific theories. The kind of connectionism described by Churchland and Clark is a distributed connectionism according to which individual nodes in the connectionist network (the so-called "hidden units") do not have a simple mentalistic interpretation. They do not, for instance, correspond to individual concepts or beliefs. But Paul Thagard has used the resources of local connectionism to model scientific theories. Thagard writes that he will "treat concepts and propositions as mental representations, with concepts corresponding to predicates and propositions corresponding to sentences." (1992, 21) He goes on to develop a view of scientific theories that relies crucially on a linguistic mode of representing theories and in which individual nodes in the connectionist network correspond to individual tenets of a theory or statements of evidence for that theory. To illustrate this approach, consider a portion of one of Thagard's case studies, Darwin's theory of evolution. Here is his restatement of some of Darwin's evidence (1992, 143):

E1 The fossil record contains few transitional forms.

E2 Animals have complex organs.

E3 Animals have instincts.

And here are the restatements of three of Darwin's main hypotheses (DH1-3) followed by the main creationist hypothesis (CH1):

DH1 Organic beings are in a struggle for existence.

DH2 Organic beings undergo natural selection.

DH3 Species of organic beings have evolved.

CH1 Species were separately created by God.

To model theory change, Thagard has developed a program that encodes inhibitory and excitatory relations between the individual pieces of evidence (E's) and the hypotheses (DH's and CH's), in such a way as to show why, given the evidence, Darwin's hypotheses are more strongly supported than the Creationist hypothesis. The level of activation of a node (e.g. DH3) after the network has settled corresponds to its level of acceptability. Thus, if the activation level of DH3 is higher than that of CH1, a tenet of Darwin's theory is favored over a tenet of the Creationist theory on the basis of the available evidence.

The use of weighted links between nodes and of variable activation levels is characteristic of the connectionist paradigm, but the crucial difference between this localist version of connectionism and standard distributed connectionism is that the nodes in Thagard's scheme are sententially expressed "propositions". The local connectionist scheme vitiates one of the main attractions of connectionism, namely its neurological plausibility (since it is highly unlikely that individual neurons will be interpretable propositionally in this simple fashion), but more important, it begs the crucial question about conceptual change. That is because Thagard must decide independently which propositions he attributes to each scientific theory. And he does not say how he decides to attribute these theoretical tenets, particularly how he judges that certain crucial terms mean the same in two competing theories. In the above example, he does not tell us how he has decided that the Darwinian term 'species' means the same as the creationist term 'species'. Even if this is a plausible assignment in this particular case, such assignments must be justified in general. Before he can even run his program, Thagard must already have made the difficult interpretive decisions. Therefore, his theoretical framework must presuppose a way of resolving the thorny questions about sameness and difference of meaning.

Despite the compatibilities that I have mentioned, at least some of the work in cognitive science conceives of concepts as independent, relatively isolated things, even as concrete objects with well-defined physical properties, attitudes that do not sit well with the interpretive approach. It would not be surprising if the study of mental processes eventually resolved itself into two or more levels of description. Meanwhile, my claim is a modest one: much of the recent work in cognitive science can be reconciled with the account of concepts that emerges from the interpretive approach, according to which

concepts are not thought of as concrete physical objects, but rather as theoretical posits ascribed on the basis of an overall interpretation of a rational agent. A reconciliation is possible with the Theory Theory of concepts in cognitive psychology, despite the claims of conceptual incommensurability made by some of its advocates. When it comes to the Prototype Theory or the connectionist view of concepts, they can be reconciled simply by saying that they aim to isolate entities at different levels of description than the interpretive approach. The fact that the way I have been using the term 'concept' conforms to aspects of commonplace usage and is compatible with much cognitivist research, lends some justification to treating concepts as meanings in the sense being advocated here. Ultimately, there may be room for pluralism, though there may be a squabble over the term 'concept' itself.

6.9. Change of the Conceptual Repertoire

This chapter has tried to justify further the claim that one compares scientific theories by matching up their concepts or the meanings of their terms. The meaning of a term from another theory is given by its translation in our theory and whenever there is such a translation the two theories are said to share a concept or meaning. A concept marks a certain semantic feature common to all the beliefs in which it occurs. For example, our concept mass drops out of all the (potentially infinite) beliefs we have about mass. On the interpretive approach, concepts constitute an indispensable finite basis for characterizing a potential infinitude of mental states; they are theoretical posits and abstract entities rather than concrete objects.

In this chapter, I have also tried to effect a partial reconciliation between some recent cognitivist views of concepts and the view that emerges from the interpretive approach. Still, there are implications of my view that are likely to offend some of the (expert or folk) psychologist. A prominent one, which has not hitherto been given due weight is the claim that the ascription of concepts is an all-or-nothing affair. In other words, a concept is either shared among two thinkers (or a single thinker at two different times) or that it is not, and there is no halfway house. This is bound to jar with the usage of some psychologists, philosophers, historians, and others who are prone to say, in some cases, that two people only partially share a concept, or that their concepts overlap only

partly. By the same token, my view also rules out talk of concepts themselves 'changing', 'evolving', 'developing', and so on--at least if that is construed as a transformation of the concept itself or an alteration in its very content, which nevertheless somehow preserves its identity.

There are two points that need to be argued here. First, that these are indeed inevitable consequences of my account of concepts, and second, that these consequences are ones that we can live with, that is, that we can justify parting company with expert and common parlance for the sake of philosophical cogency. As to the first point, recall that on the interpretive approach, our informant's term either gets translated by one of our terms or it does not, and there is no other alternative. It is a central assumption of this work that when a term gets so translated we rule that a concept is shared and when it does not we conclude that the associated concept is not common to the two thinkers or theories. Every interpretive effort involves a series of decisions to translate an alien term by a home term, and such decisions issue in determinate judgments as to whether a concept is shared or not. This is what gives concepts their fixity. Clearly, every such decision is a bivalent one since an alien term either gets matched up with one of our terms or else it does not. In the first case, the concept is (wholly) shared and in the second it is (wholly) unshared. On the interpretive account, the translation of terms is the only purchase we have on the ascription of concepts, and for that matter, on the individuation of concepts. Thus, the all-or-nothing character of concept sharing is a genuine consequence of the interpretive account of concepts.

When it comes to the second point, that this consequence of the interpretive approach is one we can live with, I will advance two considerations which should be convincing enough not only to make the position palatable, but rather to render it downright appetizing. The first thing to notice is that insisting on the complete coincidence or non-coincidence of concepts allows concepts to assume their proper role in reasoning and inference. When we judge that a particular concept is shared among two scientists, that gives us grounds for framing an agreement or disagreement between them. Once we rule that Thomson and Bohr shared the concept <u>electron</u>, we can go on to declare that they agreed that electrons had negative charge, but disagreed as to whether electrons were particles. In other words, we can isolate the shared beliefs and contrast them with the

unshared beliefs. But if we had ruled instead that there was only partial overlap between Thomson's concept and Bohr's, it is no longer clear what we could have said about their areas of agreement and disagreement. Perhaps we should say that they only partially agreed that electrons have negative charge and partially disagreed that they were particles (since their concepts only partially coincided). Then there would only have been partial consistency on the first point and and partial contradiction among their theories as to the second point. But since we do not have a logic of partial contradiction, it becomes impossible to compare their theories directly for consistency on particular points. Therefore, partial conceptual overlap wreaks havoc with our ability to ascertain the inferential relations among theories. And it is easily avoidable by talking instead about partial overlap in theory (instead of concepts), which enables us to pinpoint the area of overlap exactly by singling out the shared beliefs and segregating them from the unshared ones. The interpretive approach foregoes talk of partial conceptual overlap in favor of partial overlap of theories or sets of beliefs, where this simply means agreement on some particular tenets and disagreement on others.

A second reason for eschewing talk of conceptual overlap pertains to the very intelligibility of such a locution. In discussing the possibility, in the previous paragraph, of finding partial overlap among Thomson's concept and Bohr's concept, I explicitly avoided saying that they might have partially shared the concept electron. For that would imply that the content of the concept is at once the same and different among the two thinkers, which is incoherent. To say that it is the concept electron in both cases, but that there is only partial coincidence among the two concepts, is clearly not an option. What individuates a concept is its content and we are assuming that it has the self-same content when we identify it as the concept electron. However, we are implying that the content is different again when we say that there is partial overlap, and it seems clear that we cannot have it both ways.[19] This point has been made, albeit less explicitly, elsewhere in this work,

---

[19] Closely related points have been made by Andrew Woodfield. In a series of papers, Woodfield has distinguished between concepts, which should not be thought to exhibit internal change, and conceptions, which can be viewed as changing, developing, and so on. See Woodfield (1991), (1993), and (1996).

notably in section 5.2. with reference to MacIntyre's discussion of the medieval concept man, and again in section 6.7., in the context of Carey's discussion of the child's concept baby.

Much of what is said by psychologists and others about conceptual overlap can be rendered compatible with the view being urged here if one reinterprets it as talk of theoretical difference or difference of belief. Furthermore, to say that there is no conceptual overlap or partial conceptual coincidence among agents is not to deny that there may be some vague or indeterminate cases, typically ones in which the interpretee has few if any beliefs associated with a term. But such cases are being ignored for these purposes, which concern the comparison of full-blown scientific theories with substantive and well articulated tenets, as argued in the Introduction. A good deal of the psychological research on concepts focuses on cases in which the beliefs in question are minimal and inexplicitly articulated, such as the conceptual systems of children, rather than the precisely and exhaustively articulated theoretical systems that are the main interest of this work. Because of this discrepancy, at least one psychologist working in this field has registered reservations about treating children's unsystematic responses to questions as bona fide theories of a certain domain. Keil has indicated discomfort with children's 'theories', which only emerge in response to questions posed by an adult experimenter. (1989, 48) Notwithstanding some points of contact articulated in section 6.7., this is one of the reasons why the cognitive development of children may have to be studied somewhat differently than theoretical change in science.

Is this, then, a book on conceptual change which prohibits talk of conceptual change? No, the claim that concept sharing is an all-or-nothing affair does not invalidate talking about conceptual change; it recasts it as change of concepts rather than change in concepts. Conceptual change or evolution can be thought of as change or evolution of the whole conceptual repertoire, which involves either the introduction of new concepts or the elimination of old concepts (or both). Concepts themselves do not change, but the whole palette of concepts does by virtue of conceptual addition or subtraction. Moreover, even when no concepts have been added or discarded, an indefinite number of theoretical

changes can take place, that is, an indefinite number of changes in beliefs or theoretical tenets.

Chapter 7: Realism

> Judging by what we can observe, nature is not a mere series of episodes, like a bad tragedy.
>
> Aristotle, Metaphysics 1090b19

> [W]e see a complicated network of similarities overlapping and criss-crossing: sometimes overall similarities, sometimes similarities of detail.
>
> Ludwig Wittgenstein, Philosophical Investigations §66

7.1. Two Challenges

The incommensurability thesis has been taken to have significance for the attitudes of rationality and realism towards science. The rationality of science is an important concern, but it will not be specifically addressed in this book. If successive theories cannot be directly compared, then it is not clear how the choice between them can be based on purely rational considerations. But if they can be, it obviously does not follow that rationality is guaranteed. Once a method has been proposed for comparing theories, there is a further question as to whether rational principles can be worked out governing the choice between scientific theories. There is also a question as to whether and to what extent scientific practice can and does live up to those standards. These important questions will not be discussed here.

Although the tension between incommensurability and realism seems more apparent and is often taken for granted[1], it is perhaps in need of further comment. Whether local or global, incommensurability alleges a substantial amount of ontological replacement at each stage of scientific change. It may be possible for an advocate of incommensurability to maintain that scientific theorizing is gradually converging on the right set of entities and that the real entities will emerge at the end of inquiry. The holder

---

[1] For example, Putnam asserts without further ado that "the principle that reference can be preserved across theory change... seems to me to be central to any realist philosophy of science..." (1979, 284)

of such a position might contend that although theories cannot be put in the same terms and compared directly, there are still other ways of comparing them (as we saw in sections 1.4. and 1.5., both Feyerabend and Kuhn hint that there are such ways, without spelling them out in full). If so, there may be some scope for saying that successive theories are better at identifying the furniture of the universe. If we grant for the sake of argument that some other way of comparing scientific theories can be found, then it is possible for this position to escape anti-realism. Hence, the answer to the question whether the incommensurability theorist is committed to anti-realism awaits an answer to the question of whether an alternative means of comparison can be found. If a means of comparing theories can be devised which can rule one theory to be a better description of reality than another, then there may be a way of reconciling incommensurability with realism. But it has been a central tenet of this book that the most natural means of comparing theories is the linguistic one and that it is the mode of comparison that enables us to pinpoint the specific agreements and disagreements among scientific theories (rather than, say, greater simplicity or overall aesthetic superiority). It is safe to conclude, at least for the time being, that incommensurability about theories is incompatible with scientific realism.

The denial of incommensurability may be necessary to vindicate realism (bearing in mind the qualification of the above paragraph), but it is certainly not sufficient. Even if a method can be outlined for the direct comparison of scientific theories, it might still be said that the conception of scientific theories inherent in it is not a realist one. In this chapter, I will focus on the charge that the particular method being articulated here, the interpretive approach, fails to deliver some aspect of realism about science. Two objections will be considered, corresponding to two counts on which this approach may be accused of anti-realism. The first begins by drawing attention to the fact that I have claimed that scientific knowledge can yield a number of "crosscutting taxonomies" of the same subject matter. That position was mentioned in the course of my criticisms of the causal theory of reference in Chapter 2, since I accused it of presupposing a contrary claim, that there is a bedrock of non-overlapping, basic natural kind categories. I argued in section 2.5. that one of the main obstacles facing an attempt to deploy the causal theory of reference to account for the reference of scientific terms is, roughly that any attempt to use an initial baptism to ground the reference of a term makes a false presupposition about scientific taxonomy. It

assumes that the exemplar proffered at the baptism exemplifies a single scientific kind. Since any exemplar will typically be a representative of a multiplicity of crosscutting kinds, the brute causal relation cannot be used to ground the reference of scientific terms in a metaphysical realist fashion. But if categories drawn from different theories cut across one another, the possibility arises of two scientific theories which classify the same phenomena (in a sense to be specified) in very different ways. Indeed, it may be said that two such theories are none other than incommensurable rivals.

The second challenge to my position would have it that the denial of the metaphysical realist view of reference with its strong anchors to the world threatens to leave scientific theories radically adrift. If, as I have argued, meaning and theory are in the same boat, that boat may be Neurath's notorious vessel, which never makes contact with terra firma. After a number of successive theory changes, each theory may be interpreted in terms of its successor in such a way that we might be floating further and further from our starting point. That is because meaning assignments are given by a translation function constructed between theories, rather than a relation between theories and the world. This threatens to leave us with something like an idealist picture of scientific theories, according to which each scientific theory can be said to be about its predecessor rather than about the world of phenomena, much as literary texts are nowadays said to be about other texts rather than the world itself. To make matters worse, we seem to have no way of noticing any putative conceptual drift from within our theories themselves. These two challenges will be examined in turn in this chapter.

7.2. Crosscutting Taxonomies

The claim that scientific categories can cut across one another, which was mentioned in Chapter 2, is one that I have defended in more detail elsewhere.[2] I will recapitulate that argument here in order to try to determine its relevance to the question of incommensurability. A number of philosophers have claimed that natural kinds are arranged in a hierarchy, such that higher categories in the taxonomic system do not

[2] See Khalidi (1993a) and (1998a), for details and references to other relevant literature.

trespass on the boundaries between the categories at the lower levels. In other words, two kinds can only overlap if one of those kinds is wholly subsumed under the other. So an individual can belong to two or more kinds, only if they can all be put in subsumption relations with each other. For example, if humans are classified together with gorillas as primates, and gorillas are classified with cows as mammals, then humans and cows should also be classified together under one of those two categories (intuitively, whichever one is higher). In this case, they are classified together as mammals. The natural kind categories should form a nested hierarchy of categories that are disjoint, or do not crosscut.

This claim is false, at least if it is applied to scientific categories in general.[3] There are many examples from science of bona fide categories that cut across another set of scientific categories that serve to classify the same type of entity. To take a simple example, a tiger is a member of two categories, mammal and carnivore[4], which crosscut one another. The two kinds mammal and carnivore are not disjoint, for there is overlap between mammals and carnivores (the class of tigers being one member of the overlap) and yet neither category is subsumed under the other (since some non-mammalian birds are also carnivores and some non-carnivorous herbivores are also mammals). To illustrate, if cows are classified together with tigers as mammals, and tigers are classified together with hawks as carnivores, neither of these categories, mammal or carnivore, include both cows and hawks. If one accepts the standard Linnaean taxonomic system with its nested hierarchy of natural kinds going all the way from species and genus at the lower end to

---

[3] I enter this qualification because someone might say that the thesis of disjointness was not meant to apply to scientific categories but only to "natural kinds", which are different entities. While it is doubtful that the thesis holds even there, and there is no agreement on just which categories are the natural kind ones, I will not discuss this issue here, since the main concern is with scientific categories in general.

[4] By 'carnivore', I intend the category of meat-eating animals, rather than the phylogenetic family Carnivora. As Simpson points out, some Carnivora are strictly herbivorous. (1961, 33)

phylum and kingdom at the higher, one seems committed, on the hierarchical view, to dismissing the category carnivore as a non-scientific one.

While the category carnivore may be dismissed as being not truly scientific, there are many other examples of categories that cut across the Linnaean system of categories and are quite respectable scientific categories in their own right. The category parasite behaves in this way with respect to the category insect (the phylogenetic class Insecta): both tapeworms and fleas are parasites and both fleas and flies are insects, but tapeworms and flies are neither both parasites nor both insects. It should be pointed out that there are even some categories that cut across the most basic categories in the Linnaean system, the species taxa, so that the species taxa cannot be considered a bedrock of basic categories either. For instance, in entomology, organisms belonging to different insect species can belong to the categories, larva, pupa, and imago, all of which crosscut species taxa. Similar remarks may be made for categories in physics and chemistry. The difference between the crosscutting view of scientific taxonomy and the hierarchical view can be illustrated by way of two simple diagrams (see figure 7.1.).

The introduction of this claim can be used to issue in the following challenge: If the categories from some scientific theories are related in the way that I have outlined, how does one distinguish these sorts of crosscutting theories from incommensurable theories? This is not just a hypothetical challenge, for it owes something to an argument by Hacking who draws on some of Kuhn's unpublished writings[5], in which he uses a similar idea to recast the claim of incommensurability. Hacking has put forward a definition of a scientific 'taxonomy', which is his term for a hierarchy of non-overlapping, non-subdividing categories that culminate in a set of basic categories. The taxonomic thesis is tantamount to the idea that scientific categories are arranged in a hierarchy, such that higher categories in the taxonomic system do not trespass on the boundaries between the categories at the

---

[5] Hacking cautions that some of these writings are marked: "Draft: not for distribution, quotation or paraphrase," and suggests that these may not be Kuhn's considered opinions. Be that as it may, it is not particularly important that these views should be Kuhn's or even Hacking's. I am raising them to see if they enable us to construct a plausible defense of incommensurability and to see if they create problems for my own view.

lower levels. In addition, taxonomies terminate in a bedrock of basic categories that cannot be further subdivided. In Hacking's terminology, scientific categories cannot overlap; in my terminology, scientific categories cannot crosscut. I prefer to use the terminology of crosscutting rather than overlapping, since superordinate categories that wholly include others might be said to overlap with their subordinate categories but not to crosscut them.[6]

Hacking conjectures that scientific categories belong to taxonomic hierarchies in order to explicate Kuhn's claim that one scientific theory can structure the world differently from another theory, the claim of "conceptual disparity" among theories first encountered in section 1.5. and revisited in section 3.4. According to him, whenever we have two scientific taxonomies, their categories either overlap one another, or the categories of one taxonomy subdivide the lowest categories of the other, or else their categories coincide. In the first two cases, overlapping and subdividing, Hacking uses the taxonomic thesis to rule out the possibility of translation. In the case of overlapping

---

[6] Hacking mentions Fred Sommers as having proposed a theory of predicates in natural language that is formally analogous to this one (see Sommers (1963)). However, the crucial difference with Sommers' theory of types is that it does not consider sets of things to which a predicate applies or fails to apply, but those to which it would or would not be significantly predicable. A predicate is said to be significantly predicable of something if and only if it would not be a category mistake to apply it. The predicate 'hard' is significantly predicable of chairs and blankets (it spans the set of chairs and blankets), though it may only be true of members of the former set. This issues in a hierarchy of types, in which all apparent cases of crosscutting show in fact that a term is ambiguous. Thus, chairs and questions are of different types, and although 'hard' seems to span both, it is actually ambiguous as applied to members of the two sets. Sommers' claim would seem to be weaker than Hacking's because it relies on the notion of a category mistake rather than mere lack of applicability: the latter implies the former and the denial of the former implies the denial of the latter. Although it has been regarded by some cognitive psychologists as a constraint on language-learning (for example Keil (1979)), Sommers' claim has also met with criticism in those quarters. Carey (1986) has come up with purported counterexamples to Sommers' claim, but I will not assess her conclusions, since for my purposes it is enough that Hacking's weaker claim is false as applied specifically to scientific categories.

categories, there cannot be a translation between new and old, since they belong to different taxonomies. In the case of subdivision, the kind in the old science is a category with no scientific subkinds, "so the old name cannot be translated into any expression in the new science that denotes a scientific kind." (1993, 295) Only in the third case, when categories coincide, can there be translatability. In short, he uses the idea that categories within a particular taxonomy cannot crosscut one another to explain why categories from different taxonomies can be incommensurable.

There is an obvious problem with this reconstruction of incommensurability that Hacking clearly recognizes, namely that scientific kinds are not all taxonomic in the requisite sense. Just as I argued above that scientific categories crosscut one another, Hacking admits that there can be scientific kinds that "overlap" or "subdivide" each other. As an example, he gives the category poison, which overlaps the categories vegetable and mineral, but is surely not incommensurable with them.[7] But while he admits that poison is a legitimate scientific category (after all, the entire scientific field of toxicology is based upon it), he argues that it is not a "real Kind" (in John Stuart Mill's sense). The notion of "real Kind" is introduced in order to come up with a modified version of the Kuhnian claim, thereby saving the revised formulation of incommensurability. To qualify as a real Kind, there must be an inexhaustible number of things to find out about a category. While this is true of arsenic in Hacking's opinion, it is not true of poison, since "There is nothing much common to poisons except what puts them in the class in the first place, namely the potential for killing people after being ingested." (1993, 300) By contrast, he quotes Mill as saying that when it comes to real Kinds, we discover "new properties which were by no means implied in those we previously knew." (1993, 301) Hacking holds that the distinction between real and non-real Kinds has some application in science and does not think that it should be rejected merely on the grounds that it appears to carry a commitment to the analytic-synthetic distinction. One can now say why some real Kind

---

[7] Incidentally, it is doubtful that 'vegetable' and 'mineral' are bona fide scientific kinds, but I will grant Hacking this claim for the sake of argument. At any rate, one could replace it with the claim that poison crosscuts the categories organic and inorganic.

categories from different theories are incommensurable: they belong to different taxonomic trees. To put it differently, incommensurable theories are those that carve the world into real Kinds that "overlap" or "subdivide" one another.

Does the notion of a "real Kind" rescue the Kuhn-Hacking claim? Hacking admits that he does not have a proof that all real Kinds are taxonomic in the requisite way, and he is not sure that a proof should be sought. Still, he thinks that the notion of a real Kind is a useful one and that taxonomies "still carry some cachet" in the sciences. (1993, 303) However, it is neither clear that the notion of a real Kind is a useful one, nor that all exceptions to taxonomic trees are non-real. Thus, to cite one of the examples I adduced above to illustrate the claim of crosscutting categories, the category parasite seems to be a real Kind in Hacking's sense, although it trespasses on the taxonomic Linnaean tree. That is, there are certain things that have been discovered to be common to parasites that were not built into the category in the first place. To quote some typical findings from a standard textbook on parasitology: "Appropriate triggering mechanisms initiate the change from infective stages to parasitic stages. Once the parasite has begun its existence in a new host body, other triggering mechanisms initiate each change of the parasite during its development."[8] Such information was not part of what was initially put into the concept of parasite. In fact, there do not seem to be too many scientific categories that have this property of having nothing more discovered of them than what was put in, and Hacking is right to be concerned that his claims will bring on a waving of the "denunciatory placard 'Analytic/synthetic'!" (1993, 302) A category that had such a property would seem to be one tied to an irrevocable definition, of the kind that does not survive in the context of inquiry. Moreover, it is quite clear that this cannot be regarded as a proper case of incommensurability since these two taxonomies coexist comfortably, whereas incommensurable theories are generally regarded as rivals.

Hacking has neither made a strong case for the existence of what he calls (following Mill) "real Kinds" as opposed to non-real Kinds, nor has he made a case for the claim that all such real Kinds are also taxonomic ones. Therefore, one cannot say that incommensurable

---

[8] For more details, see Khalidi (1993a, 105).

theories are those involving overlapping taxonomies that are not inter-translatable. This helps to show that the existence of crosscutting kinds in science is more of a problem for someone claiming incommensurability than someone denying it. Notice that Hacking is quite willing to admit that there are quite respectable categories deployed by scientists such as poison, which overlap the real Kind categories. Yet he does not seem to consider such categories to be incommensurable with categories such as vegetable and mineral. Since they can coexist alongside them, the question is: Why can some such categories cohabit our global theory with the others, whereas other such categories are incommensurable with them? He might say that these categories are incommensurable too, but then he would need to explain why some incommensurable theories are considered to be rivals and some are not. On my view, all schemes that are crosscutting in this way should be capable of coexisting in our total theory of the world[9], whereas on Hacking's approach, some will and others will not coexist, and he does not say which are which. I would agree with Hacking that crosscutting taxonomies cannot be translated into one another, but nor should we expect them to be--no more than we should expect genetics to be translatable into cosmology.

But surely, it may be protested, there is a difference in the two cases: the relationship of genetics to cosmology is not the same as that of parasitology to zoology. There is still something of a problem in determining whether two theories are in competition with one another and can be translated into one another, or whether they merely coexist comfortably. Mere failure of translatability cannot be used as a demonstration that two theories or classification schemes can coexist side by side in our global theory. I need to explain why there is no temptation in the case of parasitology and zoology to say that the two theories are incommensurable, despite the fact that they pick out some of the same entities. It seems at first sight that these two theories have the same subject matter whereas genetics and cosmology do not, since parasitology and zoology

[9] To say that they are capable of coexisting is not to say that they do. Some crosscutting categories may be rejected for the same reasons that scientific categories are generally rejected.

concentrate on living organisms and individuate them in roughly the same way. Why, therefore, are they not rivals? Their ability to coexist can be explained by saying that they pertain to different interests. The idea that there are crosscutting taxonomies seems to be closely related to the view that scientific classification is interest-relative. If classification is always relative to certain interests, we would expect successive taxonomies to organize the world in different ways without displacing their precursors. I am not sure how to individuate interests or how to specify exactly their role in grounding crosscutting classification schemes, but this is what makes some of these schemes capable of cohabitation, unlike say the phlogiston theory and the oxygen theory in chemistry, which had roughly the same interests. By contrast, theories that treat the same phenomena relative to different interests can coexist comfortably in a single scientific account of the world; they are not incommensurable rivals.

Therefore, I need not rest with the brute fact of failure of translation to explain why certain theories do not come into conflict even when they appear to have same subject matter. This discussion provides me with a way of vindicating a claim made in section 3.2., where I allowed that a range of new concepts may be introduced with a subject-altering scientific change, as occurs with the introduction of a new scientific discipline or sub-discipline, but denied that this would be a case of incommensurability. That is justified by the fact that a true change of subject is distinguished from a change of theory about the same subject by the existence of different interests which guide the inquiry.[10] It is not enough to point to a set of entities or phenomena in order to specify the subject matter of a particular discipline or sub-discipline. The "domain" of a theory (as I have dubbed it elsewhere) is picked out partly by the specification of certain interests relative to which one undertakes the inquiry. For example, pharmacology and toxicology investigate some of the same chemical compounds and biochemical processes, but while the interest of the first is the use of chemicals to make humans healthier, the interest of the second is to determine

---

[10] For further details, see Khalidi (1998a). My views on this issue resemble those of William Wimsatt, who talks about theories having different "perspectives". Wimsatt (1994) develops similar ideas in much greater detail and with a wealth of evidence, while allowing that important problems remain concerning the individuation of perspectives.

the properties of chemicals that have a tendency to harm humans. Different interests may generate more than one classification scheme, sometimes within the same sub-discipline. Within toxicology itself, poisons may be classified in terms of the target organ they affect, in terms of the chemical mechanism they exploit, in terms of their poisoning potential, in terms of their route of absorption into the body, and so on. None of these examples constitute cases of incommensurable sets of theories.

This claim of interest-relativity also helps to explain why commonsense and scientific classification schemes often organize the same set of entities in a crosscutting fashion without being rivals, as I suggested in section 6.4. Since the folk often have different interests from the experts, their classificatory schemes tend to crosscut. Conversely, when they do not, they tend to coincide with one another, or else the folk classification tends to be ousted by the expert one.

7.3. Explanatory Efficacy

The second objection to my view which accuses it of anti-realism is related to an issue aired in section 6.6., in discussing the failure of transitivity when ascribing concepts. The failure of transitivity may leave the impression that, in the absence of a metaphysical realist view of reference, scientific theories will be insufficiently fastened to the world. It suggests that there is no single common subject matter for a succession of scientific theories, only a kind of invariance between pairs of successive theories. That being the case, one might well wonder what prevents scientific concepts from roaming all over the map after successive theory changes. In this section, I will investigate what ultimately tethers scientific theories, so that we do not end up with free-range theories being compared only to other theories.

The failure of transitivity implies that we cannot say that a series of agents with different theories all have the same relation to a single substance or property in the world. But this way of putting things need not be central to realism. We are generally only interested in saying whether any given theory succeeds in picking out those things that are real according to our current theory. The important thing is to be able to compare two theories and tell which concepts they share and where they agree. But, the objector may persist, the worry is deeper: on the interpretive approach, to say that Newton's term 'mass'

should be interpreted as rest mass, or that Priestley's term 'dephlogisticated air' should be interpreted as oxygen, merely marks a decision to interpret these theories in a certain way. It does not amount to saying, for example, that there are real physical quantities "out there", rest mass and oxygen, which are the common subject matter of two successive theories. In saying that a theory has the concept of rest mass or of oxygen, a realist about science surely intends to convey that some relation obtains between the holders of that theory and some real magnitude or substance in the world.

It cannot be claimed that the interpretive approach does not credit a concept to a theory partly on the basis of a relation that is thought to exist between the holders of that theory and the world, for it clearly does. Priestley was credited with the concept oxygen on the basis, inter alia, of his ability to isolate the gas in the laboratory, to identify many of its properties, and to specify its role in reactions with other substances. As I argued in section 6.2., there are various ways of ascertaining the extension of a scientist's term, though none of these are theory-independent. When a concept is ascribed to a scientist, this helps determine the extension of that scientist's term, and the extension of a term helps, in turn, in the ascription of a concept. There are a variety of relations that obtain between, say, Priestley and the extension of the concept oxygen; and it was partly on the basis of these relations that the concept was ascribed to him. On the interpretive approach, there is no single relation that must obtain between Priestley and oxygen that would lead us to translate one of his terms with our term 'oxygen'. A concept is not ascribed to an agent when that agent has a brute physical relation or a causal-intentional relation to something in the environment, as in metaphysical realist theories of meaning or reference. But that is not the only way to make sure that terms are anchored to the world. There can be a variety of robust but complex and indirect relations to the environment that can be used as evidence for crediting a concept. That is how we are able, for example, to credit Bohr with the concept hafnium and Dirac the concept positron, despite the absence of causal interaction (see section 2.5.). Although these scientists did not have direct causal contact with these entities in particular, they were obviously drawing in part on empirical data in predicting their existence.

Moreover, unlike orthodox descriptional theories of reference, the interpretive approach does not imply that the same set of properties is associated with a common

referent in two theories. If T1 and T2 share the concept 'electron', and T2 and T3 share the concept 'electron', it does not follow that there are some core properties of electrons that all three theories claim electrons have. Denying the existence of a single definition or a common core of definitional sentences which attach to a scientific concept across successive theory changes was consecrated in the Principle of Undefinability in section 5.6. This view is now widely accepted among philosophers and historians of science, some of whom have identified a kind of non-definitional or non-criterial continuity as being characteristic of successive changes of theory in science. Shapere has explicated the idea of what he calls a "chain-of-reasoning connection" across multiple theory changes[11]:

> [A]ccording to the views I have been presenting, it is in principle possible that every aspect of an idea (e.g., every property attributed to electrons) might be rejected and replaced, for good reasons. But as long as there are such reasons, an understandable relationship holds between the two uses--a 'chain-of-reasoning connection'. (1984, xxxviii)

It is indeed conceivable that all the beliefs once associated with a given term can be given up in the course of the history of science, but they surely cannot all be renounced at once. Shapere would seem to agree. As he says in reference to a case study he has carried out concerning the concept of 'observation'[12]: "[D]espite the initial oddity of the use of the term 'observation' in a certain sophisticated scientific context, a coherent interpretation of that usage can be given according to which it constitutes a reasoned extension of and departure from ordinary uses..." (1984, xxxvii) Or, as I would say, it constitutes a change in the theory

---

[11] Shapere's views have gained acceptance and have been employed by Nancy Nersessian in her (1984), where she applies them to a case study of the concept of field from Faraday to Einstein.

[12] The concept may be considered meta-scientific rather scientific proper, but Shapere shows how it is implicated in certain scientific theories, much as any ground-level concept would be. See his (1982b), in which he analyzes how scientists study neutrinos emanating from the sun to "observe" the hot solar core.

pertaining to the concept observation, rather than a change in the concept itself--otherwise, one should have chosen a different term in reporting the relevant beliefs.

Shapere's views are a welcome contribution from the point of view of the interpretive approach, since he takes a non-criterial line towards conceptual continuity and avoids a metaphysical realist view of reference, but he can be faulted for refraining from saying very much about what constitutes the "reasoned extension" or "chain-of-reasoning connection" at each stage of theory change. He does not say what kinds of considerations lead one to judge that two theories share certain concepts and sometimes suggests that reasons could be given for any change in the beliefs in which a certain concept features (i.e. for any theory change without a meaning change). This encourages the objection that the term 'electron' could come to refer to protons, or indeed that science itself could turn into football.[13] Leplin objects, against Shapere, that "chain-of-reasoning connections link virtually everything that happens in science with everything else that happens." (1988, 509) However, I believe that such a protest does not apply to the interpretive approach with its constraints on interpretation and the condition that some beliefs must always be shared for conceptual stability across any single theory change. Moreover, their explanatory efficacy is what guarantees a connection between our concepts and the world, and since some theories are better than others, some classifications will be better than others. This is how scientific concepts are anchored. I will now try to illustrate this claim with three different examples; the first two have been mentioned in previous chapters, but the third is new.

I have already stated that we can decide to use our words whichever way we wish, but that we cannot do the same for our explanatory concepts. We may link some terms indefeasibly to certain definitions, but if we take the empirical evidence seriously, we will find that we need to introduce different concepts to do the real work of science. In section 5.6., a hypothetical case was considered in which a classical physicist decided to reserve the term 'mass' for whatever quantity was given by the ratio of momentum to velocity. But

---

[13] See Shapere (1984, 246), where Gary Gutting makes the latter objection in a discussion of one of Shapere's articles.

the imaginary physicist could not make the concept into a truly explanatory one in the face of the special theory of relativity, because this quantity varies with the frame of reference and cannot be used to make the generalizations we need. Therefore, this concept can be retained, but another one will be needed to do the explanatory business of science; and this second one (rest mass) will correspond to our earlier concept mass if a comparison is carried out.

The constraint that external phenomena put on our use of concepts is perhaps stronger for concepts in the natural sciences than the social sciences, but it applies to the latter with some force also. A good illustration was provided by some of the concepts that Skinner discussed, which were considered in section 4.6. In one of his examples, the Puritans tried to extend the application of the term 'religious', so that 'religious' actions came to include ones involving strictness, punctuality, cleanliness, and so on. It was seen how this attempt was rejected by the linguistic community; rather than bring such actions into the extension of the term 'religious', the term itself became ambiguous when applied to some of the latter actions. This shows how theological concepts can even be constrained by certain theological "facts", which presumably include other doctrinal beliefs, evidence from scripture, and so on. Since the alternative classification is rejected for a good explanatory reason, this indicates that the way the concept is applied is not arbitrary, but is sensitive to theological reality.

These claims might seem to be challenged by a third example, which is derived from Imre Lakatos' account of the development of mathematics in Proofs and Refutations, but the example can in fact be used to corroborate my position. Lakatos suggests that a common situation in the development of mathematics unfolds as follows. A theorem is proposed about a certain class of mathematical objects. An exception is then found to this theorem among the objects to which it was thought to apply. Rather than rule that the theorem has been refuted, some mathematicians propose a modification of the original class of mathematical objects in such a way that the offending exceptions are excluded from its range of application. Notice that this method of "monster-barring" alters the extension of the concept with which we started in such a way that we manage to preserve a mathematical result that is a descendant of the one first introduced. Lakatos illustrates this strategy with reference to the history of Euler's conjecture concerning polyhedra, which

states that the number of vertices (V), edges (E), and faces (F), are related by the following formula: V - E + F = 2. When counterexamples are found to this formula, these are dismissed by the monster-barrers by delimiting the extension of 'polyhedron' in such a way that the counterexamples are excluded from the theorem. Lakatos describes the method of monster-barring thus: "Using this method one can eliminate any counterexample to the original conjecture by a sometimes deft but always ad hoc redefinition of the polyhedron, of its defining terms, or of the defining terms of its defining terms." (1976, 23)

The very fact that this monster-barring strategy is possible might be used to argue that the scientific phenomena (in this case, mathematical) do not always place very strenuous constraints on our concepts, thus showing that our concepts can roam freely without constraint and that we can just decide which explanatory concepts to adopt. Since it is possible to jump in either direction when faced with contradictory evidence, this may be seen to jeopardize my claim that explanatory efficacy is what guarantees a tight connection between our concepts and the world. In Lakatos' examples, it seems that we can decide either to say that polyhedra are whatever satisfy Euler's conjecture, or to say that polyhedra are a wider class of solids, some of which violate the Eulerian formula. If this is a feasible strategy, the phenomena we are investigating would not seem to constrain our concepts in the way that I have suggested (in this case, the concept polyhedron). But this strategy is only plausible when both classes of phenomena, the ones with the monsters barred and those that include the monsters, yield some interesting properties, which is to say that the monster-barring move is not completely ad hoc. To take the most extreme monster-barring move, imagine that someone decides to define polyhedra as those solid figures that obey Euler's rule. The only way that this proposal could be taken seriously would be if such objects also had other interesting properties in common; otherwise the move would be an empty one. By contrast, the actual monster-barrers[14] Lakatos discusses try less controversial moves, limiting the class of polyhedra using genuine explanatory properties, rather than defining the offending monsters out of existence in a completely ad

[14] Such moderates are not described as "monster-barrers" by Lakatos; instead, he calls them "exception-barrers".

hoc way. One such proposal does it as follows: "For all polyhedra that have no cavities, tunnels, 'multiple structure', V - E + F = 2."[15] (1976, 21) Still, if this restricted class of polyhedra does not turn out to be significant in its own right and does not generate interesting results, it does not have much to recommend it. As one of Lakatos' characters puts it:

> You have fallen in love with the problem of finding out where God drew the boundary dividing Eulerian from non-Eulerian polyhedra. But there is no reason to believe that the term 'Eulerian' occurred in God's blueprint of the universe at all. What if Eulerianness is merely an accidental property of some polyhedra? In this case it would be uninteresting or even impossible to find out the random zig-zags of the demarcation line between Eulerian and non-Eulerian polyhedra. Such an admission however would leave rationalism unsullied, for Eulerianness is then not part of the rational design of the universe. So let us forget about it. One of the main points of critical rationalism is that one is always prepared to abandon one's original problem in the course of the solution and replace it by another one. (1976, 67-8)

This little diatribe implies that there are joints even in the mathematical realm and that our concepts are designed to carve them rather than to figure in seemingly neat formulae that turn out to be restricted in scope, such as Euler's conjecture. If the category of Eulerian polyhedra does not have other explanatory properties, it is duly discarded and our concepts are made to track more significant distinctions.

The claim that the categories of science correspond to nature's own is one that is capable of various philosophical construals. On one reading, it can only be corroborated by a direct physical or metaphysical relation that is thought to obtain between scientific terms or concepts and entities or types of entity in nature. This is the kind of view associated with metaphysical realist theories of reference (criticized in the previous chapter). On

---

[15] This is a proposal put forward not by a monster-barrer, but by a proponent of the "exception-barring method", who uses monster-barring only to find "the domain of validity of the original conjecture", rather than using it as a "linguistic trick for rescuing 'nice' theorems by restrictive concepts." (1976, 26)

another, less rigid construal, it can be satisfied without positing a direct and privileged connection between words and the world. That is the view that I have been trying to advance. I do not pretend to have provided a complete account of the alternative connection that obtains between the categories of science and the divisions of nature. However, I have argued that scientific concepts have the content they do because of their explanatory efficacy, and that we are not at liberty to specify their content at will. As I argued in section 5.6., physicists could have decided by fiat to define the concept mass as momentum divided by velocity, but they cannot guarantee that this very concept will play a central role in the theory or continue to explain the phenomena in the face of theory change (indeed, it did not). Similarly, mathematicians may decide to define the concept polyhedron as whatever satisfies Euler's theorem, but this does not ensure that that concept will be crucial to their theorizing (indeed, it was not). Both examples illustrate the non-arbitrariness of our concepts and the fact that a mind-independent reality places constraints on our conceptual apparatus.

In addition to the manner in which their explanatory efficacy constrains our concepts, the interpretive approach can also appeal to a straightforward notion of extension which can be used to establish its realist credentials. Though different from a metaphysical realist understanding of reference, I argued in Chapter 6 that the extension of a scientist's term can be determined in a variety of ways, both linguistic and non-linguistic. The extension of a term helps the interpreter to determine the content of the associated concept, and the ascription of a concept, in holistic fashion, provides the interpreter with a clue to the extension of the relevant term. Moreover, as I argued in section 6.2., the interpreter can tell whether a concept is under- or over-extended by a scientist, so there is some distance between the concept and its extension.

These two aspects of my account, the explanatory efficacy of concepts and the relation beween concepts and extensions, together provide a dose of healthy realism concerning the categories of science. Moreover, it would be counter-productive to seek any more, say by demanding a demonstration that scientific categories really correspond to entities "out there". As Arthur Fine has argued, it is futile to try to satisfy "the realist's demand that we justify the existence claims sanctioned by science... as claims to the existence of entities 'out there'." (1984, 99) Rather, he suggests we adopt the "natural

ontological attitude" (NOA) and "accept the certified results of science as on a par with more homely and familiarly supported claims." (1984, 96) That attitude is compatible with the interpretive approach to comparing scientific theories.[16]

7.4. Incommensurability and NOA

In this chapter, I have tried to identify certain concrete ways in which the view of the meaning of scientific terms presented in this work might be accused of anti-realism and have proceeded to see whether such accusations were warranted. For better or worse, I have sidestepped much of the familiar debate about scientific realism in the course of developing this line. The reason for this is that it is often unclear what the nature of the disagreement is in that debate. Inasmuch as it is framed in terms of the relation of reference, it sometimes seems as if realists are distinguished by the fact that they take reference to be a theory-independent relation between an agent's use of a certain word and an object or set of objects in the agent's environment. In other words, what I have called a metaphysical realist view of reference is sometimes taken as necessary and sufficient to yield a realist view of science. If that is how it is generally understood, then the position being defended here is obviously not a realist one. However, that is by no means always the case. In a paper on the subject subtitled "Confessions of a Metaphysical Realist," Clark Glymour has written:

> The realist's conception of reference is a conception, not an analysis, and it is clearly utopian to suppose that one might define "refers to" or "signifies" in causal or physical terms. At most one should expect natural explanations of aspects of reference, or perhaps useful constraints, in like terms, on reference or co-reference, and fulfilling such expectations is as much a matter of science as of philosophy. (1982, 179)

---

[16] What may be thought incompatible with interpretivism is Fine's claim that "NOA sanctions ordinary referential semantics..." (1984, 98) However, he also dubs this a "Davidsonian-Tarskian referential semantics" (1984, 101), indicating that he does not have a metaphysical realist theory of reference in mind.

On the basis of such remarks, even a self-styled metaphysical realist like Glymour might find nothing objectionable in the position being advocated here--at least as far as realism goes.

The topic of incommensurability comes up explicitly in the debate surrounding scientific realism. However, there is an apparent willingness on the part of some philosophers who are not anti-realists to tolerate some degree of incommensurability, at least some of the time. This is illustrated by Fine's attempt to stake out the middle ground in the standoff between realists and anti-realists. Fine's position NOA, which was mentioned in the previous section, claims to consist in what the realist and anti-realist share. He summarizes it thus: "When NOA counsels us to accept the results of science as true, I take it that we are to treat truth in the usual referential way, so that a sentence (or statement) is true just in case the entities referred to stand in the referred-to relations." (1984, 98) So far, there is nothing that should alarm a proponent of the interpretive approach. What Fine goes on to say is more surprising:

> NOA is perfectly consistent with the Kuhnian alternative which counts such changes [i.e. paradigm shifts] as wholesale changes of reference. Unlike the realist, adherents to NOA are free to examine the facts in cases of paradigm shift, and to see whether or not a convincing case for stability of reference across paradigms can be made without superimposing on these facts a realist-progressivist superstructure. (1984, 98)

Here, Fine sends an ambivalent message: he indicates that stability of reference may obtain across theory-change but adds that every case must be decided on its own merits, so he fails to rule out the possibility of incommensurability. Clearly, his position is weaker than the position that I have been advocating, which claims to be able to defeat incommensurability and finds it damaging to the scientific enterprise precisely because it has anti-realist implications, at least pending an alternative method for comparing scientific theories.

The position of the present approach is, in this respect, more realist than the attitude Fine calls NOA. But that is not the main point. The willingness of Fine and other philosophers of science who are not anti-realists to tolerate incommensurability (at least on occasion) might be directly correlated with a reluctance to embrace a metaphysical

realist theory of reference similar to the causal theory. Since that course is often perceived to be the only one capable of defeating incommensurability, and since some philosophers of science feel uneasy with a metaphysical realist account of the reference of scientific terms, they might think that (occasional) incommensurability is a fair price to pay. If that is so, there is a certain irony in this, since the argument of this book has tried to show that the causal theory of reference and other metaphysical realist theories are incapable of defeating incommensurability. When it becomes clear that metaphysical realism about reference is not the way to thwart incommensurability, realist philosophers of science should cease to tolerate it. Fortunately, an alternative method of comparing theories is available, in the form of the interpretive approach.

Bibliography

Aristotle 1930. The Works of Aristotle: Vol. II, ed. W. D. Ross, trans. R.P. Hardie and R.K. Gaye, Oxford: Clarendon Press.

Atran, Scott 1990. Cognitive Foundations of Natural History: Towards an Anthropology of Science, Cambridge: Cambridge University Press.

Austin, J.L. 1975. How to Do Things with Words, Cambridge, Mass: Harvard University Press (second edition).

Bacon, Francis 1620/1985. The New Organon, ed. Fulton H. Anderson, New York: Macmillan.

Bechtel, William and Adele Abrahamsen 1991. Connectionism and the Mind: An Introduction to Parallel Processing in Networks, Oxford: Basil Blackwell.

Berger, Alan 1989. "A Theory of Reference Transmission and Reference Change," in Midwest Studies in Philosophy 14, Minneapolis: University of Minnesota Press.

Bilgrami, Akeel 1992. Belief and Meaning, Oxford: Blackwell.

Bostock, David 1991. "Aristotle on Continuity in Physics VI," in Aristotle's Physics: A Collection of Essays, ed. Lindsay Judson, Oxford: Clarendon Press.

Boyd, Richard 1979. "Metaphor and Theory Change: What Is 'Metaphor' a Metaphor For?" in Metaphor and Thought, ed. Andrew Ortony, Cambridge: Cambridge University Press.

Burge, Tyler 1979. "Individualism and the Mental," in Midwest Studies in Philosophy vol. 4: Studies in Metaphysics, ed. P.A. French et. al., Minneapolis: University of Minnesota Press.

Campbell, N.R. 1919/1957. Foundations of Science, New York: Dover.

Carey, Susan 1985. Conceptual Change in Childhood, Cambridge, Mass.: MIT Press

Carey, Susan 1986. "Constraints on Semantic Development," in Language Learning and Concept Acquisition: Foundational Issues, eds. W. Demopoulos and A. Marras, Norwood, NJ: Ablex.

Carey, Susan 1989. "Conceptual Differences Between Children and Adults," Mind and Language 3, 167-181.

Carnap, Rudolf 1955. "Meaning and Synonymy in Natural Languages," in Meaning and Necessity, Chicago: University of Chicago Press (first published 1955).

Carnap, Rudolf 1956. "The Methodological Character of Theoretical Concepts," in Minnesota Studies in the Philosophy of Science vol. 1, eds. Herbert Feigl et. al., Minneapolis: University of Minnesota Press.

Carnap, Rudolf 1963. "Replies and Systematic Expositions," in The Philosophy of Rudolf Carnap, ed. P.A. Schilpp, La Salle, Il: Open Court.

Chomsky, Noam 1989. "Language from an Internalist Perspective," ms, Conference on Method, New York City.

Chomsky, Noam 1992. "Language and Interpretation: Philosophical Reflections and Empirical Inquiry," in Earman (1992).

Chomsky, Noam 1995. "Language and Nature," Mind 104, 1-61.

Churchland, Paul M. 1979. Scientific Realism and the Plasticity of Mind, Cambridge: Cambridge University Press.

Churchland, Paul M. 1989. A Neurocomputational Perspective, Cambridge, Mass.: MIT Press.
Churchland, Paul M. 1992. "A Deeper Unity: Some Feyerabendian Themes in Neurocomputational Form," in Giere (1992).
Clark, Andy 1989. Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing, Cambridge: MIT Press.
Davidson, Donald 1968. "On Saying That," in Davidson (1986) (first published 1968).
Davidson, Donald 1974a. "On the Very Idea of a Conceptual Scheme," in Davidson (1986) (first published 1974).
Davidson, Donald 1974b. "Belief and the Basis of Meaning," in Davidson (1986) (first published 1974).
Davidson, Donald 1975. "Thought and Talk," in Davidson (1986) (first published 1975).
Davidson, Donald 1977a. "Reality Without Reference," in Davidson (1986) (first published 1977).
Davidson, Donald 1979. "The Inscrutability of Reference," in Davidson (1986) (first published 1979).
Davidson, Donald 1986. Inquiries into Truth and Interpretation, Oxford: Oxford University Press.
Dennett, Daniel C. 1987. The Intentional Stance, Cambridge: M.I.T. Press.
Donnellan, Keith 1989. "Belief and the Identity of Reference," in Midwest Studies in Philosophy 14, Minneapolis: University of Minnesota Press.
Dummett, Michael 1974. "The Significance of Quine's Indeterminacy Thesis," in Dummett (1978) (first published 1974).
Dummett, Michael 1978. Truth and Other Enigmas, Cambridge, Mass.: Harvard University Press.
Dupré, John 1981. "Natural Kinds and Biological Taxa," Philosophical Review 90, 66-90.
Dupré, John 1993. The Disorder of Things, Cambridge: Harvard University Press.
Earman, John 1977. "Against Indeterminacy," Journal of Philosophy 74, 535-538.
Earman, John ed. 1992. Inference, Explanation, and Other Frustrations: Essays in the Philosophy of Science, Berkeley: University of California Press.
Earman, John 1993. "Carnap, Kuhn, and the Philosophy of Scientific Methodology," in Horwich (1993).
Earman, John and Michael Friedman 1973. "The Meaning and Status of Newton's Law of Inertia and the Nature of Gravitational Forces," Philosophy of Science 40, 329-359.
Evans, Gareth 1973. "The Causal Theory of Names," in Schwartz (1977) (first published 1973).
Feyerabend, Paul K. 1962 "Explanation, Reduction, and Empiricism," in Feyerabend (1981) (first published 1962).
Feyerabend, Paul K. 1965a "On the 'Meaning' of Scientific Terms," in Feyerabend (1981) (first published 1965).
Feyerabend, Paul K. 1965b. "Reply to Criticism: Comments on Smart, Sellars, and Putnam," in Feyerabend (1981) (first published 1965).
Feyerabend, Paul K. 1970. "Consolations for the Specialist," in Criticism and the Growth of Knowledge, ed. I. Lakatos and A. Musgrave (Cambridge: Cambridge University Press, 1970), pp.197-230.

Feyerabend, Paul K. 1975. Against Method, London: Verso.
Feyerabend, Paul K. 1981. Realism, Rationalism, and Scientific Method: Philosophical Papers vol. 1, Cambridge: Cambridge University Press.
Feyerabend, Paul K. 1987. Farewell to Reason (London: Verso).
Field, Hartry 1973. "Theory Change and the Indeterminacy of Reference," Journal of Philosophy 70, 462-481.
Fine, Arthur 1967. "Consistency, Derivability, and Scientific Change," Journal of Philosophy 64, 231-240.
Fine, Arthur 1975. "How to Compare Theories: Reference and Change," Noûs 9, 17-32.
Fine, Arthur 1977. "Appendix," Journal of Philosophy 72, 538.
Fine, Arthur 1984. "The Natural Ontological Attitude," in Leplin (1984).
Flanagan, Owen 1984. The Science of the Mind, Cambridge, Mass.: M.I.T. Press.
Fodor, Jerry 1987. Psychosemantics: The Problem of Meaning in the Philosophy of Mind, Cambridge, Mass.: M.I.T. Press.
Foucault, Michel 1973. The Birth of the Clinic, trans. by A.M. Sheridan Smith, New York: Vintage Books.
Gazdar, Gerald 1979. Pragmatics: Implicature, Presupposition, and Logical Form, New York: Academic Press.
Giere, Ronald N. ed. 1992. Cognitive Models of Science: Minnesota Studies in the Philosophy of Science vol. 15, Minneapolis: University of Minnesota Press.
Giere, Ronald N. 1994. "The Cognitive Structure of Scientific Theories," Philosophy of Science 61, 276-296.
Gleitman, Lila R., Sharon Lee Armstrong, and Henry Gleitman 1983. "On Doubting the Concept 'Concept'," in E.K. Scholnick ed., New Trends in Conceptual Representation: Challenges to Piaget's Theory? Hillsdale, NJ: Erlbaum.
Glymour, Clark 1982. "Conceptual Scheming or Confessions of a Metaphysical Realist," Synthese 51, 169-180.
Glymour, Clark 1992. "Invasion of the Mind Snatchers," in Giere (1992).
Grandy, Richard E. 1973. "Reference, Meaning, and Belief," Journal of Philosophy 70, 439-452.
Grandy, Richard E. 1992. "Theories of Theories: A View from Cognitive Science," in Earman (1992).
Hacking, Ian 1982. "Language, Truth and Reason," in Rationality and Relativism, ed. M. Hollis and S. Lukes, Cambridge, Mass: M.I.T. Press.
Hacking, Ian 1984. "Experimentation and Scientific Realism," in Scientific Realism, ed. Jarrett Leplin, Berkeley, Calif.: University of California Press.
Hacking, Ian 1986. "The Parody of Conversation," in Truth and Interpretation, ed. Ernest LePore, New York: Blackwell.
Hacking, Ian 1993. "Working in a New World: The Taxonomic Solution," in Horwich (1993).
Harman, Gilbert 1973. Thought, Princeton, N.J.: Princeton University Press.
Haugeland, John 1978. "The Nature and Plausibility of Cognitivism," Behavioral and Brain Sciences 2, 215-260.
Haugeland, John 1981. Mind Design, Cambridge, MA: MIT Press.

Hempel, Carl G. 1963. "Implications of Carnap's Work for the Philosophy of Science," in The Philosophy of Rudolf Carnap, ed. Paul Arthur Schilpp, La Salle, Illinois: Open Court.

Hempel, Carl G. 1966. Philosophy of Natural Science, Englewood Cliffs, N.J.: Prentice-Hall.

Hempel, Carl G. 1969. "Formulation and Formalization of Scientific Theories" and "Discussion" in Suppe (1977) (delivered in 1969).

Horwich, Paul 1993. World Changes: Thomas Kuhn and the Nature of Science, Cambridge, MA: MIT Press.

Hussey, Edward 1991. "Aristotle's Mathematical Physics: A Reconstruction," in Aristotle's Physics: A Collection of Essays, ed. Lindsay Judson, Oxford: Clarendon Press.

Irzik, Gürol and Teo Grünberg 1995. "Carnap and Kuhn: Arch Enemies or Close Allies?" British Journal for the Philosophy of Science 46, 285-307.

Kant, Immanuel 1787/1933. Critique of Pure Reason, trans. Norman Kemp Smith, London: Macmillan.

Kaplan, David 1989. "Afterthoughts," in Themes from Kaplan, ed. Joseph Almog et. al., Oxford: Oxford University Press.

Keil, Frank C. 1979. Semantic and Conceptual Development: An Ontological Perspective, Cambridge: Harvard University Press.

Keil, Frank C. 1986. "The Acquisition of Natural Kind and Artifact Terms," in Language Learning and Concept Acquisition: Foundational Issues, ed. W. Demopoulos and A. Marras, Norwood, NJ: Ablex.

Keil, Frank C. 1989. "Spiders in the Web of Belief: The Tangled Relations Between Concepts and Theories," Mind and Language 4, 43-50.

Khalidi, Muhammad Ali 1993a. "Carving Nature at the Joints," Philosophy of Science 60, 100- 113.

Khalidi, Muhammad Ali 1993b. Review of Holism by Jerry Fodor and Ernest Lepore, Mind 102, 650-4

Khalidi, Muhammad Ali 1995. "Two Concepts of Concept," Mind & Language 10, 402-422.

Khalidi, Muhammad Ali 1998a. "Natural Kinds and Crosscutting Categories," Journal of Philosophy 95, 3-50.

Khalidi, Muhammad Ali 1998b. "Incommensurability in Cognitive Guise," Philosophical Psychology 11, 29-43.

Kitcher, Philip 1978. "Theories, Theorists, and Theoretical Change," Philosophical Review 87, 519-547.

Kitcher, Philip 1982. "Genes," British Journal for the Philosophy of Science 33, 337-359.

Kripke, Saul A. 1979. "A Puzzle About Belief," in Meaning and Use, ed. A. Margalit, Dordrecht: Reidel.

Kripke, Saul A. 1980. Naming and Necessity, Cambridge, MA: Harvard University Press.

Kuhn, Thomas S. 1964. "A Function for Thought Experiments," in Kuhn (1977) (first published 1964).

Kuhn, Thomas S. 1970a. The Structure of Scientific Revolutions, Chicago: University of Chicago Press (second edition).

Kuhn, Thomas S. 1970b. "Reflections on My Critics," in Criticism and the Growth of Knowledge, ed. Imre Lakatos and Alan Musgrave, Cambridge: Cambridge University Press.

Kuhn, Thomas S. 1976. "Theory-Change as Structure-Change: Comments on the Sneed Formalism," Erkenntnis 10, pp.179-199.
Kuhn, Thomas S. 1977. The Essential Tension, Chicago: University of Chicago Press.
Kuhn, Thomas S. 1983a. "Commensurability, Comparability, Communicability," PSA 1982, 669-688.
Kuhn, Thomas S. 1983b. "Response to Commentaries," PSA 1982, 712-716.
Kuhn, Thomas S. 1990. "Dubbing and Redubbing: the Vulnerability of Rigid Designation," Minnesota Studies in the Philosophy of Science, Vol.14, Scientific Theories, ed. C. Wade Savage, Minneapolis: University of Minnesota Press, pp.298-318.
Lakatos, Imre 1976. Proofs and Refutations: The Logic of Mathematical Discovery, ed. John Worrall and Elie Zahar, Cambridge: Cambridge University Press.
Leplin, Jarrett 1969. "Meaning Variance and the Comparability of Theories," British Journal for the Philosophy of Science 20, 69-80.
Leplin, Jarrett, ed. 1984. Scientific Realism, Berkeley, Calif.: University of California Press.
Leplin, Jarrett 1988. "Is Essentialism Unscientific?" Philosophy of Science 55, 493-510.
Levi, Isaac 1960. "Must the Scientist Make Value Judgments?" Journal of Philosophy 57, 345- 357.
Levi, Isaac 1980. The Enterprise of Knowledge, Cambridge, Mass.: M.I.T. Press.
Levinson, Stephen C. 1983. Pragmatics, Cambridge: Cambridge University Press.
Lewis, David 1970. "How to Define Theoretical Terms," Journal of Philosophy 67, 427-446.
Lewis, David 1974. "Radical Interpretation," Synthese 23, 331-344.
Lowe, E.J. 1995. "The Metaphysics of Abstract Objects," Journal of Philosophy 92, 509-524.
MacIntyre, Alasdair 1981. After Virtue, Notre Dame, Indiana: University of Notre Dame Press.
Malt, Barbara 1994. "Water Is Not H2O," Cognitive Psychology 27, 41-70.
Mayr, Ernst 1963. "Species Concepts and Their Applications," in Conceptual Issues in Evolutionary Biology, ed. Elliott Sober, Cambridge, MA: MIT Press, 1984 (chapter first published 1963).
Murphy, Gregory L. and Douglas L. Medin 1985. "The Role of Theories in Conceptual Coherence," Psychological Review 92, 289-316.
Nagel, Ernest 1979. The Structure of Science, Indianapolis: Hackett.
Nersessian, Nancy 1984. Faraday to Einstein: Constructing Meaning in Scientific Theories, Dordrecht: Nijhoff.
Nola, Robert 1980. "Fixing the Reference of Theoretical Terms," Philosophy ofScience 47, 505- 531.
Papineau, David 1979. Theory and Meaning, Oxford: Oxford University Press.
Papineau, David 1996. "Theory-Dependent Terms," Philosophy of Science 63, 1-20.
Parsons, Kathryn Pyne 1971. "On Criteria of Meaning Change," British Journal for Philosophy of Science 22, 131-144.
Parsons, Kathryn Pyne 1975. "A Criterion for Meaning Change," Philosophical Studies 28, 367- 396.
Partington, J.R. 1937. A Short History of Chemistry, New York: Harper and Row.
Putnam, Hilary 1962. "The Analytic and the Synthetic," in Putnam (1975a) (first published 1962).
Putnam, Hilary 1970. "Is Semantics Possible?" in Putnam (1975a) (first published 1970).

Putnam, Hilary 1973a. "Explanation and Reference," in Putnam (1975a) (first published 1973).
Putnam, Hilary 1973b. "Meaning and Reference," in Schwartz (1977) (first published 1973).
Putnam, Hilary 1975a. Mind, Language, and Reality: Philosophical Papers vol.2, Cambridge: Cambridge University Press.
Putnam, Hilary 1975b. "The Meaning of 'Meaning'," in Putnam (1975a) (first published 1975).
Putnam, Hilary 1975c. "The Refutation of Conventionalism," in Putnam (1975a) (first published 1975).
Putnam, Hilary 1979. "Comments," on Kripke (1979) in Meaning and Use, ed. A. Margalit, Dordrecht: Reidel.
Putnam, Hilary 1986. "Meaning Holism," in The Philosophy of W.V. Quine, eds. Lewis Edwin Hahn and Paul Arthur Schilpp, La Salle, Ill.: Open Court.
Putnam, Hilary 1988. Representation and Reality, Cambridge, Mass.: M.I.T. Press.
Quine, W.V. 1960. Word and Object, Cambridge, Mass.: M.I.T. Press.
Quine, W.V. 1961. From a Logical Point of View, second edition, Cambridge, Mass.: Harvard University Press.
Quine, W.V. 1969. Ontological Relativity and Other Essays, New York: Columbia University Press.
Quine, W.V. 1975. "On Empirically Equivalent Systems of the World," Erkenntnis 9, 313-328.
Quine, W.V. 1984. "Relativism and Absolutism," The Monist 67, 293-295.
Rawls, John 1971. A Theory of Justice, Cambridge: Harvard University Press.
Rorty, Richard 1980. Philosophy and the Mirror of Nature, Princeton, N.J.: Princeton University Press.
Rumelhart, David E. and James L. McClelland 1986. Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1: Foundations, Cambridge, Mass.: M.I.T. Press.
Salmon, Nathan U. 1981. Reference and Essence, Princeton, N.J.: Princeton University Press.
Scheffler, Israel 1967. Science and Subjectivity, Indianapolis: Bobbs-Merrill.
Schwartz, Stephen P., ed. 1977. Naming, Necessity, and Natural Kinds, Ithaca, N.Y.: Cornell University Press.
Searle, John R. 1969. Speech Acts: An Essay in the Philosophy of Language, Cambridge: Cambridge University Press.
Searle, John R. 1983. Intentionality: An Essay in the Philosophy of Mind, Cambridge: Cambridge University Press.
Shapere, Dudley 1982a. "Reason, Reference, and the Quest for Knowledge," Philosophy of Science 49, 1-23.
Shapere, Dudley 1982b. "The Concept of Observation in Science and Philosophy," Philosophy of Science 49, 485-525.
Shapere, Dudley 1984. Reason and the Search for Knowledge, Dordrecht: Reidel.
Singer, Peter 1983. Hegel, Oxford: Oxford University Press.
Skinner, Quentin 1980. "Language and Social Change," in Tully (1988) (first published 1980).

Sklar, Abe 1964. "On Category Overlapping in Taxonomy," in Form and Strategy in Science, ed. John R. Gregg and F.T.C. Harris, Dordrecht: Reidel.
Smith, Peter 1981. Realism and the Progress of Science, Cambridge: Cambridge University Press.
Smith, Edward E. and Douglas L. Medin 1981. Categories and Concepts, Cambridge: Harvard University Press.
Sommers, Fred 1963. "Types and Ontology," Philosophical Review 72, 327-363.
Stegmüller, Wolfgang 1979. The Structuralist View of Theories, Berlin: Springer-Verlag.
Strawson, P.F. 1985. Skepticism and Naturalism, New York: Columbia University Press.
Suppe, Frederick 1979. The Structure of Scientific Theories, Urbana, Illinois: University of Illinois Press.
Tarski, Alfred 1956. "The Concept of Truth in Formalized Languages," in Logic, Semantics, Meta-Mathematics, trans. J.H. Woodger, Indianapolis: Hackett.
Tarski, Alfred 1969. "Truth and Proof," Scientific American 20:6, 63-77.
Thomason, Richmond 1969. "Species, Determinates, and Natural Kinds," Noûs 3, 95-101.
Tully, James ed. 1988. Meaning and Context, Princeton, N.J.: Princeton University Press.
Wallace, John 1979. "Only in the Context of a Sentence Do Words Have Any Meaning," in Contemporary Perspectives in the Philosophy of Language, ed. P.A. French et. al., Minneapolis: University of Minnesota Press.
White, Stephen L. 1982. "Partial Character and the Language of Thought," Pacific Philosophical Quarterly 63, 347-365.
Williams, Raymond 1976. Keywords: A Vocabulary of Culture and Society, London: Fontana.
Wimsatt, William C. 1994. "The Ontology of Complex Systems: Levels of Organization, Perspectives, and Causal Thickets," ms, University of Chicago.
Woodfield, Andrew 1991. "Conceptions," Mind 100, 547-572.
Woodfield, Andrew 1993. "Do Your Concepts Develop?" in Philosophy and Cognitive Science, ed. Christopher Hookway and Donald Peterson, Cambridge: Cambridge University Press.
Woodfield, Andrew 1996. "Which Theoretical Concepts Do Children Use?" Philosophical Papers 25, 1-20.
Zemach, Eddy 1976. "Putnam's Theory on the Reference of Substance Terms," Journal of Philosophy 73, 116-127.
Zwicky, Arnold and Jerrold Sadock 1975. "Ambiguity Tests and How to Fail Them," in Syntax and Semantics, vol. 4, ed. J.P. Kimball, New York: Academic Press.