# Conceptual Systems

Harold I. Brown

Routledge Studies in the Philosophy of Science

# Conceptual Systems

It is argued that the introduction of new concepts and the abandonment of older concepts are persistent features of human thought as we discover new phenomena and re-examine familiar phenomena in the light of developments in science, technology and society. In recent years conceptual change and any consequent incommensurability have become important topics in philosophy and the philosophy of science. *Conceptual Systems* seeks to understand how radically new concepts are introduced into our thinking while maintaining sufficient continuity with older concepts to ensure intelligibility.

The book provides a unified account of the nature of concepts, with particular emphasis on the development of scientific concepts. Harold I. Brown establishes a database of examples of conceptual change in science, mathematics, society, and philosophy, and critically examines the influential theories of concepts in modern philosophy, documenting the way in which different theories of concepts provide different criteria for a successful conceptual analysis. The author then constructs a new theory of concepts that builds on the work of Wilfrid Sellars. The theory is applied to two types of problems: rethinking the nature and purpose of conceptual analysis, and studying conceptual change in the history of science – a task that requires analysis of the concepts being examined. *Conceptual Systems* then presents two new studies of conceptual change in physics, developments in the seventeenth century from Galileo to Descartes to Newton and the conceptual framework of the "standard model" in late twentieth-century high-energy physics. These studies illustrate how the theory of concepts developed here can guide historical studies while providing further tests of the adequacy of the theory.

This book will be welcomed by philosophers, philosophers of science and cognitive scientists interested in concepts.

**Harold I. Brown** is Professor Emeritus at Northern Illinois University, USA. His previous books include *Rationality* also published by Routledge.

## Routledge Studies in the Philosophy of Science

**1 Cognition, Evolution and Rationality**
A cognitive science for the twenty-first century
*Edited by António Zilhão*

**2 Conceptual Systems**
*Harold I. Brown*

# Conceptual Systems

**Harold I. Brown**

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2007
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Simultaneously published in the USA and Canada
by Routledge
270 Madison Ave, New York, NY 10016

Routledge is an imprint of the Taylor & Francis Group

This edition published in the Taylor & Francis e-Library, 2007.

"To purchase your own copy of this or any of Taylor & Francis or Routledge's collection of thousands of eBooks please go to www.eBookstore.tandf.co.uk."



*British Library Cataloguing in Publication Data*
A catalogue record for this book is available from the British Library

*Library of Congress Cataloging in Publication Data*
A catalog record for this book has been applied for

ISBN 0-203-96790-9 Master e-book ISBN

ISBN13: 978-0-415-70182-2 (hbk)
ISBN13: 978-0-203-96790-4 (ebk)

# Contents

# Preface

> The motto of the age of science might well be: Natural philosophers have hitherto sought to understand "meanings"; the task is to change them.
>
> (CDCM 288)

I have a long-standing interest in the ways our conceptual repertoires change as knowledge develops. It is, I think, clear that human adults in all societies and all historical periods do not somehow already possess the concepts needed to think about all discoveries throughout the past and future history of science, all the various economic, social, and political arrangements that we may come up with, and all of the other endeavors that may engage us. It seems equally clear that the concepts people use to think about aspects of the world often turn out to be inadequate; sometimes the items we think about do not exist at all. But conceptual variation raises serious questions about the evaluation of fundamental scientific theories, as well as about our ability to understand the thought of other cultures, earlier periods of our own culture, and even our neighbors. It also raises questions about the nature of conceptual innovation. While history provides powerful evidence of radical conceptual innovation, any innovation requires substantial continuity with older concepts in order to be intelligible. Thus to understand the development of human knowledge we must understand this interplay between innovation and continuity.

How we deal with these questions depends (in part) on our understanding of the nature of concepts. Attempts to understand new concepts, concepts from other times and places, and even our own concepts, point to the need for conceptual analysis – a central concern of philosophers in the community in which I work. Yet it is also clear that how we pursue this endeavor, and how we assess the adequacy of a proposed analysis, depends on our view of conceptual content. Reflection on conceptual analyses also raises questions about the significance of such analyses. Do analyses clarify the mode of thought of a culture, sub-culture, or individual, or do they have some wider scope? If we take the latter to be the case, how do we know this?

Over the years I became convinced that Wilfrid Sellars provides the best available approach to an account of conceptual content. Sellars is also a scientific realist who recognizes that finding the correct concepts to describe aspects of the world is a task for scientific research – so that realism requires conceptual innovation. Sharing many of Sellars' views, I set out to write a book in which I would explicate Sellars' theory of concepts and then apply it to case studies in the history of science, and to the analysis of two central concepts: causation and truth. I chose these concepts partly because they are central philosophical topics, but also because Sellars discusses these concepts in many places without using the resources of his own theory of concepts. My original plan was to write three papers and then take these as the basis of a book. Two of these papers have appeared (Brown 1986, 1991), but my work on causation encountered major roadblocks. Eventually I became convinced that Sellars' theory of concepts was not adequate as he left it. To pursue the project I would have to do more than just explicate Sellars' account; modifications and extensions were required. Continued work on causation, truth, and the conceptual development of science convinced me of the need for even more drastic modifications of Sellars' approach than I had previously considered. I am still convinced that Sellars provides the best starting point for a theory of concepts, and he remains the central figure in this book. I have attempted to go beyond him in a number of respects and to use my results in ways that he never pursued, but I believe that these attempts to develop and apply his ideas are wholly in tune with the Sellarsian spirit.

# Acknowledgments

Several individuals and institutions provided support at various stages of this project. Institutional support includes a Summer Research Stipend from the National Endowment for the Humanities (1990) and National Science Foundation STS Research Grant #9818094 (1999). Northern Illinois University provided a Sabbatical and a Summer Research Stipend. Tomoji Shogenji and Herman Stark commented on the entire manuscript and are responsible for significant improvements. I also received important comments on one or more chapters from Paul Bowen, Raymond Brock, Xiang Chen, Paul Hoyningen-Huene, Howard Sankey, and Michael Shaffer. Alas, I have not always followed their advice.

I wish to thank the copyright holders for permission to print excerpts from the following books:

Descartes, R. *Principles of Philosophy*, trans. V. Miller and R. Miller, Dordrecht: Kluwer, 1991, with the kind permission of Springer Science and Business Media.

Galileo, *Dialogue Concerning the Two Chief World Systems*, trans, S. Drake, Berkeley: University of California Press, 1967.

Newton, I., *The Principia: mathematical principles of natural philosophy*, trans. I. Cohen and A. Whitman, Berkeley: University of California Press, 1999.

# Abbreviations

Sellars' work is frequently cited throughout this book using the abbreviations below. Most of Sellars writings consist of articles that appeared in journals or collections of papers. Sellars' books are often collections of previously published papers; NO and SM are exceptions. When citing Sellars I generally use the original publication although this has sometimes been overridden by considerations of accessibility. This is particularly relevant to the collection SPR, which approximates a unified book and is a vital source. Papers in this volume are noted below; in citing these papers I give page references to SPR. Publication dates of other pieces are given as they occur in the reference list at the end of the present book. Other, more local, abbreviations are given in the relevant chapter or section.

| | |
|---|---|
| CC | "Conceptual Change" (1973) |
| CDCM | "Counterfactuals, Dispositions, and the Causal Modalities" (1958) |
| CIL | "Concepts as Involving Laws and Inconceivable Without Them" (1948a) |
| EAE | "Empiricism and Abstract Entities" (1963a) |
| ENNW | "Epistemology and the New Way of Words" (1947a) |
| EPM | "Empiricism and the Philosophy of Mind" (in SPR) |
| GE | "Grammar and Existence" (in SPR) |
| IILO | "Imperatives, Intentions, and the Logic of Ought" (1963b) |
| IM | "Inference and Meaning" (1953) |
| IV | "Induction as Vindication" (1964) |
| LRB | "Language, Rules, and Behavior" (1950) |
| LT | "The Language of Theories" (in SPR) |
| LTC | "Language as Thought and Communication" (1969) |
| ME | "Mental Events" (1981) |
| MFC | "Meaning as Functional Classification" (1974a) |
| MGEC | "More on Givenness and Explanatory Coherence" (1979a) |
| NO | *Naturalism and Ontology* (1979b) |
| OM | "Obligation and Motivation" (1952) |
| P | "Phenomenalism" (in SPR) |

| | |
|---|---|
| PPE | "Pure Pragmatics and Epistemology" (1947b) |
| PSM | "Philosophy and the Scientific Image of Man" (in SPR) |
| PT | "Particulars" (in SPR) |
| RM | "Reply to Marras" (1974b) |
| RNWW | "Realism and the New Way of Words" (1948b) |
| SAP | "Is There a Synthetic A Priori?" (in SPR) |
| SE | "Science and Ethics" (1967a) |
| SK | "The Structure of Knowledge" (1975) |
| SM | *Science and Metaphysics* (1968) |
| SPR | *Science, Perception and Reality* (1963c) |
| SRII | "Scientific Realism or Irenic Instrumentalism" (1965) |
| SRLG | "Some Reflections on Language Games" (in SPR) |
| SRTT | "Some Reflections on Thoughts and Things" (1967b) |
| SS | "Sensa or Sensings: Reflections on the Ontology of Perception" (1982) |
| TA | "Thought and Action" (1966) |
| TC | "Truth and Correspondence" (in SPR) |
| TE | "Theoretical Explanation" (1963d) |
| TWO | "Time and the World Order" (1962) |

# Notation

For the most part I depend on context to make it clear whether I am discussing a concept, a word, or an item that is neither linguistic nor a concept. When context is not sufficient – and sometimes for emphasis – I use quotation marks to indicate a linguistic item (e.g., "word") and small capital letters for terms that refer to concepts (e.g., CONCEPT).

# 1 Studying Concepts

> Concepts are the glue that holds our mental world together.
>
> (Murphy 2002: 1)

## 1.1 Orientation

Studies of concepts are central to several disciplines including, at least, anthropology, cognitive neurobiology, intellectual history, linguistics, philosophy, psychology, and sociology. This is as it should be since concepts play a central role in human thought. Yet this last claim is fraught with ambiguities since how we understand it, and whether we think it true, depends on our view of the nature of concepts. At the same time, our view of the nature of concepts will typically be constrained by the specific questions we are asking – which, in turn, may be a function of the discipline we are coming from and the state of that discipline. For example, when the physiological psychologist Hebb (1949) wrote about concepts he was mainly concerned with identifying neural structures at the basis of what psychologists refer to as concepts. Once he identified these structures he attempted to use them as the starting point for a purely neurological account of thought. Literally, for Hebb, concepts are in the head.

Other researchers, such as Fodor (e.g., 1975, 1988, 1998), agree that concepts are in the head – in the sense that they are mental particulars possessed by individuals – but do not study them in physiological terms. Fodor's work straddles linguistics, philosophy, and psychology; much of this work is focused on language, and thus on the theory of meaning. As a result, one can easily be led to wonder if Hebb and Fodor are studying the same subject; an example will underline the contrast. One of Hebb's key claims is that the neural basis of a concept is a series of neurons that form a closed loop; one of Fodor's key claims is that concepts are semantically evaluable. It is not immediately clear how these views relate. They may be complementary, at odds with each other, or independent parts of a single account.

While Fodor and Hebb view concepts as individual possessions, others reject this thesis. One line of argument is found among philosophers and

sociologists influenced by Wittgenstein's later work (1953). On this approach concepts are social entities so that it is impossible in principle for an isolated individual to have concepts (cf., Kripke 1982; Winch 1958). For Fodor and Hebb the existence of other people is irrelevant to the question of what concepts I possess – although others may be relevant to an account of how I acquired these concepts. Others reject both psychological and sociological theories of concepts for a quite different reason. Frege (1997), for example, held that concepts are abstract entities that exist independently of what occurs in any mind. He sought to eliminate all psychological considerations from the study of concepts, and it is clear that he would have extended his views to sociological considerations had that been a subject of discussion in his day.

Consider another contrast. Students of intellectual history are often strongly impressed by differences in the concepts we find in various historical settings; many anthropologists and sociologists are equally impressed by variations across societies. But the current practice of conceptual analysis by philosophers assumes that there is some deep sense in which concepts – or, at least, certain key concepts – are universal and unchanging. Philosophers who make this assumption are content to analyze concepts by armchair reflection, and are prepared to debate such questions as whether Aristotle or Descartes got *the* concept of knowledge right.

Some of these disparities arise because of differences in the focal questions of different disciplines. It would be helpful if we had a wider perspective for examining the outcomes of these disparate approaches and assessing whether they contribute to some common project, conflict, or deal with different questions altogether. My main goal in this book is to contribute to this wider project by developing a theory of concepts and using that theory to resolve some of the problems about concepts that are currently in play. Since I do not claim to transcend normal disciplinary limitations, I think it appropriate to give the reader fair warning about the directions from which I approach the topic. My interest in understanding concepts comes largely from studies of the history of science. It seems to me that attempts to find the right concepts for thinking about various aspects of the world constitutes a major theme in the development of science. In pursuing this goal scientists invent concepts, try them out, sometimes improve them, and sometimes abandon them. We will see that such conceptual change occurs in fields besides the sciences. Thus one major task for a theory of concepts is to provide an account of how new concepts are introduced into ongoing research in a coherent manner. Those familiar with the literature of philosophy of science since the late 1950s will recognize the kinds of problems that concern me; I will say a bit more about the nature of these problems in Sec. 1.6. In my view, discussions of conceptual development typically underestimate the scope of conceptual innovation in human thought. Thus in Ch. 2 I will provide a large number of examples of conceptual change in several fields, and a preliminary discussion of some of the forms of conceptual innovation that we find.

I have a second major concern in this book that derives from my professional concerns as a philosopher. Acknowledging large-scale conceptual change in the course of human cognitive history raises fundamental problems about the nature and purpose of conceptual analysis. Studies of conceptual change require analysis of the concepts being studied, but philosophers typically hold that the outcome of a conceptual analysis is not just a description of a local mode of thought. Indeed, such historical study is an empirical endeavor, and many philosophers maintain that their studies of concepts are, in some deep sense, a priori. I examine the nature of conceptual analysis in some detail in Chs 7 and 8, after I have developed the theory of concepts I wish to propose. In the present chapter I will give a somewhat more extended sketch of the main issues that I plan to address in this book, and explain my own philosophical approach in more detail. Still, what I say in this chapter should be read as a preliminary orientation; my views on many of the topics I am now discussing will become fully clear only as my detailed theory of concepts develops. I return to several of these issues throughout the book, but I want to stress two features of my approach at the outset.

First, many studies of concepts, particularly in philosophy and psychology, focus on relatively simple concepts and on the ways in which these are learned – with special emphasis on how they are learned by young children. This is important work, but I will not pursue it here. My primary focus will be on some of the most sophisticated concepts in our repertoire, and the theory I propose will be developed to handle sophisticated adult thought.[1] This approach need not be viewed as a competitor to the more common approach since an adequate theory of concepts will have to encompass both ends, as well as the middle ground. I prefer to think of the relation between studies of conceptual development in children and studies of highly sophisticated concepts as analogous to driving a tunnel under a mountain from both ends. In modern tunnel building it is reasonable to expect that the two parts will meet, and if we are really lucky something like this will happen with studies of concepts that start from these opposite ends. At the present stage in studies of concepts it is more likely that the two strands will miss and that adjustments to each will be needed. I will not attempt anything quite so grandiose here. Although I will propose a general theory of concepts, I think of this theory as an attempt to contribute to a larger project whose completion lies in the future.

Second, I want to state where I stand on three types of questions that are commonly raised about concepts. Consider first two *ontological* questions: what kinds of entities concepts are, and where in reality they are located. In this book I will treat concepts as mental entities – items that exit in the minds of individual cognitive agents whatever minds ultimately turn out to be. (Thus I will leave the first of my two ontological questions open.) In treating concepts as mental entities I will be following a practice that is standard in psychology, but rejected by many contemporary philosophers –

although not by all (e.g., Prinz 2002; Rey 1999). Whatever role society plays in an individual's acquisition and use of concepts, there is still a distinction between individuals who have a particular concept and those who do not. Something must occur in an individual when a concept is acquired, and whatever this is, it may remain in place if that individual leaves the society in which that concept was acquired. *Next*, given this view of the ontological status of concepts, the key question in dispute is the nature of *conceptual content*. Thus the expression "theory of concepts" should be read as an abbreviation for "theory of conceptual content" unless explicit reasons are given for some other reading. *Finally*, there is an *epistemological* question: What reasons do we have for believing that concepts, understood as mental entities, exist? In my view concepts are a theoretical postulate introduced to explain a variety of cognitive phenomena; the explanatory success of this postulate provides the grounds for accepting it. Thus I will propose a theory of conceptual content and defend that theory on the basis of its explanatory power. The assumption that concepts are mental entities will be central to that theory, and the argument for this theory will thus constitute an argument for the claim that concepts exist.

## 1.2 Conceptual Variation

Even brief reflection suggests that new concepts are introduced both in the course of individual lives and across human history. That individuals acquire concepts as they mature from infancy seems beyond doubt. Even if one holds that there is some set of basic, perhaps innate, concepts that all humans share, it seems clear that people are not born with full mastery of such concepts as boson, isotope, fuel injector, split infinitive, corn futures, standard deviation, transcendental argument, coming-out party, royal flush, or balk. These concepts and many others are acquired in the course of a life. Moreover, these examples include concepts that are not learned by all people, and that are not found in all contemporary cultures or in all historical periods of our own culture. As already indicated, this study will focus on those who are sufficiently mature to have acquired a native language and a body concepts that is rich enough to deal with the objects and situations they encounter in the normal course of their lives. But even adults enter into situations in which they acquire new concepts, for example, as they learn a vocation, adopt an avocation, pursue a wider education, or encounter people from different cultures and sub-cultures. In a society of any complexity there will be considerable variation in the conceptual repertoires of various people. Those in a particular profession – say, electricians, arbitragers, sculptors, neurosurgeons, or astrophysicists – will have specialized bodies of concepts for dealing with objects, situations, materials, tools, and processes they encounter in their professional activities. In a similar way, those interested in opera, stamp collecting, antiques, horse racing, and so forth will also acquire specialized concepts that are not universally shared. Since human beings are

social creatures, a significant part of our conceptual repertoires will be concerned with social arrangements and practices. Examples include capitalism, freshman, citizen, legislature, secretary of state,[2] prime minister, commissar, civil right, and eminent domain. Which of these concepts each of us acquires depends on the society we live in, the depth of our understanding of that society, and the scope of our education concerning other societies.

The introduction of new concepts is especially striking as we run our gaze over the course of human history. From an historical perspective we encounter myriad examples of concepts that are not part of contemporary thought and that will be familiar only to those who have studied the relevant history. Examples include phlogiston, telegony, radioactive induction, N ray, vassal, and the god of war. Different fields of human endeavor have different developmental histories. Some fields have a history that goes back well before we have any clear records, but some appeared within historical time and have a documentable history in which new concepts were introduced by creative individuals and passed along to their successors. Often new concepts were introduced as part of an attempt to solve outstanding problems, and when we look at the contemporary world we can reasonably project that the resolution of some currently recalcitrant problems will require ways of thinking that are not yet available.

Typical adults living in a society have a body of concepts and beliefs that allow them to deal fairly successfully with the common situations they are liable to encounter. The exact relation between concepts and beliefs is one of the topics to be explored in this book, but we should be able to agree that beliefs about a particular topic require concepts for thinking about that topic. Many of our concepts concern items we can detect with the senses we evolved on the surface of this planet, senses that allow us to pick out objects, properties of objects, and processes that occur in the environments in which humans have lived for most of our history.[3] But people also introduce concepts for items that are not available to normal perception. Common examples include deities, spirits, angels, and worlds beyond the range of common experience. The development of science led to the massive postulation of items that cannot be detected by unaided perception as it became strikingly clear that the world is full of such items. These include X rays, bacteria, specific toxins (e.g., in mushrooms or the soil on which a housing sub-division was constructed), genes, and electrons, among others. Every such postulation involves the introduction of a concept, and the fact that I can direct the thoughts of many readers to these items just by using a word or phrase is powerful evidence that we share the relevant concepts. The means by which such concepts are introduced, and the ways in which adults can learn them, are among the topics to be addressed by a theory of concepts.

To be sure, not everybody associates a concept with every expression I have used. For each of us there are subjects about which we lack concepts

and thus have no beliefs at all. It is easiest to illustrate this point by contrasting earlier people with ourselves, although the point applies to us as well. Consider just a few examples of subjects on which many of us have beliefs that could not be formulated using the concepts available to an ancient African or Athenian or Australian: the use of radiation to sterilize food, the amount of RAM needed to run Windows XP efficiently, the pitfalls of investing in complex derivatives, the imbalance between matter and anti-matter in the universe, the difference between ordinary and partial differential equations, the constitutionality of using sampling techniques in a national census, and the significance of solar neutrino experiments for the question of whether neutrinos have mass. These examples all derive from modern western society, but it is more than likely that people living in non-western societies have concepts that I cannot presently describe. Someone who is capable of surviving without modern technology in the African or Australian bush, or in the Arctic, has a great deal of knowledge that I lack, and this knowledge may well involve concepts that I do not possess.

Some of the concepts I have mentioned in the course of these introductory remarks have no corresponding instances in the world; as philosophers are wont to say, they are not *instantiated*. However, an uninstantiated concept may still be a genuine concept. It will be a persistent theme of this book that we must distinguish an account of the content of a concept from an assessment of whether it has instances. Indeed, any attempt to show that a concept lacks instances requires a grasp of the content of that concept. At the same time, the fact that some group has a well-developed practice of using and teaching a particular concept does not guarantee that this concept has instances. While both of these points strike me as obvious, there are important philosophical theories of concepts that challenge these claims; I will consider such theories as we proceed, especially in Chs 3 and 6.

I have been illustrating the enormous range of conceptual variation among people within a society, in various parts of the world at a given time, and in the course of human history. It is an immediate corollary that conceptual change occurs as people learn – both in the course of history and in the course of an individual life. Before proceeding I want to emphasize that I am using the phrase *conceptual change* to cover any change in a conceptual repertoire; the expression is intended to be neutral on the question whether such change always involves replacement of one concept by another, or if there is a significant sense in which concepts can themselves be altered. Now, one major task – and test – for a theory of concepts is to provide a basis for understanding how conceptual repertoires change. Two problems must be addressed in considering this topic. One of these is a psychological problem: it concerns the cognitive means by which individuals invent and acquire new concepts. Since I will discuss only those who already have a substantial conceptual repertoire, one approach to this question is to show how new concepts can be constructed out of previously available concepts. How this occurs will depend on the details of a theory of conceptual content. For

example, some theories of concepts postulate a set of basic concepts that is largely shared by human beings. New concepts are introduced by constructing them out of subsets of these basic concepts; people learn the new concepts by following out this construction. Other theories of concepts that we will encounter reject the existence of such basic concepts, but still hold that new concepts are constructed out of previously existing concepts. Advocates of these different theories will give different accounts of this construction process, and of how newly introduced concepts are learned.

The second problem arises because we can also consider concepts as abstract structures, apart from their embodiment in individuals. (I will return to this topic in Sec. 1.5.) We adopt this perspective, for example, when we compare the content of concepts in order to clarify ways in which they are the same, and ways in which they differ. Questions of this sort typically arise in situations where we have competing concepts for dealing with the same subject matter; the concepts of space and time found in classical physics and in relativity theory provide a much-discussed example. How we carry out this comparison – and whether such a comparison can be carried out at all – depends on our view of conceptual content.

## 1.3 Conceptual Analysis

Conceptual analysis is a major philosophical industry, especially in the twentieth century English-speaking world where many hold it to be the only legitimate philosophical endeavor. Whatever one's view on this strong claim, conceptual analysis is an important philosophical concern, and is important in other fields as well. For example, those who study the conceptual development of a science must engage in conceptual analysis in order to compare the content of concepts at various points in time. Those who seek to understand the thinking of people from other cultures must also carry out conceptual analyses as part of their research. In addition, those who propose a conceptual innovation must engage in analyses of the existing concepts and of the new concepts they seek to introduce. But any attempt to carry out a conceptual analysis requires a theory of how conceptual content is determined. Without such a theory we have no way of deciding what counts as an analysis and no way of judging whether a proposed analysis is adequate. Competing theories of conceptual content often give different answers to these questions. I want to mention some preliminary examples, subject to more detailed discussion in later chapters.

One issue is the relation between concepts and propositions. A common view is that concepts are fundamental and that propositions are built out of concepts. The general point can be seen with particular clarity if we look at the analogous relation between words and sentences. The common view holds that words have meaning independently of the sentences in which they occur, and that the meaning of a sentence is determined by the meanings of its words plus the grammatical rules of the language. A contrasting view

holds that sentences are the fundamental bearers of meaning, and that words acquire meaning from the roles they play in various sentences. The verification theory of meaning championed by logical positivists is an example of the latter view since it is propositions that are verified or falsified. This analogy between words and concepts raises further questions about the relation between language and concepts; I will postpone this topic until the next section.

A doctrine of the classical empiricists will introduce another point of disagreement. These philosophers drew a sharp distinction between *simple ideas* and *complex ideas*. Simple ideas are acquired directly from experience, cannot be broken down into simpler components, and provide the material for all of our thinking. Complex ideas are built up, in various ways, out of simple ideas. Only complex ideas are subject to analysis, and the analysis of a complex idea consists of resolving it into its component simple ideas. An alternative view, found for example in C. I. Lewis (1946, 1956), rejects any distinction between simple and complex concepts. In Lewis' view conceptual content is constituted out of relations to other concepts. Conceptual analysis requires mapping out the relations between concepts, not their dissolution into simpler parts, and all concepts are equally subject to analysis.

Another debate turns on whether concepts are structured by necessary and sufficient conditions or have some form of "open texture." The vast majority of those who practice conceptual analysis assume the necessary-and-sufficient-conditions view, which provides one set of criteria for a successful analysis. Analyses are typically presented in the form "X is C if and only if . . . "; critics of a particular analysis can challenge either the necessity or the sufficiency of the conditions stated. This view was challenged by Wittgenstein (1953), and more recently by work in psychology where it is argued that people often behave in ways that are not compatible with the view that concepts are constituted by necessary and sufficient conditions. For example, respondents are quite clear that a robin is a better example of a bird than a turkey, but a necessary-and-sufficient-conditions view has no room for such considerations of degree; an item either falls under a concept or it does not. Those who hold that concepts are open textured require a different account of the aims of conceptual analysis than is currently typical among philosophers.[4]

A further issue concerning the aims of conceptual analysis arises when we recognize that the concepts we currently use for thinking about a subject may not be adequate. One approach holds that the aim of analysis is to describe concepts as we find them. The significance of this view is illustrated by Sen's discussion of economic inequality. Sen takes it for granted that we already have the relevant concept of inequality in mind, and that his task is to provide an appropriate measure of this inequality. Responding to the proposal that we should be able to give a complete ordering of levels of inequality, Sen writes: "It is, however, possible to argue that the implicit notion of inequality that we carry in our mind is, in fact, much less precise

and may correspond to an incomplete quasi-ordering" (1997: 5–6). A bit further down the page he adds,

> There are reasons to believe that our idea of equality as a ranking relation may indeed be inherently incomplete. If so, to find a measure of inequality that involves a complete ordering may produce artificial problems, because *a measure can hardly be more precise than the concept it represents* [italics added].

In a later review of Sen's text, Foster and Sen write:

> If a concept has some basic ambiguity (as ideas of what constitutes 'inequality' tend to have), then a *precise* representation of that ambiguous concept must *preserve* that ambiguity, rather than try to remove it through some arbitrarily completed ordering. This is quite central to the need for *descriptive accuracy* in inequality assessment, which has to be distinguished from fully ranked, unambiguous assertions (irrespective of the ambiguities in the underlying concept).
>
> (1997: 121)

Many analytic philosophers will agree with the view that an analysis of an imprecise concept should share that imprecision.[5]

A rather different view is found in Carnap's classic account of explication. For Carnap an explication does not just provide an explicit formulation of an available concept. Instead: "The task of *explication* consists in transforming a given more or less inexact concept into an exact one or, rather, in replacing the first by the second" (1950: 3). Development of this replacement concept involves a tradeoff among four criteria: similarity to the concept being explicated, precision, fruitfulness in the sense of being useful for the formulation of universal statements, and simplicity. With regard to the first of these criteria Carnap writes,

> The explicatum is to be *similar to the explicandum* in such a way that, in most cases in which the explicandum has so far been used, the explicatum can be used; however, *close similarity is not required, and considerable differences are permitted* [italics added].
>
> (1950: 7)

Thus, for Carnap, philosophical reflection on concepts aims at improving our conceptual situation, not just at describing it with all of its current imperfections. Such improvement is one of the forms of conceptual innovation that we will explore as we proceed.[6]

Sometimes the need for conceptual improvement can be extremely compelling. Russell's discovery of an inconsistency in the concept of a set used by Cantor and Frege provides a classic case in which the inadequacy of

a concept was discovered by pure reflection. Other cases occur when new information undermines conceptual boundaries we have drawn. In biology, for example, the European discovery of Australian monotremes undermined the prevailing concept of a mammal because monotremes mix together features that were considered characteristic of mammals with other features considered characteristic of birds and reptiles; we will encounter many similar examples in Ch. 2. For the moment I want to emphasize that we should expect challenges to existing concepts as long as we recognize that there is a great deal about the universe that we do not know, as well as the fallibility of our present beliefs. A theory of concepts should provide some insight into how innovations are produced, as well as a guide to analyzing available concepts.

## 1.4 Concepts and Language I

Many philosophers identify concepts with linguistic entities so that "conceptual analysis" and "linguistic analysis" are two names for a single enterprise. This view is particularly prevalent in the twentieth century, but has been under discussion at least since Plato considered the hypothesis that thinking is talking to oneself (*Sophist* 263E, *Theaetetus* 189E). There are important reasons for this practice. Many hold that thought takes place in language, so the study of cognition is encapsulated in the study of language. In addition, language is a public phenomenon that seems more easily accessible for study than concepts viewed as mental entities. Others, however, draw a sharp distinction between concepts and language. During the seventeenth and eighteenth centuries both empiricists and rationalists held that ideas are the medium of thought and language is a superstructure used for communication and as an aid to memory. Empiricists proposed theories of language, but held that the meanings of words *are* the ideas with which they are associated. Misuse of language was treated as a major source of errors, and one reason for discussing language was to learn how to avoid these errors. Rationalists devoted considerably less attention to language. For those who do not treat concepts as linguistic entities, the study of language may be a source of evidence about concepts, but will not be the entire story. I will not identify concepts with linguistic entities in this book; I have several reasons for this decision.

First, whether non-linguistic animals have concepts is an important question that we should not attempt to settle by fiat. A better approach is to develop a theory of human concepts and then consider whether the relevant evidence supports attribution of concepts to other species. I will discuss some of the literature on animal cognition in Ch. 5, but only in order to clarify what is involved in attributing concepts; I will not take a stand on whether other animals have concepts. For the most part I will be concerned with human concepts, and with conceptual analysis and conceptual innovation – activities which, as far as we know, only humans pursue. In this case

the standard practice of using linguistic information as one source of evidence about underlying concepts is appropriate and I will adopt it.

Second, the case studies in Ch. 2 indicate that linguistic change generally lags behind conceptual change. This provides a positive reason for distinguishing studies of language from studies of concepts. Moreover, the attempt to construct an independent theory of concepts may provide considerable insight into how the two relate. It is even possible that such an attempt might fail, and thereby support the view that languages and concepts are intimately connected.

Third, treating concepts as linguistic, and conceptual systems as languages, may be useful at many points, but when pushed too far this practice encourages us to lose sight of two important issues. First, it is clear that multiple, even competing, conceptual systems are expressible in a single natural language; this point is obscured when we treat a natural language as a single conceptual system.[7] In addition, whether different natural languages are capable of expressing the same concepts is an open question that can be approached most clearly if we have an account of concepts that does not presuppose a particular relation between concepts and language.

Fourth, a brief look at one recent development in the theory of meaning will provide an additional reason for separating conceptual matters from linguistic matters. When Putnam argues that meanings are not (solely) psychological entities he is careful to distinguish concepts from meanings (1975: 217–19, 226–27, 245, 248). Putnam argues that a significant part of the meaning of any term that refers to a natural kind is in the world, and that we learn the meaning of the term through scientific research. I have no intention of endorsing this view, but I want to note one of its consequences for understanding scientific research. As we study some presumed natural kind we need to think about it in order to formulate questions and hypotheses, and develop means of testing those hypotheses. In other words, we need some *mental representation* of that kind. This representation is our *concept* of that item, and one aim of empirical research is to improve the accuracy of this concept.[8] Thus, even given the theory of meaning that Putnam defends, we still need to introduce concepts, understood as mental entities, to make sense of the research process that leads to an understanding of the meanings of our terms. Nor need we adopt Putnam's account of linguistic meaning in order to recognize that one role of empirical research is to formulate descriptions of items in the world, and to improve these descriptions as research develops. As this process proceeds we seek to improve the conceptual repertoire we use to think about these items. Thus an understanding of the nature of concepts, and of the ways in which a conceptual repertoire is altered, are central to any epistemology that acknowledges a role for ongoing research in the development of human knowledge. Note also that the methodological difficulties involved in studying mental entities are no greater than those involved in studying other items that are not easily detected by casual observation. The history of

science has surely taught us that the ease with which an item can be studied is not a reliable indicator of its theoretical importance.[9]

I will not adopt any general position on linguistic meaning in this book, and I will not pursue the vagaries of the theory of meaning except when discussing thinkers whose theories of meaning and of concepts are inseparable. Towards this end, *when discussing my own views* I will reserve the term "meaning" for cases in which I am explicitly discussing linguistic items, and I will talk about the *content* of concepts as I have already been doing. (Cf. Harman 1982: 243–44, 1999: 208 for similar terminology.) I will henceforth follow the common practice of putting names of linguistic items in quotation marks, and I will generally depend on context to make it clear when I am discussing a concept and when I am discussing some item that is neither linguistic nor a concept. When context is not sufficient – and sometimes for emphasis – I use small capital letters for terms that refer to concepts (e.g., CONCEPT). Still, given the widespread identification of concepts with words, I will not always be able to follow this practice when discussing the views of others. In those cases I will usually adopt the practice of the philosopher under discussion.

## 1.5 Biology, Psychology, and Abstract Descriptions

In order to understand the specific project that I am undertaking we must distinguish three perspectives from which we can study concepts: biological, psychological, and abstract. From a *psychological perspective* concepts are mental entities that exist in individual minds, but without any concern about how these concepts are implemented in a neural system (I include the brain under this rubric). Working from this perspective psychologists examine the role that concepts play in individual thought and in various forms of human behavior. Studies of this sort can be carried out in many ways, including experiments on subjects in the psychologists' laboratory, and studies of the thought of people from various historical periods and societies even though those individuals are not currently available. All human cognitive activities are products of human psychology and provide evidence about the nature of human minds that may be relevant to the psychological study of concepts.

We move to a *biological perspective* when we examine the physical embodiment of concepts in organisms. Whatever else we may say about an individual's concepts, they must have some neural embodiment if they exist at all. Thus an account of the neural basis for concepts is a necessary component of a complete account. Psychological and biological research on concepts are deeply interrelated. One relation arises because psychological studies provide data that must be accounted for by an adequate biological account, but psychological studies of concepts are not limited to accumulating data for biologists. There are theories of concepts that are developed in psychological terms without any concern for their physiological embodiment; examples include the "idea" theories of classical empiricists and

contemporary "language of thought" theories. Still, a correct psychological theory must be implementable in human biology, so psychological theorizing is ultimately constrained by our biology. As we learn more biology we may find that a theory that accounts for a wide variety of human behavior in psychological terms must be modified or rejected. In addition, growing understanding of cognitive neurobiology may point to new directions for psychological research. In general, the relation between biological and psychological approaches is one of mutual fertilization and mutual constraint.

An *abstract perspective*, and its distinction from a psychological perspective, can be introduced by considering two different ways of thinking about logic. One tradition holds that logic investigates "the fundamental laws of those operations of the mind by which reasoning is performed . . . " (Boole 1958: 1); this is a psychological perspective. It can be given a normative turn as the study of the laws to which thought ought to conform, but is still a psychological approach as long as it is concerned with actual thinking. Following Frege, logicians now generally accept the alternative view that logic studies relationships between propositions independently of their embodiment in actual thought; this is an abstract perspective. Logic is still viewed as providing norms, but these norms are applied to products of thinking, not to processes by which these products were produced.

We can get some further clarification by considering an analogy: the relation between a computer program and what physically occurs in the computer. Although physical events in a computer are analogous to what occurs in our biology (a psychological approach is not relevant in this case), reflection on computers will help clarify the notion of an abstract perspective. (Colburn 1999 provides a useful discussion.) From a physical perspective any inputs we provide – whether a program or a body of data – are present in the computer as electronic states, such as a set of charges on capacitors. Processes that take place in the computer are physical processes involving these states. In the earliest computers programs and data were entered by connecting wires or flipping switches. In later computers, such as those that most of us use, the keyboard provides a more convenient way of accomplishing the same end. Pressing a key closes a switch that sends an electric current to a specific unit in the computer. That unit generates other currents that change the charge states of various elements in the machine. An electrical engineer will be interested in a description of these charge states; programmers do not work with a description in these terms.

Descriptions in terms of programs and data are abstract descriptions. They describe what is going on in the machine, but do so using different concepts than those required for a physical account and leave out many details of the machine's workings – with the result that this description applies equally well to a variety of machines that implement it in different ways. Consider the two descriptions in the case of a simple program that could be written in a high-level language such as Fortran or Basic. For the

program to run, all the operations it specifies must be replaced by operations that are hard-wired into the machine. This is done by other programs, such as compilers and assemblers that have already been implemented in the computer; they take the program as input and generate electronic states that the computer can process. The programming language allows us considerable freedom from any concern with hardware details, but this is possible only because these details have already been taken care of by those who designed and implemented the language. The details of the implementation may be different in different computers, but programmers can ignore such matters. Still, programming languages have to be designed in a way that allows programs to be implemented in actual machines, which is why we can say that the program describes what is occurring in the machine.

A program is a description of the processes going on in the machine that allows us to *abstract from* – that is, to ignore – some aspects of the task we are engaged in and to focus our attention on other aspects. Abstract descriptions typically use different concepts than machine-level descriptions and allow us to study features of a program – such as its logical structure – that might not be apparent from a description of a sequence of electronic states. Abstraction is a matter of degree: different descriptions of a process may be *more or less abstract* – where a more abstract description includes less detail about what occurs in the machine. A program written in Fortran or Basic is more abstract than a program that carries out the same task but is written in the assembly language for a particular machine. A flow chart that gives just the logical structure of a program is more abstract than a program written in a specific programming language. Translating the flow chart into an implemented language requires adding considerable detail.

Note that I have been discussing abstract *descriptions*, not abstract *entities* existing in some non-physical world. Every abstract description will be embodied in some objects in the world in which we live (e.g., in a brain), but we ignore this embodiment when working from an abstract perspective.[10] Consider one more example: the distinction between *sentences* and *propositions*. I will treat propositions as abstract descriptions of sentences. Propositions are physically embodied in sentences, but there are cases – such as logical studies – where such things as the color of the ink or the particular language in which the sentence is formulated are irrelevant. Thus I was writing from an abstract perspective when I described logic as dealing with relations between propositions. In general, studies from an abstract perspective have two key characteristics: they ignore properties of their subject matter that would be included in a biological or psychological perspective, and they may use concepts that would not appear when working from one of the other perspectives.[11]

In this book I will study concepts primarily from an abstract perspective. This will allow us to discuss such topics as implicational relations among concepts, the consistency of a conceptual system, and logical consequences of a conceptual system independently of whether anyone has noticed them.

This perspective will become especially salient in Ch. 4. We will also encounter situations (beginning in Sec. 5.8) in which we must take thought processes into account; I leave further discussion of this topic until it is needed. I have nothing to say here about neurobiology or about the ultimate relation between biology and psychology. Still, the theory of concepts I propose is anchored in human biology and psychology because my data come from actual cases of our cognitive history. In Ch. 2 I establish an initial database for the development of the theory; I further test the theory by considering new examples in Chs 9 and 10. The large number of cases I consider provides an important reason for believing that the theory I propose captures actual features of human thought. In addition, I will be concerned throughout with the question of how we can introduce and learn new concepts while maintaining the continuity required for intelligibility. This is a constraint on theorizing that is imposed by human psychology. The account I give is thus subject to empirical evaluation.

Peacocke (1992) advocates a rather different view of the relation between philosophical and psychological studies of concepts.[12] Peacocke rejects the view, held by many philosophers, that philosophical and psychological studies of concepts are utterly disjoint activities, so that practitioners of the two types of study need have no professional interest in each others' work. Instead, Peacocke holds that there is a one-way relation between the two fields. It is up to philosophy to provide the possession condition for a concept; once this condition has been specified, it is the task of psychology to determine how this concept is implemented in the individual: "When a thinker possesses a particular concept, an adequate psychology should explain why the thinker meets the concept's possession condition" (177). Moreover, "For any particular concept, the task for the psychologist is not fully formulated until the philosopher has supplied an adequate possession condition for it" (190). Indeed, the relation can go only in this direction because a philosophical study of a concept proceeds by a priori, not by empirical, methods (179). Peacocke acknowledges that a priori methods are fallible, but challenges to conclusions arrived at by a priori methods can come only from other a priori considerations; no empirical study can ever challenge the results of an a priori analysis. Since I reject this priority thesis, I want to offer a preliminary account of my reasons for doing so.

Consider an analogy that Peacocke uses to defuse the apparent arrogance of his approach. The view that correct philosophical analyses of concepts provide one-way constraints on psychological studies of concepts is, he tells us, "no more objectionable than the principle that a good micro theory of gasses should explain the macro truth that pressure increases with temperature for a given volume" (179). However, while the macro law holds for a significant range of pressures, temperatures, and volumes (and a particular degree of instrumental precision), it is not true in general. The microtheory of gases explains why this law fails, and yields a more accurate replacement for this law. In general, microphysical theories do not just explain established macroscopic

laws; situations are common in which lead us to revise macroscopic laws. Sellars provides a description of the general situation: microtheories

> explain empirical laws by explaining why observable things obey to the extent that they do, these empirical laws. . . . Furthermore, theories not only explain why observable things obey certain laws, they also explain why in certain respects their behaviour obeys no inductively confirmable generalization in the observation framework.
>
> (LT 121)

Peacocke's analogy, then, suggests that psychological studies of concepts may challenge philosophical analyses. One example is provided by the evidence (mentioned above) which suggests that possessing a concept should not be identified with possessing a set of necessary and sufficient conditions for instances of that concept. I will have much to say about the nature and purpose of conceptual analysis in Chs 7 and 8. For now I adopt the working hypothesis that abstract and psychological studies of concepts constrain and fertilize each other in the same ways that psychological and biological studies do.

## 1.6 Naturalism

It should be clear that my approach to concepts is thoroughly naturalistic. The central idea of naturalism is that humans – including human cognitive abilities – are part the natural world. (Giere 2000 provides a recent summary of the naturalist position.) This is a major departure from the view of human thought in much of our cultural history, including much earlier epistemology. The thesis that our minds are not part of nature is central to, among others, the epistemologies of Plato, Descartes, Locke, Berkeley, and Kant. It can be argued that some pre-twentieth century philosophers were naturalists – Hume is an especially attractive example. But it is really in the twentieth century, under the influence of the theory of evolution, that we came to fully conceive of ourselves as part of nature, and in the last half of that century that we began to explore how this understanding should affect epistemology.

The key thesis of naturalism for epistemology is that we must study human knowledge in the same ways that we study other domains: by examining evidence, and formulating and testing hypotheses. A major aim of such study is to learn about human cognitive abilities, for an account of human knowledge requires that we understand the nature, scope, and limits of these abilities. This point applies even to the development of a normative epistemology, for the goal of such an epistemology is to develop norms that are appropriate for human knowers. Ignoring human limitations we could easily put forward *be omniscient* as the central epistemological norm. But omniscience is not within our capabilities, so we face the double task of

discovering appropriate methods for evaluating knowledge claims, and understanding the limits of those methods. It is only within this framework that we can formulate epistemic norms that are relevant for us. Descartes clearly recognized this point; it is our epistemic limitations that lead to the quest for reliable means of acquiring knowledge. However, Descartes considered the key issues to be metaphysical and his attempt to establish the appropriate methods proceeded a priori. Yet the empirical study of human belief systems indicates pretty clearly that we have no a priori insight into any features of the world – including our own epistemic abilities.

The thesis that we do not have a priori knowledge of the world is a familiar theme of philosophical empiricism. Historically, empiricists have been much more cautious in attributing cognitive abilities to human beings than have philosophers in some other traditions. Hooker (1987: 74) offers the following summary of the twentieth century empiricist view of human cognition: "Man is a sensory experience reception chamber together with a generalized logic machine," where *generalized logic* includes "the theory of truth functions . . . first order predicate calculus and systems of what are called inductive logic, these days we may consider also N-order predicate calculi, various forms of modal logics, many-valued logics . . . " (1987: 71).[13] The key feature of logic, whatever its detailed scope, is that it deals only with formal relations, abstracting from experience; the content of knowledge comes from experience. In my view the caution characteristic of the empiricist tradition is appropriate, but empiricists were too cautious in their account of our cognitive capabilities. (For discussion see Brown 1978, 1988, 1994b, 2000c.) I will not offer a comprehensive theory of our cognitive capacities in this book, but we will encounter reasons for admitting a somewhat greater range of human cognitive abilities than has been typical in the empiricist tradition. Throughout the discussion I will adopt the older empiricist, and contemporary naturalist, view that an account of the nature of conceptual content is an empirical theory.[14] In doing so I will consider a much larger variety of concepts than is typical in the philosophical literature. There is also a theory of concepts that is characteristic of philosophical empiricism: our ability to form concepts is limited by our perceptual and introspective experience, which provides all the content for our concepts. While seventeenth and eighteenth century empiricists treated this thesis as an empirical claim, many twentieth century empiricists adopted this view while denying its empirical status. I will examine both versions of this theory in Ch. 3.

## 1.7 Incommensurability and Relativism

Discussions of conceptual change lead directly to the incommensurability thesis and then on to concerns about relativism. The incommensurability rubric was introduced by Kuhn and Feyerabend in 1962; claims associated with this rubric have been subjected to many interpretations, refutations, and defenses. (For recent discussions and an extensive bibliography see

Hoyningen-Huene and Sankey 2001.) It is important that we be clear on the historical context in which the notion arose since ideas from that context remain central to these debates. The dominant view of concepts in philosophy of science – developed in terms of a theory of meaning – was a version of the empiricist view: we have a basic vocabulary made up of terms that derive their meaning directly from experience. This vocabulary, dubbed the "observation language," provides the meaning of all other terms – although exactly how this occurs was subject to debate. All human beings with normal sense organs can share this observation language (variations are a matter of the particular experiences one has had), and the meanings of its terms are established independently of any of our beliefs – and *a fortiori* independently of any theories we may hold. Different natural languages associate different phonemes and graphemes with experienced items, but terms that are associated with qualitatively identical bits of experience have the same meaning. Thus the terms of the observation vocabulary are precisely translatable among all languages. The non-observation terms of a language are strictly auxiliary; they are introduced for convenience and can be eliminated. All cognitively meaningful discourse can be expressed in the observation language. As a result, if two theories compete we can state their points of disagreement in the observation language and see precisely what evidence would decide between them. Thus the observation language plays a double role: in addition to providing the source of all linguistic meaning, it also provides a medium for comparing competing theories.

Kuhn and Feyerabend proposed a different view of the relations between theories and empirical evidence. They denied the existence of a theory-independent observation language. Instead, they argued, a theoretical language gets its meaning from the internal structure of the theory, independently of any association with experience. Meaning then flows from theory to observation, not in the reverse direction. A theory provides a language in terms of which sensory experience is reported and understood, and any empirical evidence that is relevant to the evaluation of a theory must be expressed in the language of that theory. An immediate consequence would seem to be that no single body of evidence can provide an independent basis for comparing competing fundamental theories since the evidence that is relevant to each of these theories is already laden with the language of that theory.[15] If this view is correct it undermines the account of theory comparison that was standard in the late 1950s. Many thinkers move directly from this rejection of empiricist theories of meaning and evidence to an epistemic relativism. Rejecting the doctrine of an observation language, they conclude that there is no neutral medium in which to carry out an objective comparison of competing theories, and thus that there are no epistemically compelling grounds for preferring one theory over another. The preferences that individuals and groups do have, they conclude, are based on a variety of personal and social factors that have nothing to do with any form of epistemic superiority. It is, however, far from clear that objective theory

comparison *requires* that competing theories be expressed in a common language.[16] More generally, it is far from clear that rejection of the empiricist account of theory comparison eliminates all means of objective comparison. I will postpone further discussion of this issue until the final chapter of this book. For the moment I want to emphasize that how we deal with this question will depend on our accounts of conceptual content and of the cognitive abilities that we bring to bear in evaluating theories.

There is one more issue that is easily solved, at least in principle, by empiricist accounts of meaning, but provides a major challenge on the alternative proposed by Feyerabend and Kuhn: How do we learn the language of a new theory? On the empiricist approach the cognitive content of all auxiliary terms can be formulated in the observation language, and we can make the transition from one language to another by means of such formulations. The alternative we are considering blocks this route and leads to such claims as that we must learn a new system of concepts as a whole, and to Kuhn's famous gestalt-shift metaphor. However, our account of how adults learn new language – or, returning to my preferred idiom, new concepts – depends, again, on our account of human cognitive abilities and our theory of conceptual content. I will speak to these issues as I develop my theory of concepts.

# 2 Conceptual Journeys

> After all, it is characteristic of modern science to produce deliberately mutant conceptual structures with which to challenge the world.
>
> (IM 337)

Contemporary interest in conceptual change developed out work in history-based philosophy of science, particularly the work of Kuhn and Feyerabend, and much of the philosophical literature has focused on the examples they used. These are important examples; they include the development of Copernican astronomy and Newtonian mechanics along with the replacement of their Aristotelian, Ptolemaic, and Brahean predecessors; the introduction of relativity and quantum mechanics, and their contested conceptual relations to classical mechanics; the advent of Lavoisier's chemistry which superseded the phlogiston theory; and a few others. I will consider these cases in the course of this book, but the continued focus on just these examples leaves the impression that conceptual change is an isolated phenomenon in human cognitive history and in the course of an individual life. In this chapter I will endeavor to broaden the base of our discussion by describing a number of additional cases of conceptual innovation in science, mathematics, technology, society at large, and philosophy. I have two aims in these discussions: to underline the pervasive role of conceptual innovation in the development of knowledge and to provide a working database that any theory of concepts must address. Further examples will be introduced throughout this book.

Some philosophers will object that many of the issues I discuss in this chapter do not involve *conceptual* change, but rather change of belief in which the concepts involved remain constant. However, one main thesis of this book is that what counts as a conceptual change depends on the theory of concepts we adopt, so I urge the reader to withhold judgment on this topic. In coming chapters we will examine several different accounts of conceptual content and, as a result, of what counts as conceptual change.

## 2.1 Physical Science[1]

> From time to time in the sciences a new field may open which shortly before had been inconceivable. An unnoticed phenomenon comes to attention, a novel concept is formulated, and what had previously been a matter for speculation is brought within the range of experimental research.
>
> (Romer in Rb 3)

I will begin by examining some of the changes involved in the journey from the ancient Greek concepts of atoms and elements to the modern version. Greek atomism holds that material objects are ultimately made up of tiny indivisible particles characterized by their size and shape. The doctrine of elements that was dominant after Aristotle holds that all items in the universe are constituted out of earth, water, air, and fire in the terrestrial realm, and ether in the celestial realm.[2] Although the two views are not mutually exclusive, Aristotle rejected atomism; I will begin with a non-atomistic account. The distinction between a terrestrial and celestial realm, made up of different materials and following different laws, was central to the Aristotelian cosmology that remained dominant until the seventeenth century. Given this distinction, no one ever had a sample of ether in hand; once the fundamental division of the world into two realms was rejected the original notion of ether also vanished – although the word reappears throughout the history of science.[3] Let us consider the four terrestrial elements.

Two distinct lines of thought converge in the ancient notion of an element; we can view these, perhaps anachronistically, as coming from physics and chemistry. The physical notion of an element comes from Aristotle's dynamics. Aristotle held that terrestrial space is organized into a set of *natural places*, one associated with each element. An unconstrained sample of an element moves spontaneously to its natural place: earth to the center of the universe, water above earth, air above water, and fire to the sphere of the moon. These are *natural motions*, one associated with each element. From the perspective of these motions, air and fire share the property of being light; earth and water are both heavy. Thus the elements "fall into two pairs which belong to the two regions, each to each; for Fire and Air are forms of the body moving towards the limit, while Earth and Water are forms of the body which moves towards the centre" (1995b, 330b: 541). This account explains several familiar features of the world: the pattern in which we find water above earth and air above water, why stones and water fall while fire rises, and the spherical shape of the earth. For the most part I will postpone further discussion of Aristotelian dynamics until Ch. 9, but I want to raise one question here: Why haven't the elements separated from each other, with each element having settled long ago into its natural place? Aristotle's reply is that this separation does not occur because the elements

are continually changing into each other (1995b, 337a: 552). To understand this reply we must consider Aristotle's chemistry.

From a chemical perspective elements are natural kinds that cannot be resolved into different, more basic, constituents. For example, if you attempt to resolve pure air into its constituents all you will find is more air. At no point will you reach some different, more elementary, constituents out of which air is constructed – and similarly for the other elements. Actual samples that we might study are blends of all the elements (1995b, 334b–335a: 547–48), but distinguishing these elements is as far as we can go in resolving a sample into basic kinds. In addition, there are four fundamental qualities divided into two pairs of polar opposites: hot and cold, dry and wet. Each element is characterized by two qualities, one from each pair, imposed on an underlying matter (see Figure 2.1). As a result, elements can be transformed into each other by changing one quality at a time while the substratum endures. Note that there is no commitment to atomism here – no suggestion that division will end at a minimal unit of an element.

It is way beyond the scope of this book to attempt a detailed history of all the conceptual transformations that took place on the route to twenty-first century chemistry, but I want to note several key steps and then consider one set of changes in some detail. Doubts about the status of fire as an element – that is, about whether fire should be classified with air, earth, and water for chemical purposes – arose quite early, although they did not prevail.

> Despite persistent criticisms by Theophrastus (371–286 BC), Aristotle's pupil and successor at the Lyceum, that fire was different from the other elements in being able to generate itself and in needing other matter to sustain it, the theory of the four elements was to remain the fundamental basis of theoretical chemistry until the eighteenth century.
>
> (B 12–13)

The theory was invoked to explain observed transformations. For example, residues left behind when water was evaporated were taken as evidence of the transformation of water into earth, while the evaporation of water seemed a

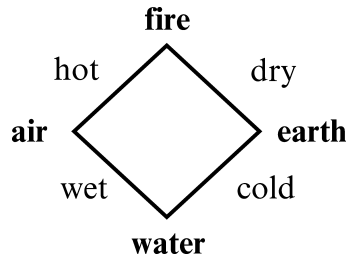


*Figure 2.1* Elements and Qualities

clear case of the transformation of water into air. The following remarks from Newton indicate some features of the prevailing framework. After noting that "Water by frequent Distillations changes into fix'd Earth, as Mr. *Boyle* has try'd . . . " (1952: 374), Newton adds:

> Nature . . . seems delighted with Transmutations. Water, which is a very fluid tasteless Salt, she changes by Heat into Vapour, which is a sort of Air, and by Cold into Ice, which is a hard, pellucid, brittle, fusible Stone; and this Stone returns into Water by Heat, and Vapour returns into Water by Cold. Earth by Heat becomes Fire, and by Cold returns into Earth. Dense Bodies by Fermentation rarify into several sorts of Air, and this Air by Fermentation, and sometimes without it, returns into dense Bodies.
>
> (1952: 374–75)

Still, the exact significance of these transformations was subject to debate. For example, in 1746 Eller argued:

> that water could be changed into both earth and air by the action of fire or phlogiston. For Eller this was evidence that there were only two elements, fire and water. The active element of fire acted on passive water to produce all other substances.[4]
>
> (B 96)

Other developments led to deeper challenges. As the existence of a large number of different solid materials became clear, doubts developed about treating earth as an element. In the case of the transformation of water to earth, "by the 1760s most chemists could no longer credit that such an apparently simple pure substance as water could be transmuted into an incredibly large number of complicated solid materials . . . " (B 96). General recognition that there are different gases, as opposed to different forms of air, took longer. Still, by 1773 Lavoisier recognized that solid, liquid, and gas are three distinct states in which a single kind of natural body can occur (B 98). This is a major departure from the older framework and was part of Lavoisier's attempt to build a new framework for chemistry. This new framework recognizes 33 elements, but does not include any of Aristotle's elements, although different members of that original set are treated differently: earth and air are viewed as heterogeneous mixtures, water is recognized as a compound of hydrogen and oxygen, and fire is excluded from this discussion altogether. It is the task of chemical analysis to identify the elements: "Lavoisier defined the chemical element pragmatically and operationally as any substance that could not be analyzed by chemical means" (B 119). But what can be analyzed by chemical means depends on the state of chemistry, so it is no great surprise that Lavoisier's list includes some elements that are not on later lists. The distance between Lavoisier's

understanding of elements and later views can be indicated by his inclusion of light and caloric (the matter of heat, not to be confused with phlogiston) among the elements. We are not dealing just with different views of the extension of an established concept, but with different conceptual frameworks for chemistry. These examples also indicate that continuous use of such words as "air" and "element" does not guarantee that the same concept is associated with these words throughout.

The story of water is worthy of further comment. In the 1780s Cavendish, Priestly, and Watt had all observed that water appears when an electric spark is passed through a mixture of common air and inflammable air (hydrogen), but interpreted this result in terms of the phlogiston theory.[5] Cavendish, for example, concluded that water is a compound body made up of pure air and phlogiston. Lavoisier came closest to the modern view of water as a compound of hydrogen and oxygen (B 109–10), although Lavoisier's understanding of oxygen was rather different than ours. He believed that oxygen was the principle of acidity – that the presence of oxygen makes a compound an acid. This view is reflected in the German term for oxygen, *Sauerstoff* (B 107). In addition, the sense in which Lavoisier uses the term "principle" does not exist in current science, although it was pervasive in the early modern period.

Although Aristotle viewed the theory of elements and atomism as opposed alternatives, atomistic thinking continued to develop, e.g., in the work of Boyle and Gassendi. By the end of the eighteenth century atomism was integrated with Lavoisier's account of the elements. This new atomism took a major step forward in the early nineteenth century when Dalton argued that the determining feature of each element is the relative weight of its atoms. Dalton's new theory includes the introduction of the laws of fixed and multiple proportions

> when elements combined to form more than one compound, the weights of one element that combined with a fixed weight of the other were bound to be small whole numbers. For example, if:
>
> $A + B = AB$ and $2A + B = A_2B$
>
> then the weights of A combined with the weight B are in the simple ratio 1:2
>
> (B 143–44).

While the emphasis in this passage is on whole-number ratios, the existence of different combinations of the same elements was a significant innovation. It included cases in which two or more atoms of one element combine with two or more atoms of another element – e.g., combinations such as 2A + 2B or 2A + 3B. Dalton allowed for this possibility from the beginning, but Berzelius, for one, resisted until 1831 (B 158). Eventually this work led to the introduction of a new concept VALENCE (B 241–45), along with the recognition that some elements exhibit multiple valences. By the end of the

century it was also recognized that many naturally occurring elements are molecules composed of two atoms of the same element, and even that a single element can occur in more that one molecular form – e.g., common oxygen $O_2$ and ozone $O_3$.

I am going to pass over many important nineteenth century developments, but a few must be mentioned to set the stage for more detailed discussion of some developments in the late nineteenth and early twentieth centuries.[6] New discoveries in electricity were integrated into chemical theory through the work of Davy, Berzelius, and Faraday. One important outcome was the introduction of the concept of an ION for the parts of a compound deposited at the electrodes in electrolysis. This is a new concept: an ion of oxygen or sodium is not identical with an atom of familiar samples of these elements (B 371–82). The term "ion" is a neologism introduced by Faraday and Whewell.

In the 1820s cases were discovered in which two substances with the same chemical composition have different properties due to different structural arrangements of their components. Such cases are a natural possibility on an atomistic view, and were already suggested by Boyle. Once they were found to exist another new concept was required. In 1830 Berzelius coined the term "isomer" to refer to this concept (B 214). Mendeleev's periodic table was announced in 1869. Each element is characterized by a distinct atomic weight and these weights serve as the ordering principle for the table.

I now want to consider in somewhat greater detail a number of developments at the interface between chemistry and physics that occurred in a period of less than twenty years during which conceptual innovations came at a rapid rate. The story begins in 1895 when Röntgen discovered X-rays. He was working with a cathode-ray tube when he "was quite startled to notice a fluorescence on his detector" (P 1); he labeled the unknown cause of this fluorescence "X-rays." I will not describe how the cathode-ray tube was developed, or why Röntgen was working with it, or how he came to be using this particular detector, but the opening lines of the paper in which he announced his discovery indicates the array of concepts that would have to be introduced in the course of such an account.

> If the discharge of a fairly large Rühmkorff induction coil is allowed to pass through a Hittorf vacuum tube . . . and if one covers the tube with a fairly close-fitting mantle of thin black cardboard, one observes in the completely darkened room that a paper screen painted with barium platinocyanide placed near the apparatus glows brightly or becomes fluorescent with each discharge, regardless of whether the coated surface or the other side is turned toward the discharge tube. This fluorescence is still visible at a distance of two meters from the apparatus.
>
> It is easy to prove that the cause of the fluorescence emanates from the discharge apparatus and not from any other point of the conducting circuit.
>
> (Quoted in P 37–38, ellipses in P)

Pais underlines the conceptual distance between Röntgen's position and our own.

> In 1895 Roentgen could not yet know that X-rays may be considered as a stream of particles – photons – with zero mass. He did not know then that cathode rays consist of electrons; those were discovered only two years later. Nor could he have anticipated that within a few months X-rays would be the spur to the discovery of radioactivity.
>
> (P 3)

A few pages later Pais adds: "cosmic rays had not yet been discovered, the only accelerator in captivity was a cathode ray tube [although it was not thought of as a particle accelerator], and relativity theory and quantum theory were yet to come" (8).[7]

Within four months the discovery of X-rays led to the discovery of another unanticipated phenomenon, radioactivity, by Becquerel.[8] To understand Becquerel's thinking as his research proceeded we must keep in mind that the source of Röntgen's X-rays was a fluorescent spot on the wall of his cathode-ray tube.[9] This relation between fluorescence and X-rays led Becquerel to suspect that fluorescent crystals would also emit X-rays. After some searching he found a crystalline uranium salt that exhibits fluorescence and also gives off penetrating rays that affect a photographic plate. Becquerel initially concluded that his crystals were absorbing energy from the sun and giving off X-rays; his earliest experiments dealt only with uranium crystals that had been exposed to sunlight. A typical set-up consisted of a photographic plate wrapped in black paper to protect it from sunlight, with the uranium salt resting on a copper cross that was sitting on the covered plate. The developed plate would have an image of the cross. But an unplanned observation showed Becquerel that his interpretation of the role of sunlight in generating the penetrating radiation was mistaken. He had prepared some of his usual experiments on 26 and 27 February 1896. However, he reports,

> as on those days the sun appeared only intermittently, I held back the experiments that had been prepared, and returned the plate-holders to darkness in a drawer, leaving the lamellas of the uranium salt in place. As the sun still did not appear during the following days, I developed the photographic plates on the first of March, expecting to find very weak images. To the contrary, the silhouettes appeared with great intensity. I thought at once that the action must have been going on in darkness. . . .
>
> (Ra 11)

It is not known why Becquerel decided to develop these plates, but his immediate response was to do further experiments which confirmed that sunlight

was not required. The upshot was the recognition of a new phenomenon that was soon labeled *radioactivity*.

The next step was made when Marie Curie and Gerhard Schmidt independently discovered another element, thorium, that gave off the same kind of radiation as uranium (P 54; Ra 1). A short time later Marie and Pierre Curie, working together, discovered two more radioactive elements in pitchblende, which they named polonium and radium (P 55–56; Ra 1).

About the same time, Thomson discovered that X-rays ionize gases; he turned further research on the topic over to his student Rutherford who continued this work for two years and extended it in new directions. "As he mastered the details of gas ionization, Rutherford moved from x rays as the agent to ultraviolet light and then to uranium" (Rb 9). In the course of this research Rutherford found that uranium was emitting two different kinds of rays. One type, which he labeled *alpha rays*, produce most of the ionization but can be easily blocked by intervening material. The other type, *beta rays*, are much more penetrating but produce little ionization. Others found similar radiations from radium and added exploration of the effects of a magnetic field on this radiation. In particular, Pierre Curie found that the penetrating rays were easily deflected by a magnetic field, while the field had no detectable effect on the non-penetrating rays (Rb 11–12). In 1900 Becquerel showed that beta rays are streams of electrons. Pais notes that the recently discovered electrons were "another novelty not anticipated theoretically" (10). The nature of alpha rays remained unclear at this time, although Becquerel – drawing on the fact that streams of electrons cause X-rays – thought that alpha rays were a secondary X-ray caused by the beta rays.[10]

Late in 1898 Rutherford moved to his first professional position at McGill University in Canada. He now focused his research on radioactivity and shifted his methodology from using radioactivity as a means of investigating ionization, to using ionization as the basis for investigating radioactivity (T 21). At McGill Rutherford met Soddy, a young chemist also in his first professional position. They collaborated for a period of about eighteen months from 1901–03; during this period they developed a theory of radioactivity that had wide implications for our understanding of the nature of the "elements."

Rutherford and Soddy met when they agreed to engage in a debate on the chemists' and physicists' concepts of an atom (T 24–28). The discovery that radioactive elements eject electrons had raised, among physicists, the possibility that the chemical atoms are not the most fundamental units of matter. Soddy defended the chemists' view of atoms as fundamental, non-composite particles, and questioned whether electrons should be considered material objects. Rutherford defended the physicists' view and maintained that chemists had to adapt to the new discoveries. The interaction was one of mutual respect, but the entire field was new to Soddy – who was taking Rutherford's course on the effects of the various radiations on ionized gases. Soddy acknowledged that Rutherford had taught him what he knew of these developments (T 27).

Rutherford had been studying thorium emanation – a gaseous product of thorium's radioactivity – and convinced Soddy to collaborate and study the chemical nature of this emanation (T 40). One major outcome of their collaboration was the conclusion that radioactivity involves a transformation of one element into another; for Soddy this was a repudiation of the view he had defended in their debate. There were two major versions of their transformation theory; I will consider both, beginning with the route to the first version.

Rutherford and Soddy began their joint study with the hypothesis that the emanation is produced by thorium, but their research, plus parallel research by others, led them to reject this claim. The purity of the available thorium had been questioned by some chemists. Moreover, Crookes and Becquerel, working independently, had used standard chemical procedures to separate an intensely radioactive substance from uranium, and they found that the residual uranium was not radioactive. Crookes thus concluded that "the radioactive property ascribed to uranium and its compounds is not an inherent property of the element, but resides in some outside body which can be separated from it" (Ra 75); he labeled this substance *UrX*.[11] Soddy succeeded in carrying out an analogous procedure with thorium, separating out the new substance *ThX*. This result led Rutherford and Soddy to conclude that thorium is not itself radioactive (Ra 115–16), and that ThX is the source of the emanation. "There remains only one step to prove beyond doubt that the radioactivity and emanating power of thorium are not specific properties of the thorium molecule – the preparation of thoria free from these properties – and on this problem we are now engaged" (Ra 116).

However, Rutherford and Soddy ran into a pair of anomalies. First, they found that over time the radioactivity of the remaining thorium increased. Becquerel reported a similar increase in the activity of his purified uranium samples, and Soddy verified Becquerel's results. Soddy also found that when the activity of the thorium had increased he could remove more ThX. This supported the conclusion that some kind of change was taking place in the non-radioactive thorium which resulted in the chemically distinct radioactive ThX, which was the source of the emanation:

> The results therefore find their simplest expression on the view that just as a chemical change is proceeding in thorium whereby a non-thorium material is produced, so the latter undergoes a further transformation, giving rise to a gaseous product which in the radioactive state constitutes the emanation.
>
> (Ra 137)

Second, Soddy found that no matter how much ThX he removed from the thorium, approximately 25 percent of the radioactivity remained. Moreover, the residual activity consisted solely of alpha particles, while ThX gave off both alpha and beta particles. (In fact, Soddy's ThX was a mixture of several isotopes.) Soddy now repeated Crookes analysis of uranium, and here too he

found a residual radiation consisting only of alpha particles, while UrX (also a mixture of isotopes) gave off both kinds of radiation. For a while Soddy was somewhat bewildered about why Crookes – a superb chemist – found no radiation from the purified uranium, but Soddy eventually realized that he and Crookes were using different techniques for detecting radioactivity. Crookes had been using a photographic method that required wrapping the plates to protect them from light; the wrapping absorbed the alpha particles. Soddy used an electrical detector that was especially sensitive to the highly ionizing alpha particles. It is important to be clear that Crookes had not made a careless mistake. New phenomena were being studied using new methods; no properties of the radiation or of the detecting instruments were instantly knowable. All of these had to be worked out as research proceeded.

To understand the next step in Rutherford and Soddy's thinking we should keep in mind that they began with the hypothesis that thorium is radioactive and the source of the emanation. Then, we have seen, they believed that they had refuted this hypothesis, and were now convinced that thorium is not itself radioactive. To account for the residual radioactivity, they postulated a new radioactive substance produced by thorium; this second substance was inseparable from thorium by known chemical procedures. Thus we arrive at the view I referred to above as Rutherford and Soddy's *first theory*: non-radioactive thorium undergoes two different kinds of changes yielding two radioactive substances, ThX which is chemically separable from thorium, and a second substance that is chemically inseparable (Ra 142–43; T Ch. 4).[12]

An interesting sequence of events followed. Soddy was less well established as a researcher than Rutherford, so in order to boost Soddy's career two joint papers on the analysis of thorium were published in a chemistry journal, with the transformation theory in the second paper. They then rewrote the papers for publication in a physics journal, and while they were rewriting the second paper Rutherford and Soddy concluded that thorium is radioactive after all, and that there is no need to postulate a distinct non-separable component. They added an addendum to the new version of the paper in which they announced their new transformation theory. Here is their description of the difference between the two theories.

> So far it has been assumed, as the simplest explanation, that the radioactivity is *preceded* by chemical change, the products of the latter possessing a certain amount of available energy dissipated in the course of time. A slightly different view is at least open to consideration, and is in some ways preferable. Radioactivity may be an *accompaniment* of the change, the amount of the former at any instant being proportional to the amount of the latter. On this view the non-separable radioactivities of thorium and uranium would be caused by the primary change in which ThX and UrX are produced.
>
> (Ra 149)

This theory is far from the modern account in many respects, but I want to emphasize just one. The primary event is viewed as a change in the *structure* of a thorium atom; the emission of a radioactive ray is caused by this change. It will help us understand what is going on in both of these theories if we consider views current at the time on the nature of atoms.

Pais (178) introduces his discussion by noting that model-building is one of the projects physicists regularly pursue. "Whatever blocks they have, models they must build. At the turn of the century they had only one species of block: electrons. Accordingly they set out to build atoms from electrons only." This project begins with Thomson in a paper of 1897 in which he recalls Prout's hypothesis that all atoms are built up out of hydrogen. Although this hypothesis fails, Thomson argues that the underlying idea is tenable if we take electrons as the elementary constituents (P 178–79). Rutherford was Thomson's student, and was familiar with models in which a hydrogen atom contains hundreds of electrons.[13] It was recognized that some positive charge was required to neutralize this large negative charge. Initially Thomson preferred to evade the issue but around 1900 Larmor concluded that if the electrons are moving in a ring around a centrally located positive charge, then a dynamically stable equilibrium can be achieved (P 181–82, see 180–83 for other early attempts at modeling the atom). Becquerel's 1900 discovery that beta radiation consists of electrons supported the view that electrons are the essential constituent of atoms.[14] From this perspective it might be possible for an atom to change its chemical type by shifting from one equilibrium state to another – perhaps with the emission of electrons or alpha particles. In particular, some thorium atoms could change to ThX, and ThX atoms could then change into atoms of the emanation.

This view of the atom will help us understand the motivation for a concept that Rutherford introduced in 1904, RAYLESS DECAY (Ra 217–18). There were some cases in which it was clear that a transformation had taken place, but no radiation was detected. One possible explanation is that the experimenters failed to detect the radiation. For example, low-energy beta rays produce little ionization and cannot be detected by the electrical methods that Rutherford preferred. Rutherford notes this possibility, but also considers a different explanation: That the change results from a change in the atom's equilibrium state without the emission of any radiation. From this perspective atomic transformation is the fundamental phenomenon and radioactive emission is only one mode by which such transformations occur.

I now want to digress from the main line of my discussion in order to press an important point. The period we are examining is one in which researchers are attempting to understand a new phenomenon; as this attempt proceeds many new concepts are introduced, applied, and modified or abandoned. It is easy to find historical examples of cases in which new concepts appeared in the course of our cognitive history. X-rays and radioactivity are two such examples, and we will encounter many more in the course of this

chapter. It is harder to find examples of concepts that were dropped – exactly because they no longer occur in our current repertoire. The case at hand provides a rich field for locating such concepts. Trenn has stressed this point with regard to the separable and inseparable components of the first Rutherford-Soddy theory:

> Historically, such constituents belong to the same category as phlogiston and caloric. This is another example of a conclusion drawn on the basis of available evidence and within a theoretical framework ultimately found to be erroneous. But in April 1902 Rutherford and Soddy, having overcome their initial scepticism, fully embraced active constituents, both separable and inseparable, produced and made active in the process of transformation, as the general explanation of radioactivity.
>
> (T 75–76)

RAYLESS DECAY is another example of a concept that was introduced and soon dropped, as is INDUCED RADIOACTIVITY. In 1899 the Curies discovered cases in which a non-radioactive material placed near a radioactive material acquires a temporary radioactivity. They interpreted this as an induction phenomenon on the model of induced magnetism and induced currents, and this interpretation was widely accepted. Indeed, Becquerel attempted to interpret the reappearance of radioactivity in his uranium samples as another case of radioactive induction (Ra 118; T 51–52). However, an alternative explanation due to Rutherford and Soddy (Ra 160) prevailed: The non-radioactive material did not become radioactive, but was contaminated by a radioactive substance. As a result, the concept of induced radioactivity was dropped. But in the same paper Rutherford and Soddy introduced another short-lived concept. They noted the existence of short-lived decay products that appeared in the course of radioactive decays, and they considered these to be a special class of atoms that they labeled "metabolons": "Their instability is their chief characteristic" (Ra 162). We will encounter further examples of concepts that were introduced and then rejected as we examine other historical cases.

I want to consider one more enduring concept that Soddy introduced several years after his collaboration with Rutherford had ended: ISOTOPE.[15] This new concept, and the deep changes in chemical thinking that it involved, was a direct outcome of the discovery of radioactivity. As noted above, the thesis that a characteristic weight is the defining feature of each chemical element was central to nineteenth century chemistry. It had been introduced by Dalton, was embodied in Prout's thesis that each element is compounded out of hydrogen atoms, and provided a major part of the conceptual basis for locating elements on the periodic table. Yet anomalies appeared throughout the century so that by 1886 Crookes put forward the "audacious" but testable speculation that the weight standardly associated

with an element was that of the majority of its atoms, and that some might have slightly different weights (Bruzzaniti and Robotti 1989: 309). This hypothesis was not immediately embraced. Variant atomic weights associated with a specific element were generally interpreted as failures of chemical analysis, rather than as evidence against the principle. Nevertheless, as the study of radioactivity continued, anomalies accumulated. In particular, several radioactive substances were discovered that had different atomic weights and different half-lives – which seemed to indicate that they are chemically distinct – but which could not be separated by any known chemical procedures.

In 1910 Soddy undertook a study of a substance known as *mesothorium*.[16] Since pure mesothorium was not commercially available, he obtained a mineral that was known to contain this substance and began the process of separating mesothorium from other constituents. He found, however, that the only procedures that would yield mesothorium were those required to separate out radium. Moreover, once these procedures had been applied he could not separate the radium from the mesothorium – or from ThX, which was also included in the mixture. Fully aware that he was violating a fundamental principle of chemistry, Soddy concluded that these three substances are chemically identical, and drew the same conclusion for some other cases. Soddy now undertook two projects in order to confirm this conclusion. First, he searched the literature to learn all he could about the various radioactive elements. Second, he began to work with a young chemist, Fleck, who proposed a new technique: mixing together known quantities of radioactive elements and then trying to alter their proportions by chemical means. Fleck's results supported Soddy's conjecture about chemical identities, while Soddy's literature search showed important patterns in radioactive decays – patterns that were also noted by others. In particular, it became clear that transformations occur in which an element emitted an alpha particle and two beta particles (in any order). By this time protons and the nuclear atom had been discovered, but the neutron was still two decades in the future. It was generally believed that the nucleus contained both protons and electrons, but that electrons make no significant contribution to an element's weight (although it was recognized that electrons have mass). With beta decay treated as involving no change of weight, the transformations in question leave an element's slot in the periodic table unchanged while its weight drops by four units (see Fajans in Rb 207–19; Soddy in Rb 219–28). This makes it strikingly clear that a single element can have two different atomic weights. A new concept was required, and in 1913 Soddy introduced the term "isotope" for this concept. Other radioactive transformations could result in two different elements with the same atomic weight (isobars), so *it was no longer possible to characterize an element by its weight*.

The effect of this discovery on chemical thought was profound. A completely new basis was required for locating elements on the periodic

table. Given the prevailing view of the nucleus, Soddy proposed that the difference in the numbers of the two constituents – dubbed the "intra-atomic charge" – provides the proper criterion.[17] Clearly, the concept of an intra-atomic charge is not the same as the modern concept of atomic number: it assumes a view of the nucleus that is now rejected, and is calculated in a way that makes no contemporary sense, even though this calculation gives the same result as modern calculations of atomic number. Intra-atomic charge, as understood at the time in question, is another vanished concept.

The impact of this discovery on chemical practice was, in one respect, even more dramatic. Soddy notes that an immediate consequence of the discovery of isotopes was to change the precise determination of unique atomic weights from a central research project of chemistry to an irrelevant undertaking.

> There is something, surely, akin to if not transcending tragedy in the fate that has overtaken the life work of that distinguished galaxy of nineteenth century chemists, rightly revered by their contemporaries as representing the crown and perfection of accurate scientific measurement. Their hard-won results, for the moment at least, appears as of as little interest and significance as the determination of the average weight of a collection of bottles, some of them full and some of them more or less empty.
>
> (1932: 50)

Still, much of the existing body of chemical knowledge was unaffected. The arrangement of elements in the periodic table was not changed, even while the conceptual basis of this ordering was undercut. All results of standard chemical and spectroscopic analyses also remained unchanged, along with most of the accepted physical properties of the elements. However, those properties explicitly involving considerations of atomic weight, such as density and diffusion rates, had to be reconsidered (Soddy 1932: 44). New tests for identifying isotopes of an element were needed, tests that could detect small weight differences in chemically indistinguishable samples. The most important technique was soon embodied in Aston's mass spectrograph. In addition, the concept of the half-life of radioactive elements provides a means of recognizing different isotopes of some elements, as well as a new means of distinguishing among radioactive elements. This interplay between radical change and continuity will appear in other cases we will examine.

The concept of an isotope was published in 1913, the same year that gave us Bohr's new theory of the atom. From this point forward conceptual transformations continue to build in many directions, with relativity and quantum theory playing a central role in the story. We will encounter these theories in subsequent discussions in this book, but I have told enough of the story for my present purpose, which is to provide one set of illustrations

of the prevalence of conceptual change in human thought. The development of chemistry, physics, and their interface has required the wholesale abandonment of ways of thinking from earlier periods, and the frequent introduction of new concepts. (See Kragh 2000 for a recent summary.) In the later stages of the story we find multiple conceptual changes taking place in rather short periods of time. While many of these changes have been radical, when we examine their microstructure we also find continuous strands running through each transformation. I will return to these examples in later chapters as we develop the tools required for such microstudies. For the moment I want to build up our stock of examples of conceptual change by looking at other fields.

## 2.2 Mathematics[18]

> But you know what mathematicians are like – always meddling. No sooner does one of them come up with a definition of dimension in terms of directions, when some other smartarse has to improve on the idea by finding a completely different definition that gives the same answer when the dimension of a space is a whole number, but works for other spaces too.
>
> (Stewart 2001: 68)

The history of mathematics is an especially rich field for studying conceptual change. The extension of established concepts and the introduction of new concepts are major modes of mathematical development, and while mathematicians have great flexibility in conceptual innovation, there is usually a definite motivation for a particular move. In addition, because of the relative clarity and precision of mathematics, it is often easier to see what is going on than in other fields. I am going to discuss several examples and use them to introduce some important forms of conceptual change.[19]

### *2.2.1 Numbers*

I begin with a line of development that played a vital role in the history of mathematics: extensions of the concept of a number. Ancient Greek mathematicians focused their attention mainly on geometry and thus on numbers to which they could give a geometrical interpretation; these included positive integers, fractions, and irrational numbers such as the square-root of two which (they knew) cannot be expressed as an integer or fraction. Some Greek mathematicians encountered cases that involve other kinds of numbers but generally did not follow up on them. For example, Diophantus recognized the existence of quadratic equations that have only negative or imaginary roots, but considered these to be unsolvable (Ka 143; M 165). In general, Diophantus thought of quadratic equations as having only one solution. If an equation had a positive and

negative root he took the positive root as the solution; when he encountered two positive roots he took the larger as the correct solution. Early Chinese and Indian mathematicians made limited use of negative numbers, but were not fully comfortable with them (BM 201; Ka 185), while Arab mathematicians "were familiar with negative numbers and the rules for operating with them through the work of the Hindus," but still rejected negative numbers (Ka 192).

I will now jump to the sixteenth century and restrict discussion to European mathematics where many major developments occurred, and where we can follow these developments in some detail. I will take integers, fractions, irrational numbers, zero, and positional notation as the established basis, and consider some issues involved in admitting negative and complex numbers. (Zero seems to have been generally accepted as a number by around 1500, Ka 251–52.) We can approach the problem by considering attempts to solve different types of equations. This approach is well founded in the actual history, and will bring out several key issues.

Given just positive integers and zero we can solve all equations of the form $x - B = 0$, where $B$ is a positive integer. Including fractions extends our scope to equations of the form $Ax - B = 0$, where $A$ is also a positive integer. What is missing at this point is the ability to solve equations of the above sort in which the minus sign is replaced by a plus; to solve these equations we need negative numbers. Moreover, once we contemplate the possibility of negative solutions, the question arises whether to allow negative coefficients – that is, negative values of A and B. The admissibility of negative coefficients and solutions also arises in the case of quadratic and higher-order equations, which provided a more important locus for discussion than simple linear equations. Resistance and ambivalence with respect to negative numbers was deep and long lasting. This resistance provides a good indicator of the degree to which allowing the full legitimacy of negative numbers required new modes of thought. It also indicates differences from modes of thought that are common now. I want to canvass some important examples.

In *Arithmetica Integra* (1544), Stifel allowed negative numbers to appear as coefficients of equations, but called them *numeri absurdi* and did not allow them as solutions (BM 282). Cardan (1501–76), whose solution of cubic equations will be discussed below, also used negative numbers, but was not convinced of their legitimacy and called them *numeri ficti* (BM 287–88). Vieta (1540–63), who made major contributions to the development of algebra, did not allow either negative roots or coefficients (BM 305). In the early seventeenth century, Girard allowed negative solutions to equations (BM 305–6), but Descartes did not consider negative roots to be true roots (BM 345; D 229–30). "On the whole not many sixteenth - and seventeenth - century mathematicians felt at ease with or accepted negative numbers as such, let alone recognizing them as true roots of equations" (Ka 253). As this passage suggests, the views of mathematicians were far from

unanimous, and the history of the acceptance of negative numbers was not linear. Wallis and Newton, for example, constructed graphs with negative abscissas and ordinates (Ka 319) and Newton plotted curves in all four quadrants (Ka 548). In his *Optics* Newton takes negative numbers for granted and uses them to clarify the relation between attraction and repulsion: "And as in Algebra, where affirmative Quantities vanish and cease, there negative ones begin; so in Mechanicks, where Attraction ceases, there a repulsive Virtue ought to succeed" (1952: 395). In the eighteenth century most textbook authors "felt it necessary to dwell at length on the rules governing multiplications of negative numbers, and some rejected categorically the possibility of multiplication of two negative numbers" (BM 459). Resistance continued even into the nineteenth century where De Morgan held that "$0 - a$ is inconceivable" and maintained that if a negative number appears as the solutions of a problem, it "indicates some inconsistency or absurdity" (Ka 593, cf. Kb 155–56).

These were not just expressions of personal distaste; there were genuine difficulties and confusions about negative numbers, and arguments against their legitimacy. In the seventeenth century, for example, Arnauld, doubted that the ratios –1:1 and 1: –1 are equal because:

> –1 is less than +1; hence, How could a smaller be to a greater as a greater is to a smaller? The problem was discussed by many men. In 1712 Leibniz agreed that there was a valid objection but argued that one can calculate with such proportions because their form is correct, just as one calculates with imaginary quantities.
>
> (Ka 252)

In a book published in 1655 Wallis argued that negative numbers are "larger than $\infty$ as well as less than zero" (Kb 116, cf. Ka 253) because $a/b$ is infinite when $a$ is positive and $b$ is zero, thus if $b$ is less than zero the ratio must be larger than infinity.

De Morgan illustrated his objections to negative numbers with the following problem:

> A father is 56; his son is 29. When will the father be twice as old as the son? He solves $56 + x = 2(29 + x)$ and obtains $x = -2$. Thus the result, he says, is absurd. But, he continues, if we change $x$ to $-x$ and solve $56 - x = 2(29 - x)$, we get $x = 2$. He concludes that we phrased the original problem wrongly and thus were led to the unacceptable negative answer. De Morgan insisted that it was absurd to consider numbers less than zero.
>
> (Ka 593)

We might ask why De Morgan was bothered by the original solution since inserting –2 into the original equation will give the correct answer: the event

asked about occurred two years ago. But this response suggests that we are thinking about this situation differently than De Morgan did. Kline emphasizes that "negative numbers were not really understood until modern times" (Ka 593).

I noted above that irrational numbers were recognized by the Greeks. Euclid distinguishes between rational and irrational numbers in Book X of *The Elements* and proves theorems about irrationals on a purely geometric basis. In an algebraic context these numbers are required if equations such as $x^2 - 2 = 0$ are to have solutions, but early-modern mathematics were ambivalent (see Ka 251–52 for a brief summary). I will not pursue this case any further at the moment, but I want to note that modern definitions of irrational numbers were developed in the latter part of the nineteenth century, and make use of concepts that were not available in the sixteenth, seventeenth, and eighteenth centuries (for discussion see BM 563–65; Ka 982–87).

The most confusing problems about what count as genuine numbers focused on the so-called "imaginary" numbers.[20] These numbers are required to solve such equations as $x^2 + 1 = 0$, a point that was long known; we have already encountered the view that such equations lack solutions. The need to come to terms with square roots of negative numbers was, however, pressed upon mathematicians from a different direction: The Cardan-Tartaglia solution (developed in the mid-sixteenth century) of cubic equations of the form $x^3 = px + q$, where $p$ and $q$ are positive integers.[21] The general solution of this equation is, in modern notation:

$$x = [q/2 + (q^2/4 - p^3/27)^{1/2}]^{1/3} - [-q/2 + (q^2/4 - p^3/27)^{1/2}]^{1/3}.$$

Imaginary numbers appear because there are values of $p$ and $q$ for which $q^2/4 - p^3/27$ is negative, even though the root in question is real:

> Whenever the three roots of a cubic equation are real and different from zero, the Cardan-Tartaglia formula leads inevitably to square roots of negative numbers. The goal was known to be a real number, but it could not be reached without understanding something about imaginary numbers. The imaginary now had to be reckoned with even if one did agree to restrict oneself to real roots.
>
> (BM 288)[22]

Still, discomfort about imaginary numbers continued; many objections paralleled those we have already encountered for negative numbers. In the sixteenth century Bombelli "formulated in practically modern form the four operations with complex numbers; but still considered them as useless and 'sophistic'" (Ka 253). Descartes, who introduced the term "imaginary," considered their status to be worse than that of negative numbers. Descartes came to terms with negative roots of equations by finding a general method of transforming

equations with negative roots into related equations with positive roots (Ka 252, 271; D 230–35). Since this cannot be done with imaginary roots, he concluded that they are not genuine roots of equations (Ka 253–54).

Many earlier mathematicians distinguished carrying out formal operations with a symbol and admitting that the symbol stands for a number. Leibniz drew this distinction for both negative and imaginary numbers. Euler, who made major contributions to the theory of complex numbers, also had his doubts:

> Because all conceivable numbers are either greater than zero or less than 0 or equal to 0, then it is clear that the square roots of negative numbers cannot be included among the possible numbers. Consequently we must say that these are impossible numbers. And this circumstance leads us to the concept of such numbers, which by their nature are impossible, and ordinarily are called imaginary or fancied numbers, because they exist only in the imagination.
>
> (Quoted in Ka 594)

However, this did not stop Euler from carrying out detailed studies of these imaginary numbers (Dunham 1999: 86–87). Even Cauchy, "who founded the theory of functions of a complex variable during the first few decades of the 19th century, refused to treat expressions such as $a + b\sqrt{-1}$ as numbers." Instead he interpreted these expressions as being about real numbers: "For example, the equation $a + b\sqrt{-1} = c + d\sqrt{-1}$ tells us that $a = c$ and $b = d$" (Kb 155). Hamilton and DeMorgan included complex and negative numbers under the same anathema (Kb 155–57), although we will see below that Hamilton introduced a way of thinking about complex numbers that worked around his objections.

I now want to examine the various kinds of numbers from a different perspective. Consider four number systems (I will take negative numbers for granted in this discussion): integers, rationals, reals, and complex numbers. Beginning with the integers we can introduce other kinds of numbers by means of generalizations – although there is an important respect in which this is misleading. Properly speaking each of the following steps involves a different number concept; it is only from a specific perspective that we can view one system as a generalization of another. I will develop this point as we proceed.

Given the integers, each rational can be viewed as a pair of integers specified in a given order: instead of writing $A/B$, we could write $<A,B>$, which is not, in general, equal to $<B,A>$. With this in mind, I will use the more familiar fractional notation in this discussion; the case in which the denominator is zero is excluded. Let us consider the standard rule by which we add rational numbers. This rule requires finding a least common denominator, so the sum of $A/B$ and $C/D$ is $(AD + BC)/BD$. Although we often write expressions of the form $A/B + C/D$, use of the symbol "+" in this expression is

misleading (even though it may not generate errors in practical contexts) because this symbol is governed by a different rule when used for adding fractions than when used for adding integers. The "addition" operation for rational numbers is defined in terms of the addition and multiplication operations for integers. As a result, when we write:

$$A/B + C/D = (AD + BC)/BD$$

"+" is used in quite different ways in its two occurrences.

Note another difference between the system of integers and that of rationals. In the former system each integer is unique, but rationals come in infinite equivalence sets: each rational, $A/B$, is equal to every rational of the form $mA/mB$, where $m$ is an integer. Now consider the subset S of rationals (of the form $A/B$) in which $B = 1$. There is an infinite set of rationals equivalent to each member of S, and we can view a member of S as representing one of these sets. The members of S can be put into one–one correspondence with the integers, which yields an isomorphism.[23] It is because of this isomorphism that we can view the rational numbers with $B = 1$ as providing an image of the integers in the set of rationals. *These rationals are not the same mathematical items as the integers*. Indeed, each integer corresponds to an infinite subset of rationals. We will see as we proceed that this is one instance of a common kind of conceptual innovation in mathematics: One seeks to "generalize" a concept C by finding a different concept that specifies a class of items I such that some subset of I is isomorphic to the instances of C.[24]

The reals can now be constructed out of the rationals. The technical story is rather more complex than in the case we have just examined (see BM 563–65; Ka 982–87, or a textbook of modern algebra), but the upshot is the same: We have new definitions of the "arithmetical" operations and a new class of mathematical items with a subset that is isomorphic to the rationals. Once we have introduced real numbers, complex numbers (which have the general form $a + bi$) can be constructed as ordered pairs of real numbers $<a,b>$.[25] In this case $a$ gives the real part of the complex number and $b$ gives the imaginary part; these are not numerators and denominators as occurred for the rationals. There is also an image of the reals in the complex domain: the set of reals is isomorphic to the subset of complex numbers in which $b = 0$. Once again we need a new rule for combining our numbers – a new "addition" rule. This rule is defined in terms of operations on real numbers and is quite simple: we add the real parts and add the imaginary parts. But we must again distinguish different operations that are commonly indicated by the symbol "+." Note especially that when we write $a + bi$, "+" does not indicate the same operations as it does when adding real numbers. In fact, we need three different "addition" concepts: one to express the combination of the real and imaginary parts of a complex number, one to express the sum of two complex numbers, and one for the addition of real numbers. Since we

are now in the domain of complex numbers, I will use a simple "+" for the first operation, as I have already done in this paragraph. I will introduce "$+_C$" for the addition of complex numbers, and "$+_R$" for the addition of real numbers. Here, then, is the rule for adding complex numbers:

$$(a + bi) +_C (c + di) = (a +_R c) + (b +_R d)i^{26}.$$

Multiplication introduces further complications. Multiplication of rationals is straightforward: the new numerator is the result of integer multiplication of the input numerators; the new denominator is the integer multiple of the input denominators. The multiplication of real numbers involves technical complexities that I will pass over here. The multiplication of two complex numbers must be treated as analogous to the multiplication of two binomials, although there are two different multiplication concepts that must be distinguished. The point of the distinction is particularly clear when we calculate the square of a complex number. One kind of square is the straightforward application of the binomial multiplication rule:

$$(a + bi)^2 = a^2 - b^2 + 2abi.$$

The other kind of multiplication requires introduction of a new concept. The two complex numbers $a + bi$ and $a - bi$ are described as *complex conjugates* of each other; our second "squaring" operation consists of multiplying a complex number by its complex conjugate. In many situations this is the important version of the square of a complex number because it always yields a real number. There is no analog to this special form of the squaring operation in the other number systems we have considered.[27]

Are there are other number systems that we might introduce? We need not go any further as long as we are concerned with solutions of algebraic equations – equations of the general form

$$a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0,$$

where the *n*s are integers and the *a*s are complex numbers. Every such equation can be solved in the domain of complex numbers (Birkhoff and McLane 1953: 107–9). However, there are real numbers (and thus complex numbers) that are not solutions of algebraic equations. Such numbers are labeled *transcendental* and include $e$, $\pi$, and many others. These are still considered real numbers, but we cannot generate all real numbers by considering solutions of equations. Mathematicians have proposed other kinds of numbers as well; I will note one of these rather briefly. As indicated above, we can think of each complex number as a vector in a plane. Given this representation, the nineteenth century mathematician Hamilton sought an analogous type of number that would be represented by a vector in 3D space. This was not a search for any arbitrary item, since Hamilton required

reasonable generalizations of the arithmetic operations into this new realm. Given these constraints, it turns out that no such numbers can be found. However, he did discover a class of four-term numbers, known as *quaternions*, but only after he relaxed one of the constraints with which he began: the generalization of multiplication to quaternions is not commutative (Ka 776–82).

### *2.2.2 Exponents*

In our discussion of number systems we encountered one common form of mathematical generalization: introduction of a wider concept whose extension includes a subset that is isomorphic to the extension of the original concept. I now want to introduce another common kind of mathematical generalization that can also generate conceptual change: relaxing a restriction (implicit or explicit) on the range of some term or terms in a formula. Exponents will illustrate the process.[28]

Positive-integer exponents provide an abbreviation for iterated multiplication; thus we can abbreviate $x{\cdot}x{\cdot}x$ as $x^3$. Such abbreviations provide a powerful cognitive tool because once an abbreviation is introduced its properties can be worked out, and we can then reason directly in terms of the new notation. It is, for example, straightforward to show that $x^n{\cdot}x^m = x^{n+m}$ and $x^n/x^m = x^{n-m}$. Once we master these rules we can manipulate exponents without having to refer back to the original definitions. Now, the division rule might lead one to ask what happens if $m$ is greater than $n$, and from there it is a natural step to consider extending the range of exponents beyond positive integers. Some early mathematicians extended the range of powers, without the exponent notation, and often without anything like a modern justification for this extension. For example, in the fourteenth century Oresme developed the laws for positive integer and fractional powers (BM 263), and around 1500 Chuquet included negative numbers and zero as powers (BM 277). In 1685 Wallis used fractional, negative, and even irrational powers in specific formulas (BM 382), and "Newton used positive, negative, integral, and fractional exponents . . . " (Ka 261, cf. Newton 1999: 541). Newton's binomial theorem provides a justification for this usage.[29] I want to examine the introduction of various numbers as powers in several stages, with an eye to the conceptual innovations involved rather than to historical details.

Negative-integer exponents can be introduced by finding an interpretation that is in accord with the established laws; interpreting $x^{-1}$ as $1/x$ does the trick. Applying the multiplication law we find that, for example, $x^5{\cdot}x^{-3} = x^2$, which accords with the proposed interpretation. Expressions of the form $x^n{\cdot}x^{-n}$ should equal 1 on this interpretation, which works out correctly if we interpret $x^0$ as 1; this interpretation stands up to further exploration. The same approach leads to an interpretation of fractional exponents: treat the numerator as a power and the denominator as a root. Thus $x^{1/3}$ is the cube root of $x$, and $8^{2/3}$ is the square of the cube root of eight (or the cube root of the square, the order of the operations is irrelevant), i.e., four. These are

straightforward extensions that do not involve conceptual change. Extension of exponents to include irrational and complex numbers is somewhat trickier.

It is not immediately clear how the results already established apply to irrational exponents since we cannot settle this question by appeal to the original definition of an exponent. Strictly speaking, we cannot define irrational exponents within the confines of algebra; techniques from calculus are needed, which requires introduction of a substantial body of new concepts. However, we can make a detour through logarithms and produce a means of calculating with irrational exponents. While this will also require some additional concepts, it is a route that high school students commonly follow before learning calculus. Thus I will leave the main line of this discussion in order to introduce logarithms; then I will return to irrational exponents.

The logarithm of a number can be defined by introducing a *base* and a power to which we raise that base. In the case of so-called *common logarithms* the base is ten. Suppose that $10^L = n$; then $L$ is the common logarithm of $n$.[30] Thus logarithms are exponents, and the value of this new notion derives from the laws of exponents. By temporarily replacing numbers by their logarithms we turn multiplication into addition, division into subtraction, exponentiation into multiplication, and extraction of roots into division. For example, to multiply $x$ by $y$, we find $\log x$ and $\log y$, add these logarithms, and then find the anti-log – i.e., the number whose logarithm is our result. To calculate $x^n$ we calculate $n\log x$, and then find the anti-log. (The required logs and anti-logs can be looked up on readily available tables or many calculators.) Given this general rule, there is no special problem about substituting an irrational number for $n$.[31] In effect, we have moved through the following steps: (1) Powers of numbers are introduced independently of logarithms for the non-problematic cases. (2) Logarithms are defined in terms of powers. (3) The restriction to non-problematic powers is dropped because logarithmic calculations make sense with irrational exponents. Put differently, once we understand rational exponents we can introduce logarithms, which then provide a vehicle for introducing irrational exponents. Logarithms can also be used to introduce complex exponents, but we must explore logarithms a bit further before considering this case.

When we learn about logarithms in high school (or college) we are typically told that negative numbers do not have logarithms – a claim built into many electronic calculators. But this holds only if we restrict ourselves to the real numbers; in the complex domain negative numbers have logarithms – which are complex. To see why consider a result due to Euler. Around 1748 Euler established the following formula connecting imaginary numbers with trigonometric functions:

$$e^{i\theta} = \cos\theta + i\sin\theta, \qquad \text{(E)}$$

where $\theta$ is an angle measured in radians and $e$ is the base of natural logarithms (see note 12). Consider the case in which $\theta = \pi$. Since $\cos\pi = -1$ and

$\sin\pi = 0$, E reduces to $e^{i\pi} = -1$.[32] If we take natural logs of both sides of this equation we arrive at: $ln(-1) = i\pi$.

Another surprising consequence of E is worth a passing mention. Suppose we ask a typical trigonometric question: What angle has sine *S*? Strictly speaking, the answer is an infinite set of angles because if $\sin A = S$, then *S* is also the sine of every angle equal to *A* plus an integral multiple of $2\pi$. E, then, implies that every number, including the reals, has an infinite set of logarithms.[33] As we will see shortly, E also implies that complex numbers have logarithms.[34] I submit that learning to include negative and complex numbers among those that have logarithms, and to think of logarithms as coming in infinite sets, requires modification or replacement of the concept of a logarithm that many of us once learned. As a result, people with different levels of mathematical education do not associate exactly the same concept with the term "logarithm."

Before considering complex exponents, I want to note one historical point about logarithms. When Napier invented logarithms (c. 1594, but not published until 1614) he was not thinking in terms of a base and an exponent, but rather in terms of a correlation between a geometric series and an arithmetic series. (For details see BM 312–14; Ka 256–58.) Briggs suggested recasting this construction in terms of exponents using ten as a base, and Napier accepted this proposal. If we restructure Napier's original version in terms of a base and exponents, we find that his base was $1\text{-}10^{-7}$ = .9999999. Two reasons have been suggested for this choice. First, to avoid fractional exponents in the geometric series, Napier needed a base that was small, but not too small, because the change from the logarithm of one number to that of the next must be gradual if logarithms are to be useful for calculation. This suggested a number less than, but close to, 1. Second, Napier's original concern was to simplify calculations in trigonometry where contemporary practice divided a unit circle into $10^7$ parts. So he took one minus one part as his effective base. One consequence of this choice is that higher numbers have smaller logarithms. Further mathematical work led to new methods of calculating logarithms, for example, in terms of integrals, sums of infinite series, and limits of infinite series. Some mathematicians take these new calculation methods as definitions of the concept of a logarithm (Ka 354, 404), and this will serve to introduce a *third kind of conceptual change* that is common in mathematics: taking a consequence of a structure as the basis for a redefinition of that structure. Usually this results in a more general concept with the original version as a special case. I will return to this case in Sec. 2.5.

We are ready now to consider complex exponents. Our first step is to interpret $x^{a+bi}$ as $x^a \cdot x^{bi}$ in accordance with the established laws of exponents. Recall, however, our previous discussion of the meaning of "+" in the expression of complex numbers. Given that meaning, application of the addition law for exponents to this case involves an extension of earlier concepts. In a similar way, we can treat $x^{bi}$ as $(x^i)^b$ and focus just on the

imaginary part of this expression. But what in the world does $x^i$ mean? What sense can we make out of the operation of raising a number to an imaginary power? This requires an interpretation for $x^i$, but before we proceed I want to note that there is no a priori guarantee that such an interpretation will be found (see the remarks above on Hamilton's extension of complex numbers), or that if one is found it will be of any mathematical interest. In the present case Euler's formula E provides the basis for such an interpretation. In effect, we can use E to define complex exponents. (See M 171–72 for a useful discussion.)

Given E, calculation of imaginary powers is straightforward – but not lacking in further surprises. Sometimes the result of our calculation will itself be complex, which might be expected since we are in the complex domain. For example, if we take $\theta = 1$ we get $e^i = \cos 1 + i \sin 1 = .54 + .84i$. Taking $\theta = \pi/2$ gives an especially interesting and useful result since $\cos \pi/2 = 0$ and $\sin \pi/2 = 1$. Substituting into E gives $e^{i\pi/2} = i$. But complex exponents do not always yield complex results. A striking example occurs when we use the value of $i$ we have just derived to calculate $i^i$. This equals $(e^{i\pi/2})^i$ which, by the laws of exponents, equals $e^{i \cdot i\pi/2} = e^{-\pi/2} = 1/e^{\pi/2}$. Thus $i^i$ is a real number, equal to approximately 0.20788.

### *2.2.3 The Gamma Function*

In our discussion of number systems we saw that each extension of the number concept can be viewed as a case in which mathematicians introduce a new structure whose extension has a subset that is isomorphic to the extension of the original structure. I want to note an example in which a mathematician explicitly sought a generalization of this kind, and in which the conceptual gap between the starting point and end point is considerably more dramatic than in the case of numbers. The initial concept is a *factorial*, a relatively simple concept that requires no mathematical background beyond multiplication of integers. A specific example will provide the key idea: five factorial, written 5!, equals $5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$. In general, $n!$ is the product of descending integers from $n$ to 1.

Around 1731 Euler sought a generalization of this concept that would make sense of non-integral values of $n$ – in particular, fractional values. Within the conceptual confines of multiplication this seems nonsense, but it is no more intrinsically nonsensical than is the attempt to solve equations such as $x^2 + 1 = 0$. Although such attempts will be absurd relative to a particular conceptual repertoire, a different conceptual repertoire may remove the absurdity. The result that Euler arrived at, known as the *gamma function*, requires a considerable body of mathematical knowledge beyond the ability to multiply; it is defined as:

$$\Gamma(n) = \int_0^\infty e^{-x} x^{n-1} dx \qquad \text{(G)}$$

In addition to multiplication, *e*, and negative exponents, we also require the conceptual machinery involved in the notion of a definite integral with an infinite upper limit. This concept could not have been formulated by any mathematician before the late seventeenth century since several of the required concepts had not yet been developed.

G is a generalization of the factorial in that a subset of values of this function is isomorphic to the factorials. The mapping is not quite straightforward since $\Gamma(n)$ is equal to $(n - 1)!$. Still, we can use the gamma function to calculate factorials, so we have a generalization of the kind that Euler sought. While this new notion gives results for fractions, as desired, it can also be used for real numbers, including negative real numbers, *except* the negative integers. The gamma function is meaningless for negative integers. To my knowledge no one has provided a mathematically interesting generalization that would include these in its scope.[35]

### *2.2.4 Calculus*

I now want to examine some examples from the history of calculus, both from the period in the seventeenth century preceding the work of Leibniz and Newton, and from the eighteenth century when the new techniques received extensive development and application. These periods are of special interest because the modern understanding of the foundations of calculus would not be developed until the 1860s. As a result, seventeenth and eighteenth century mathematicians thought about this subject in ways that are significantly different from the way it is taught today. This led to the use of definitions, techniques, and arguments that are now viewed as confused, but it is doubtful that modern analysis – including the modern accounts of its central concepts – would have developed without this work. "The concepts and techniques of the infinitesimal calculus are the result of a long line of mathematical development stretching almost unbroken from antiquity to the present day . . . " (BA 253). Let me emphasize that all my examples are from the work of mathematicians whom we still consider major contributors to the development of calculus – and they are only examples. I will focus on two themes: the sum of an infinite series, which is central to the concept of an integral, but has much wider use in mathematics; and the ratio of small quantities, which is central to the concept of a derivative.

Kline identifies four major problem areas that motivated the seventeenth century work leading to calculus. One of these included "finding the lengths of curves, for example, the distance covered by a planet in a given period of time; the areas bounded by curves; [and] volumes bounded by surfaces . . . " (Ka 343). Much seventeenth century work on these topics began with Kepler: "The identification of curvilinear areas and volumes with the sum of an infinite number of infinitesimal elements of the same dimension is the

essence of Kepler's method" (Ka 348). Thus Kepler found the area of a circle by adding up the areas of:

> an infinite number of triangles, each with a vertex at the center and a base on the circumference. . . . In an analogous manner he regarded the volume of a sphere as the sum of the volumes of small cones with vertices at the center of the sphere and bases on its surface.
>
> (Ka 348)

One subject of dispute in this period was whether the elements to be summed must have the same dimension as the item they compose. Tacquet and Roberval agreed with Kepler (BO 139–42) that the dimensions must match, but others disagreed and treated curves as sums of points, areas as sums of lines, and volumes as sum of planes.[36] Wallis seems to straddle these two views; an example from his work will illustrate a common approach (see Figure 2.2).

To determine the area of a triangle of altitude H and base B, Wallis considered the triangle to be made up of infinitely many rectangles parallel to the base. But each of these rectangles is infinitely thin and thus equivalent to a line:

> Wallis did not consider the distinction between lines and parallelograms of any great importance in the sense that parallelograms whose altitudes are supposedly infinitely small, that is, having no altitude (since a quantity infinitely small is no quantity), scarcely differ from a line. He does, however, make the proviso that, the line is to be regarded as having so much thickness that, by infinite multiplication, it becomes capable of acquiring an altitude equal to that of the figure in which it is inscribed.
>
> (BA 206)

We can, then, consider the triangle to be made up of infinitely many parallel lines, each having an altitude of $H/\infty$. (It was Wallis who introduced the

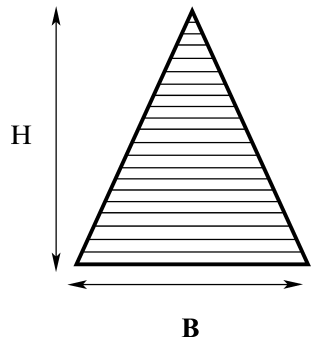


*Figure 2.2* Wallis' Triangle

symbol ∞ for infinity.) Since the lengths of the lines vary uniformly from B to zero, the average length is B/2. Adding up the lines, the area of the triangle is ∞ x B/2 x H/∞ = BH/2 after the infinities are canceled (see also BO 171). A modern approach would begin with small rectangles and consider the area to be the *limit* of the sum as the width of the rectangles approach zero. We will see that although some mathematicians had at least an inkling of the limit concept, limits did not become the basis for understanding this process until the nineteenth century.

The process of summing an infinite series played a central role in the development of mathematics in other fields besides geometrical problems. In 1685 Wallis offered the following summary of the main stages in the development of mathematicians' understanding of infinitesimal processes:

> 1. Method of Exhaustion (Archimedes).
> 2. Method of Indivisibles (Cavalieri).
> 3. Arithmetick of Infinites (Wallis).
> 4. Method of Infinite Series (Newton).
>
> (Quoted in BA 213)

One of Newton's many contributions was his discovery of the general binomial – the rule for expanding expressions of the form $(a + b)^n$. Newton (and others) routinely substituted negative, fractional, and irrational numbers for $n$, which led to infinite series, although it was not the only source of such series.[37] But there was confusion about when it makes sense to sum an infinite series.

> As Newton, Leibniz, the several Bernoullis, Euler, d'Alembert, Lagrange, and other 18th century men struggled with the strange problem of infinite series and employed them in analysis, they perpetrated all sorts of blunders, made false proofs, and drew incorrect conclusions; they even gave arguments that now with hindsight we are obliged to call ludicrous.
>
> (Kb 142)

A major source of these confusions came from the lack of a clear understanding of the difference between convergent and divergent series, along with the recognition that only convergent series can be summed. As one example of the kind of argument that resulted from a failure to understand this point consider the series (Kb 142–43):

$$1/(1+x) = 1 - x + x^2 - x^3 + x^4 \ldots . \qquad \text{(S)}$$

This series converges only for cases in which $x^2 < 1$, but early eighteenth century mathematicians discussed the case in which $x = 1$. S then becomes:

$$1/2 = 1 - 1 + 1 - 1 + 1 \ldots ,$$

which was generally considered to be correct. There were at least two additional arguments that led to this result. One argument comes from rewriting the series as:

$$S = 1 - (1 - 1 + 1 - 1 \ldots)$$

Since the expression in parentheses is equal to *S* we get

$$S = 1 - S$$

which gives S = 1/2. Second, Leibniz argued that if we take the sums of progressively larger sets of terms – i.e., the first term, the sum of the first two terms, the sum of the first three terms, and so forth – we get:

$$1, 0, 1, 0 \ldots .$$

Since 1 and 0 are equally probable, we should take the mean, 1/2, as the sum. "This argument was accepted by James, John, and Daniel Bernoulli and Lagrange" (Kb 143).

Euler provides an interesting variation on this theme: He did distinguish between convergent and divergent series and recognized that only convergent series can be summed – except for the special case in which a divergent series is equal to an explicit function: "Whenever an infinite series is obtained as the development of some closed expression, it may be used in mathematical operations as the equivalent of that expression, even for values of the variable for which the series diverges" (quoted in Ka 463). Thus he considered it legitimate to substitute 1 into S and take the value of the left-hand side as the sum of the series (see Ka 446–47 for additional examples).

Throughout the early work with infinite series "the question of convergence and divergence was certainly not taken too seriously; neither, however, was it entirely ignored" (Ka 460).

> Newton, Leibniz, Euler, and even Lagrange [who attempted to use series as the foundation for calculus, see BO 252–53; Ka 430–32] regarded series as an extension of the algebra of polynomials and hardly realized that they were introducing new problems by extending sums to an infinite number of terms. Consequently, they were not prepared to face the problems that infinite series thrust upon them; but the apparent difficulties that did arise caused them at least occasionally to bring up these questions. What is especially interesting is that the correct resolution of the paradoxes and other difficulties was often voiced and just as often ignored.[38]
>
> (Ka 460)

Next consider some issues associated with the early development of the derivative concept. We use derivatives to find maxima and minima, which is

another problem area that Kline lists as leading to the development of calculus (Ka 343). There is an approach towards our methods in a problem that Fermat discussed (BM 155–56; Ka 347–48).[39] Consider a line segment of length $a$ that is to be divided at a point $x$ (see Figure 2.3). Taking the two parts of the line as the sides of a rectangle, the problem is to find the division point such that the area of the rectangle is a maximum. The area of the rectangle, in modern notation, is:

$$A = x(a - x) = ax - x^2$$

Suppose we move $x$ a small amount to the right, $E$. The new area:

$$A' = (x + E)(a - x - E) = ax - x^2 - 2Ex + Ea - E^2.$$

Fermat maintains that at the maximum, $A = A'$. Equating the two and doing a bit of algebra we get: $2xE - Ea + E^2 = 0$. Dividing through by $E$ gives: $2x - a + E = 0$. Fermat then sets $E = 0$ and gets $x = a/2$ – that is, the rectangle of maximum area is the square.

Two steps in this argument are questionable. First, consider the step in which Fermat sets $E = 0$; we would, instead, consider the limit as $E$ approaches zero. Kline comments: "Fermat did not see the need to justify introducing a non-zero $E$ and then, after dividing by $E$, setting $E = 0$" (Ka 348). The second issue concerns setting $A = A'$. This issue was raised by Fermat's contemporaries and he "justified the equating of the two values of $A$ by remarking that at a maximum point they are not really equal but they should be equal. He therefore formed the pseudo-equality which became equality on letting $E$ be zero" (BO 156). Many decades later Berkeley asked, "by what right he took the positions $x$ and $x + E$ to be different and yet in the end said that they coincide" (BO 156). Fermat used a similar method to find tangents to curves – another problem that we handle by means of derivatives, and that is included on Kline's list (see BO 156–57; Ka 344–45).

Fermat's handling of the small quantity $E$ is one instance of confusion about infinitesimals that had a direct impact on attempts to understand the derivative concept. Many early researchers thought of a derivative as a ratio of two infinitesimals, which were taken to be smaller than any finite number, but not zero since 0/0 is undefined. Leibniz called these small quantities "differentials" and regarded the differential as the fundamental concept of calculus (BO 210–11). In Leibniz notation, $dy$ and $dx$ are differentials, the derivative is written as $dy/dx$, and the derivative is explicitly viewed as a ratio

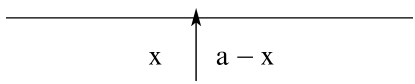


*Figure 2.3* Fermat's problem

of two items that can be manipulated independently. (Mathematicians now define the derivative as a single quantity.) Either of these might be set equal to zero at key points in a computation. For example, Leibniz defined a tangent to a curve as "a line joining two infinitely near points of the curve, these infinitely small differences being expressible by means of differentials or differences between two consecutive values of the variable" (BO 210, cf. Ka 377). But this small difference is then treated as zero since a tangent has only one point in common with the curve. In one of his attempts to justify such procedures Leibniz maintained that as long as appropriate rules were followed the results would be significant even if the meanings of the symbols were unclear. We have already encountered this view in Leibniz's approach to negative and imaginary quantities.

The use of differentials was central to the work of Leibniz's continental followers, who did the great bulk of the work of developing calculus in the eighteenth century. Euler proposed one of the more extreme interpretations. He considered the derivative to be 0/0 on the grounds that the only number smaller than all other numbers is zero. Setting $0/0 = n$, and noting that $n \cdot 0 = 0$ for any $n$, he concluded that a derivative could have any value whatsoever. The problem of calculating a derivative thus became the problem of determining the value of 0/0 in a specific case (BO 244; Kb 147–48). "Thus Euler accepts unqualifiedly that there exist quantities that are absolutely zero but whose ratios are finite numbers" (Ka 429).

Newton's practice in his earliest writings on fluxions (1669, 1671) also amounted to treating infinitesimals as zero at selected points in a computation.[40] By 1676 he had abandoned infinitesimals, criticized the dropping of small quantities, and introduced a new approach based on ratios of changing quantities.[41] In *Principia* he described this new approach as "the method of first and ultimate ratios" (1999: 433). The idea is that we begin with a ratio of two finite quantities, and consider what happens as they become *evanescent* (i.e., vanish). Newton's discussion of this method seems to involve a genuine anticipation of the later account of derivatives in terms of limits:

> Those ultimate ratios with which quantities vanish are not actually ratios of ultimate quantities, but limits which the ratios of quantities decreasing without limit are continually approaching, and which they can approach so closely that their difference is less than any given quantity, but which they can never exceed and can never reach before the quantities are decreased indefinitely.
>
> (1999: 442–43)

One can debate whether this is actually the modern notion of a limit, but for present purposes it is sufficient to note that even if it is, Leibniz and his successors did not think of derivatives in these terms. Thus even if Newton had the modern concept, this concept was not shared by many of the mathemati-

cians who made major contributions to analysis. Kline offers the following summary of the situation:

> Almost every mathematician of the eighteenth century made some effort or at least pronouncement on the logic of the calculus, and though one or two were on the right track, all the efforts were abortive. The distinction between a very large number and an infinite "number" was hardly made. It seemed clear that a theorem that held for any $n$ must hold for $n$ infinite. Likewise a difference quotient was replaced by a derivative, and a sum of a finite number of terms and an integral were hardly distinguished. Mathematicians passed from one to the other freely.
>
> (Ka 433–34)

The conceptual foundations of analysis were brought to their current state in the last half of the nineteenth century by several mathematicians; I want to underline two respects in which this final development involved major departures from earlier ways of thinking about the subject. First, clarification of the concept of a limit and its use as the basis for defining a derivative developed through the work of Bolzano and Cauchy, and culminated with Weierstrass (c. 1861). Once the derivative is defined in this way, it is clear that a derivative is a single quantity, not a ratio of two quantities (BO e.g., 253–55, 216–17, 221, 255). Contemporary mathematicians make use of differentials, but they define them on the basis of derivatives, and use them as distinct entities only to the extent that they have provided a justification for doing so.

Second, while much early thinking about derivatives and integrals was based on geometry, the new understanding was built on arithmetic (BO 273; Ka 950–52). This was a major departure from earlier thinking since the Greeks and many early moderns took geometry to be basic, and required a geometric interpretation before they would accept a new type of number as legitimate. By the end of the nineteenth century the foundations of the numbers were established independently of geometry, and there was even a movement to reconstruct geometry on the basis of arithmetic (Kb 182). But in order to achieve this goal, the various kinds of numbers that we discussed above first had to be put on an acceptable foundation. This was especially the case for irrational numbers because understanding their nature is deeply wound up with understanding continuity, which is basic for the modern concepts used in calculus. Only the integers were considered to be intrinsically intelligible, so the task was to construct the other numbers out of the integers. The historical process moved from the more questionable levels down to the basics. The first step was taken by Hamilton in 1837 when he defined the complex numbers as ordered couples of real numbers and formulated the appropriate arithmetic operations on these couples (Ka 775–76).[42] But this approach succeeds only to the extent that real numbers are not

problematic. The foundations of the real numbers were established by Weierstrass, Dedekind, and Cantor in the 1860s and 70s, leaving the rationals as the next set needing a foundation. Weierstrass accomplished this during the 1860s when he showed how to derive negative integers, positive rationals, and negative rationals from pairs of positive integers (Ka 987). This led to the further recognition that the integers needed an axiomatic foundation, which was provided by Peano in 1889. It was now possible to define all of the other numbers and deduce their properties without the need for any further axioms (Ka 988–89).

I note, finally, that in the early 1960s Robinson developed a new approach to calculus known as "non-standard analysis" in which infinitesimals and infinitely large numbers return. To a degree, proofs in this approach look like the early proofs in which infinitely small quantities are introduced, manipulated as if they were ordinary numbers, and then eliminated at the end of the argument. But this is only appearance because the new proofs are based on precisely defined concepts and are rigorously justified. Moreover, the basis for the approach lies in model theory, a part of modern logic that was developed only in the twentieth century. In other words, the conceptual basis for introducing the new version of infinitesimal and infinite numbers is quite different from that found in the seventeenth and eighteenth centuries.[43]

The key point of this discussion, given my aims in this book, is that major contributors to the formulation of calculus, and to the development of analysis for the first century of its existence, thought about their subject in different ways than do contemporary mathematicians. These early researchers associated concepts with "derivative," "integral," "differential," and so forth, but these were not the concepts that we associate with these terms. One central task of a theory of concepts is to provide a basis for understanding these transformations.

## 2.3 Biology, Technology, and Society[44]

> What was called Darwinism in 1859 was no longer considered so thirty years later, because the term had been transferred to something very different from that which it designated at the earlier period.
>
> (Mayr 1991: 91)

Developments in technology are often associated with conceptual change. The need for new concepts is clear when we consider such examples as the efficiency of a heat engine, the shape of an airfoil, the control rods in a nuclear power plant, the danger of a nuclear meltdown, and the array of concepts associated with computers. Many technological developments are driven by developments in science which themselves required conceptual innovation. In addition, new technology can generate challenges to existing social and legal concepts. I will develop these points by examining one

example in this section – *in vitro* fertilization (IVF). I will begin by summarizing some main steps on the route to our contemporary understanding of the roles that males and females play in sexual reproduction. This historical background will also provide further examples of concepts that have dropped out of our repertoire.

As one historian has noted: "Although it was clear to scientists, as it was to breeders, that sexual intercourse was necessary if the higher animals were to reproduce, there was no expert consensus until the late nineteenth century why this was so" (RI 106, n. 113). I will begin our historical sketch in the early seventeenth century when the prevailing view derived from Aristotle. His view, formulated in his own conceptual framework, was that females provide undifferentiated matter and males provide form, "including the formal, efficient, and final causes . . . " (Maienschein 1981a: 96, cf. G 29).[45] The process of individual development is initiated by the mingling of menstrual and seminal fluids; the fact that menstruation stops during pregnancy was considered evidence for this thesis.

This view of the initiation of pregnancy was challenged by Harvey in 1651; its rejection was central to his work on generation (G 21), even though he worked largely in terms of Aristotelian concepts. Harvey was convinced that all animals develop from eggs, produced in the female, that have an "innate capacity to develop after receiving the influence of male semen" (Farley 1981a: 163, cf. G 25–28). At this point sperm had not been identified, and Harvey was unclear on the exact nature of this influence, although he denied that semen makes physical contact with the egg (FA 17; G 28). Rather, development of the egg begins "when the male semen activates it by exerting some immaterial or vital influence, resulting in epigenetic development" (Maienschein 1981c). Harvey's new term "epigenesis" marks a new concept and is his most important departure from Aristotelian ideas: Harvey held that the form of a new organism develops gradually once the egg has been activated, rather than being completely available from the beginning (G 30; Maienschein 1981c). He also held that the egg makes a significant contribution to the new organism's form. At this point neither Harvey nor anyone else had observed eggs in female mammals, but general agreement on the role of eggs in development arose after de Graff published his discovery of "egg-like follicles within mammalian ovaries" (Farley 1981a: 163) in 1672, a discovery that was quickly confirmed by others (G 38–39). Some believed that they had actually seen mammalian eggs, but Von Baer first observed these in 1828 after considerable improvements in the lenses of microscopes.

*Preformationists* offered an alternative to both epigenesis and the Aristotelian view, holding that offspring already exist in some undeveloped form before reproduction. They disagreed on where this form is located. One version, *ovism*, took off from Harvey's thesis that reproduction always requires an egg, holding that a miniature preformed individual is encapsulated in the egg, and that seminal fluid somehow initiates its development. A competing preformationist view appeared after the discovery of sperm in the

seminal fluid in the 1670s (by van Leeuwenhoek and Hartsoeker). It was initially unclear what role, if any, sperm played in reproduction (Farley 1981b, G 54). The unanticipated appearance of sperm is, in many ways, like the unexpected appearance of X-rays and radioactivity; the exact role of sperm in reproduction remained a subject of dispute for some 200 years (see FA for an extended discussion). For a substantial period after their discovery many naturalists believed that sperm are parasites of the testes playing no role in reproduction – a view that survived well into the nineteenth century (FA 43–47). But *animalculists* maintained that the new organism is located in the sperm and that the egg provides only a base for the development of the organism.

> [T]hose who believed that the preformed germ was the spermatozoon – the animalculists – had the advantage over the ovists in that their view restored the male to the more important position in reproduction, and was thus in line with all tradition. In addition, they could point to a visibly moving, and therefore living, object as their postulated germ. This version of preformation, which was suggested by Leeuwenhoek in 1683 . . . , soon became popular, although it never won universal assent, and some prominent naturalists, including John Ray, continued to oppose it.
>
> (G 55–56; cf. Maienschein 1981b, e; FA 17–21)

Both of these preformationist views imply that all individuals that will ever exist were already there at the creation, as Gasking notes in a discussion of ovism. "It followed from such a view that there was no true generation; what appeared as the formation of a new individual was simply the growth of an organized living thing which had been formed at the beginning of Time" (G 42). Gasking emphasizes that in spite of its absurd appearance to us, preformationism "was accepted in slightly varying forms by such great naturalists as Leeuwenhoek, Ray, Réaumur, Haller, Spallanzani, Bonnet, and even Cuvier" (G 43).

One feature of the conceptual background made preformation plausible – even compelling: At least since Aristotle it seemed clear that *organs* form the basic building blocks of living beings. It follows that if an embryo is alive, it must have organs, and it is only a short step to the conclusion that the embryo is a miniature version of the infant. *It required a new understanding of the basis of life before this approach could lose all its attractions.* The key step was the discovery of cells and their role in life, which would only come in the middle of the nineteenth century (FA 47–54).

An important challenge to preformation came from Maupertuis who, in a series of works published between 1745 and 1757, injected studies of heredity in humans and animals into the debate (G Ch. 6). Maupertuis studied individuals with six fingers or toes (polydactyly) in four generations of a single family; this provided an important part of his argument that

males and females contribute equally to their offspring – a conclusion that was at odds with preformation (G 78–81; Olby 1981: 182). (Réaumur carried out an overlapping investigation of this phenomenon.) Maupertuis argued that both inheritance and the production of a new organism result from the merging of particles contained in male and female fluids. Yet he was still working in terms of the thesis that organs are the basic building blocks of organisms, and thought of the various particles as specific to the construction of particular organs. He did not think it possible that these particles were enclosed in a small object such as an egg or sperm. As a result, he rejected the view that an egg or Graff's follicles contained the female contribution to generation, and denied that the ovaries play any direct role in this process. Rather, Maupertuis held that the female contribution to reproduction is a fluid formed in the uterus. Semen carries the male contribution, but he believed that all of the semen is involved with the essential element consisting of solid particles in the semen. Sperm, he thought, are "motile particles whose function was to agitate the commingled mass of the two semina, and thus facilitate the mixture of essential parts" (G 83). Moreover, Maupertuis was a physicist who sought to extend the new Newtonian doctrine of attraction into biology, arguing that the synthesis of particles into organs was brought about by special forms of attraction.[46] His immediate successors generally rejected this view, along with the belief that life could be accounted for solely on a materialistic basis (G 83–87).

One of these successors was Buffon who maintained (1748) that there is an unbridgeable divide between living and non-living entities, and sought to explain this distinction by postulating two kinds of fundamental particles: organic and inorganic. Living beings are made up of organic particles which come together to form miniature versions of specific beings. "Just as a grain of salt was made up of numerous smaller grains, so organisms were composed of numerous minute replicas of themselves . . . each unit being a group of primary particles" (G 87). Buffon's views on reproduction were more sophisticated than those of earlier preformationists, but still amount to a variation on this approach. He rejected the view that each new individual is encapsulated in one of the parents, and accepted an equal role for both parents in determining the offspring's characteristics. An embryo is formed by mingling groups of organic particles contained in the male and female fluids – semen and a fluid produced by the follicles – but these groups are miniatures of specific organs. They mingle to produce the offspring whose "sex would be determined by whichever units happened to predominate . . . " (G 89). The embryo formed by this process is a miniature of the resultant offspring which "grew as new groups of vital particles were intercalated between the original ones" (G 91). Buffon's view also accounts for regeneration (in those species in which it occurs), and thus brings regeneration and reproduction under a single theory.

Although preformationism was in something of a decline after about 1730 (G 107), "from 1759 onwards to the end of the century there was a complete swing back to preformation" (G 101). This occurred even though detailed studies of the development of embryos did not provide direct evidence for preformation. Advocates of this view accepted, and even contributed to, these empirical studies, but considered preformation to be an explanatory theory in which the preformed entity was an unobservable theoretical postulate (G 102–4).

> Most of the late eighteenth century naturalists were forced back to preformation because they insisted on a causal explanation for generation and refused to believe that events in the world were due to any causes other than those reducible to physics and chemistry.
>
> (G 106, cf. FA 16–17)

Gasking suggests that this theoretical approach was more congenial to ovists than to animalculists (108), presumably because sperm were accessible to microscopic study so that some evidence of the preformed organs should appear. In addition, the discovery of parthenogenesis (in aphids and other species) strengthened the view that the preformed offspring must be in the female (G 110). The enormous waste of sperm provided an additional argument against animalculism, since it was generally held that nature does nothing without a purpose (FA 20–21). Challenges to preformation did develop in the late eighteenth century, but came from researchers who rejected the demand for mechanistic explanation. Many of these were German biologists who defended new varieties of epigenesis: "all these studies viewed epigenetic development as essentially an emergence of form which conforms to the particular morphological type of the parents" (Maienschein 1981a: 97). Maienschein also notes that Darwin challenged this view since it assumes fixity of species. Darwin's perspective was picked up by Haeckel who suggested that the development of each embryo recapitulated the species' evolutionary history.

The function of sperm continued to be elusive throughout the nineteenth century. While it came to be accepted that sperm play some role in initiating the development of an egg, it was unclear whether the sperm actually had to make contact with the egg, let alone enter it. As a result, it was also unclear if sperm just stimulate egg development, or actually make a contribution to the nature of the offspring (FA 70–71). While it seems obvious that both parents affect the characteristics of their offspring, the research we are discussing was largely carried out by laboratory-based physiologists: "Indifferent, even hostile, to the world of the amateur naturalist, [they] remained generally oblivious to the problems of inheritance – problems that were of the utmost significance to naturalists and animal and plant breeders" (FA 70).

These issues were largely settled only after the development of improved microscopes and staining techniques late in the nineteenth century. Work by

Hertwig and Fol in the 1870s was of particular importance. Hertwig described the merging of the egg and sperm nuclei to form the zygote, while Fol observed the penetration of the egg by a sperm (FA 160–65). Of course, these reports were not accepted without considerable debate.[47] Note that all these views precede the discovery of chromosomes, cell division, and the distinction between the process by which body cells divide (mitosis) and the rather different process involved in the division of reproductive cells (meiosis) – discoveries that are crucial for the developments leading to IVF.

Eventually attempts to understand the generation of new individuals had to mingle with studies of heredity. Mammals (and members of other classes) produce offspring of the same species, and children tend to have characteristics similar to those of their parents. Still, views on the mechanisms by which this occurs, and on the relative contributions of father and mother, varied. It is a considerable conceptual journey from preformation, which attributes an offspring's heredity to only one parent, to the views guiding IVF, which hold that each parent contributes 50 percent to an offspring's genetic endowment – but a different 50 percent to the endowment of different children – and that children of full siblings have a 25 percent genetic match. I will sketch some of the many stages along the way.

One common view of the mechanism of heredity was *pangenesis*: that hereditary characteristics are transmitted by particles produced by various parts of the body. This view goes back at least to Hippocrates (Mayr 1982: 635); Maupertuis should be included among its advocates. It was adopted by Buffon, who held that "the male determined the [offspring's] extremities, the female the internal parts and the overall shape and size" (Olby 1981: 182). Pangenesis was also advocated by Darwin who used it to account for, among other things, inheritance of acquired characteristics (Eiseley 1961: 217; Mayr 1982: 693–94). Here is a summary of Darwin's version.

> Darwin assumed that the cells of the body throw off minute material particles and that these particles, "gemules," he calls them, are gathered from all parts of the body into the sexual cells of the organism. Darwin thus assumes that the sexual cells contain only what is represented in the living body – or primarily so – and the particles they receive upon fertilization. Every character thus comes from the somatic, or body, tissues, and the germ cells contain only what is brought to them by the blood stream from all parts of the body. The germ is merely a device to create a new body out of the mingling of the particles of the parents' bodies.[48]
>
> (Eiseley 1961: 217)

This view has been replaced in our contemporary understanding of heredity, and PANGENESIS has vanished from the active scientific repertoire.

TELEGONY is another abandoned concept that played a role in nineteenth-century thought about heredity:

> The concept of telegony, which was almost universally believed in by nineteenth century breeders and fanciers and widely accepted within the zoological community, attributed to the "previous sire" – usually understood as the father of a female's first child – the power of influencing her subsequent offspring.[49]
>
> (RI 107–8)

Belief in telegony formed the basis for some advice found in the literature of animal husbandry where owners of a pure-bred bitch that had become pregnant by an undesirable male were advised to eliminate her from the breeding stock. Similarly, cattle breeders were advised never to start a herd with a purchased cow since they had no control over her previous mates. Belief in telegony provided an addition to the many reasons why female virginity was so much insisted on in human marriages. Darwin is included among the scientists who believed in telegony (RI 109–10).

PREPOTENCY is another concept from nineteenth century thought on heredity that is worth recalling. The idea is that members of certain groups have, for various reasons, a greater influence on the characteristics of their offspring than other groups. Those with greater influence were described as more prepotent. A highly inbred pedigree was one supposed source of prepotency. "So efficacious was social superiority, as embodied in pedigree, that it could tip the sexual scales that normally allotted the dominant role in shaping offspring to the male" (RI 115). Ewart, a professor of natural history, maintained that "the Jews, as a race, are more prepotent than the English – are better or purer bred" (quoted in RI 115). This view led some to advocate inbreeding – not only in animal husbandry, but in human mating as well (RI 119). In addition, members of "wilder" groups were held to be more prepotent than members of "civilized" groups. Consider the views of Millais, who wrote on animal breeding.

> As examples he offered not only crosses of horses with zebras and quaggas, and of wolves with various breeds of dogs, but crosses between European people and members of darker human groups, which he considered to be both relatively old and relatively wild. If the father of a white woman's child was "a Mongol, a Polynesian, a Red Indian, or a Negro," he asserted, it "will resemble the sire to a much greater extent than where the white man is the father of the dark woman's child."
>
> (RI 117)

Such views, found in the late nineteenth century, are enormously distant from the conceptual framework in which IVF was developed. Let us consider that procedure.

I assume that the reader is familiar with the basics of human reproduction and the main ideas from genetics involved in understanding human heredity. In the IVF procedure eggs are fertilized by sperm in a laboratory dish and

then transferred to the woman who is to become pregnant. If the transfer is successful one or more embryo(s) implants in the uterine wall, although pregnancy and the birth of a baby are far from guaranteed. In the early days of this procedure pregnancy was rare, although outcomes have improved with continuing research and experience. I will sketch the main steps involved in IVF since the details are relevant to some of the conceptual issues that arise.

The first step in the procedure is harvesting eggs by means of a far-from-trivial medical procedure. The woman supplying the eggs is given a variety of hormones to manipulate her menstrual cycle and cause "superovulation": bringing multiple eggs to maturity in a single cycle. The procedure begins with a hormone that halts egg development; it is continued until blood tests and ultrasound indicate that egg development has been stopped.

> When the function of the ovaries has been temporarily stopped, you will receive hormone injections for about 7 days to stimulate the development of ovarian follicles (fluid-filled sacs in which eggs mature). More blood tests and another ultrasound will be done to determine follicle growth. . . . After the appropriate degree of ovarian stimulation is reached, another hormone, called human chorionic gonadotropin (hCG), is injected to help the eggs mature and trigger ovulation.[50]
>
> (Larson 1996: 1220)

Next the eggs must be removed. The oldest technique is laparoscopy which requires general anesthesia and three small incisions in the abdominal wall. These entail the risks involved in general anesthesia and invasive surgery; Rowland (RD 26) notes three reported deaths at the time of her research. A later approach requires only local anesthetics, which are less risky than general anesthesia, but also involve some pain; the procedure is still invasive. More recent procedures are less risky, although they involve intravenous drugs to help the woman remain "comfortable and relaxed throughout the procedure" (Larson 1996: 1220). Eggs are removed by a needle through the vagina guided by ultrasound monitoring. The male contribution to IVF is considerably less demanding.

While the developments in biological concepts that led from nineteenth century ideas to IVF are manifold, the impact on social and legal thought is quite as dramatic and still in a process of resolution. Consider the impact of this procedure on the ancient concept MOTHER. Before the advent of IVF different kinds of mothers were recognized because of the many situations in which a woman conceives and gives birth to a child that she does not raise. This yields a distinction between a BIOLOGICAL MOTHER and a SOCIAL MOTHER, and has led to further social and legal distinctions in some societies, such at that between a STEP MOTHER and an ADOPTIVE MOTHER.[51] But until the advent of IVF, conception, pregnancy, and birth have always taken place in a single woman's body, and were all implicated in the concept of a

biological mother.[52] IVF has changed the conceptual landscape: conception need not take place in a body at all, and the woman who contributes half of a child's genetic endowment need not be the same woman who undergoes pregnancy and childbirth. This leads to a new distinction between a GENETIC MOTHER and a GESTATIONAL MOTHER, neither of whom need be the social mother.

The social impact of this new distinction can be illustrated by a legal case that occurred in California where the law recognized only one NATURAL MOTHER, but accepted either a blood test or the fact of giving birth as sufficient for establishing motherhood. These criteria can now conflict. When the issue arose in a specific case, "the California Supreme Court devised a new rule to break the tie by looking to the intentions expressed in the surrogacy agreement" (FK 219). Thus a contractual agreement became the key factor in determining who is the "natural" mother. Other maneuvers have occurred in other legal jurisdictions. In the Australian state of New South Wales the *Artificial Conception Act of 1984*

> provides that when a husband consents to the use of donor sperm to achieve his wife's pregnancy, he is presumed to have "caused the pregnancy" and to be the child's father. The sperm-donor is presumed *not* to have caused the pregnancy and *not* to be the child's father. . . .
>
> (S 227)

In Australia, the Victoria *Status of Children Act* carries this line of thinking further.

> In addition to creating a presumption of paternity in favour of a consenting husband, the Act covers children born from donated embryos or donated eggs. The Act creates an irrebuttable presumption that the birth mother is the mother of the IVF child and that the ovum donor is irrefutably *not* the mother.
>
> (S 228)

If some of these moves create a sense of dissonance, I urge that this is one indicator of a situation in which existing concepts have become inadequate and people are attempting to adapt.

I now want to consider the new concept SURROGATE MOTHER. Surrogacy occurs when a woman voluntarily becomes pregnant with the explicit intention of giving birth to a baby that will be reared by others. Surrogacy thus has a biological component since only a woman who has undergone pregnancy and childbirth counts as a surrogate; but the concept also involves the reasons why that woman became pregnant. A woman who becomes pregnant without this intention, and who then gives up the baby to be reared by others, is not a surrogate mother.[53] The phenomenon of surrogacy predates recent high-tech methods of initiating pregnancy. One common situation

occurs when a heterosexual couple's inability to achieve pregnancy rests with the woman. Surrogacy allows for the couple to use the male's sperm and have a baby that is genetically related to one of them. The surrogate may be impregnated either in the old-fashioned "natural" way, or – for moral and social reasons – by means of a syringe. The latter procedure is known as "artificial insemination" and may be considered a low-tech method. In either case, the woman who becomes pregnant is a surrogate only because she agreed at the outset to turn the baby over to this particular couple. It is less clear how to classify a woman who enters into a surrogacy agreement and then reneges in order to keep the baby for herself. I will return to cases of this kind in a moment, but first I want to consider some of the different combinations of egg-donor, sperm-donor, surrogate, and social parents that IVF makes possible, and that occur. In order to simplify the discussion somewhat, I will initially focus on cases in which the procedure is done on behalf of a heterosexual couple.

(1) The female member of the couple may be both the egg-donor and the woman into to whom the fertilized eggs are transferred. This occurs in the original situation for which IVF was created: the woman's fallopian tubes are blocked, or otherwise damaged, but the rest of her reproductive system is healthy. However, the procedure is also used in cases in which the woman's productive system is healthy, but there is a problem on her partner's side, such as low sperm count or low sperm motility. Surrogacy is not involved in these cases since the genetic mother and the gestational mother are identical.

(2) A woman may not be producing eggs – perhaps because she was born without ovaries – although she has an otherwise healthy reproductive system. In this case eggs may be provided by an egg-donor, fertilized in the laboratory, and transferred to the woman who wishes to become pregnant. Again, we do not have a case of surrogacy, but we do have a clear distinction between the genetic mother and the gestational mother.

(3) A woman who is producing eggs may have other problems with her reproductive system; she may, for example, have been born without a uterus. This can lead to a situation in which the couple who want a baby provide the eggs and sperm, and engage another woman as the gestational mother. Now we have a clear case of surrogacy. In other cases, both partners may have fully functional reproductive systems, but engage a surrogate for many reasons. For example:

> Pregnancy may be a serious burden or risk for one woman, whereas it is much less so for another. Some women love being pregnant, others hate it; pregnancy interferes with work for some, but not others; pregnancy also poses much higher levels of risk to health (even life) for some than for others. Reducing burden and risk benefits not only the woman

> involved, but also the resulting child; high-risk pregnancies create, among other things, serious risk of prematurity, one of the major sources of handicap in babies. Society also benefits when expensive problems like prematurity are avoided.
>
> (Purdy 1992: 304)

(4) Some cases involve three different women: the egg-donor (who is the genetic mother), a surrogate gestational mother, and the woman who will become the social mother. Given our current understanding of genetics, this leads to further variations. Often the gestational mother will have no genetic relation to her baby, but we now recognize degrees of genetic relationship between relatives. This opens up the possibility of getting a close relative to provide the eggs. To take but one example, if a woman provides the eggs leading to her full sister's pregnancy, then the gestational mother can assume a 25 percent genetic relation to the baby. Note that there is no control over which 25 percent is involved – different eggs will provide a different set of genes. None of this would make sense in terms of early accounts of heredity. We may also have cases that involve four distinct people because the sperm-donor need not be the intended social father.

(5) Surrogacy via IVF or artificial insemination may occur on behalf of a single male, a single female, a homosexual couple, or some other kind of non-traditional family group.[54]

We can now consider in more detail some of the impacts of IVF on legal and social concepts. Legal systems become involved when a surrogate mother refuses to relinquish the child. This may occur for many reasons, not the least of which is that a woman who enters into a surrogacy contract in good faith may change her mind as she finds herself bonding with the fetus she carries. This situation raises questions concerning the relative status of contracts, genetics, gestation, individual psychology, risk, and other considerations in determining who should be considered a child's parents. These questions are currently under debate in courts, legislatures, and among commentators with a wide range of philosophical, political, social, and theological interests and agendas.[55] The following are examples of *some* issues that have arisen so far.

There is dispute over the legal status of explicit surrogacy contracts when these exist. Although contracts play a central role in Western societies, there are areas in which contractual arrangements are not permitted. For example, individuals cannot enter into contracts to sell themselves or their children into slavery. It is also generally illegal to sell a baby for adoption or to sell one's organs for transplantation (another issue generated by new medical technologies). It is, however, quite legal to give a baby up for adoption and to donate organs such as one kidney or some bone marrow to another person. The point of these examples is that contractual arrangements do not automatically prevail, so that substantial issues may arise when a surrogate

mother changes her mind and challenges a surrogacy contract. These issues can involve rethinking our concept of a contract, or figuring out how to adapt that concept to new situations. Different legal jurisdictions have taken different approaches. Surrogacy for money is illegal in the UK and in some states in the US and Australia; it is legal in most US states. Recall the case mentioned above in which a US court appealed to provisions in a surrogacy contract to mediate a conflict between gestation and genetics as the basis for deciding who counts as the "natural" mother. That court could instead have reconsidered the view that there must be only one natural mother. Such diversity should not be surprising when people seek to extend existing concepts into situations that were not thought through, or even considered, in the past.

Another major topic of debate concerns which factors are relevant. Some feminists, for example, would give a substantial role to a woman's experience during pregnancy and to the major physical and psychological contribution of women to both artificial insemination and IVF. Another criterion that has arisen is the best interest of the child. Two contrasting cases will bring out some of the issues involved.

The first case to bring these issues to public notice in the US was Whitehead v. Stern. (This is also known as *The Matter of Baby M*. See Oliver 1992; RD 159–61, among the many published discussions.) In this case the baby was conceived by artificial insemination using sperm provided by Stern, the intended social father. There was an explicit, detailed contract between Stern and the surrogate Whitehead, who was the child's genetic and gestational mother. Whitehead decided to keep the baby and challenged the contract in the New Jersey courts. The contract was ruled legitimate and controlling by the trial court. Whitehead appealed, and the New Jersey Supreme Court rejected every aspect of the trial court's decision, ruled such contracts invalid, and accorded Whitehead full parental rights. But they awarded custody to Stern on the grounds that this was in the child's best interest: the Sterns were considerably more affluent, were better educated, and were better able to educate a child; in addition, the court concluded that the Sterns were likely to establish a better emotional relation to the child.

A different outcome occurred in an early case in the UK. When a surrogate refused to hand over the twins she had borne, they were made a ward of the court and left in her care pending final decision. Custody was eventually awarded to the gestational mother because of the time the infants had spent in her care: removing them from her custody was judged not to be in the children's best interest (RD 170–71).

I turn next to another cluster of legal and social issues generated by the details of IVF technology. Recall that superovulation produces several eggs. This is desirable because the probability of a single fertilized egg implanting in the uterus and yielding pregnancy is quite low; transferring multiple eggs increases the chances of pregnancy.[56] Still, several fertilized eggs often remain after the procedure, and these can be frozen and saved for later use.[57]

Questions arise about how to deal with them, and these may involve the contested conceptual issue of when an embryo is to be considered human, and thus endowed with rights. If this occurs at some stage, do these embryos have a right to be implanted in some woman? Are these eggs to be available for research? Should they be made available for use by other women? Should they be discarded? These are questions that were not explicitly addressed in the past when there were no spare embryos to consider. Although we created these spare embryos, figuring out how to think about them has much in common with trying to understand sperm or radioactivity. We should expect the resolution of the issues involving embryos to involve difficulties not encountered in the other cases because many people and groups believe they have a stake in the outcome. Some of the most challenging problems arise when the status of the prospective parents changes. Two cases will highlight some issues.

The first case is that of Mario and Elsa Ríos whose status changed drastically when both died in an airplane crash, "leaving behind two frozen embryos with no instructions for their disposition in case of their deaths" (FK 188–89). One set of questions arose because the couple were wealthy and it was not clear whether an embryo that was brought to term would have a right to inherit. "One law professor asked colleagues at a legal seminar to consider whether the embryos owned their parents' estate or the estate owned the embryos" (FK 189). Another complication arose because the Ríos' were American but had gone to Australia for their IVF procedure, so two different legal systems were involved. In this case each legal jurisdiction addressed a different question, which simplified matters considerably. In the US, where the estate issue was settled, a California court ruled that any children resulting from these embryos had no right to inherit. The disposition of the fertilized eggs was decided by the Victoria legislature; the embryos were donated anonymously for use by some other individual. But it is not difficult to imagine a case in which a California court decrees a right to inherit, and also decrees that the court has an interest in the ultimate disposition of the embryos. Or, an Australian court might claim an interest in the future financial status of the resulting children.[58] I submit that these questions cannot be settled just by becoming clear on the content of already available concepts such as inheritance and an estate. At a minimum they involve adapting existing concepts to situations that were never considered when these concepts were deployed in the past.

The second case concerns Mary Sue and Junior Davis who divorced leaving behind seven frozen fertilized eggs, with no agreement as to their disposition (FK 189–90; Shevory 1992: 232–45). This led to an extended legal battle over custody of the embryos. As these proceedings were going on both parties remarried. Mary Sue was no longer interested in becoming pregnant with the frozen embryos, but wished to donate them to another couple. Junior objected to having fatherhood imposed on him. The case eventually involved three levels of the US judicial system. The original trial court awarded custody to Mary Sue, but this decision was reversed on appeal by

the Tennessee State Supreme Court that propounded three ranked criteria to be followed in the absence of an explicit agreement:

> (1) Ordinarily, the party wishing to avoid procreation should prevail, assuming that the other party has a reasonable possibility of achieving parenthood by other means. (2) If no alternative means of achieving parenthood reasonably exists, then the argument of the party desiring to use the preembryos should be considered. (3) An intention to donate the preembryos to another couple should never prevail over an opposing gamete donor.[59]
>
> (FK 190)

The US Supreme Court allowed this last ruling to stand and the embryos were destroyed.

Another issue arises because there is some time lag between the point at which eggs are fertilized and the best time for transfer to the prospective gestational mother. This provides an opportunity for genetic screening. Some embryos are rejected because they are found to harbor genetic defects – and some because they do not have the desired gender. Gender preference is not a new idea, although it has not been explicitly endorsed or openly practiced in Western societies. But the ability to select a child's gender before pregnancy occurs requires new ways of thinking about this possibility. Screening for genetic disease is a new issue that requires the present understanding of genes and their role in heredity. Since both kinds of screening can occur as part of the same process, distinctions between them may be difficult to enforce. Deciding how to think about these possibilities requires attention to the way we decide on conceptual boundaries, and these decisions can be influenced by a variety of economic, ethical, political, social, and perhaps other agendas.

Another consequence of the use of multiple fertilized eggs is multiple pregnancies. (This also occurs with other, less drastic, fertility treatments.) Multiple pregnancies raise a host of problems for both the mother and the resulting children. The mother faces a higher risk of early delivery preceded by a period of hospitalization, an increased risk of medical conditions that are sometimes induced by pregnancy, and a considerably more uncomfortable pregnancy. The resulting infants face a greater risk of early delivery and low birth weight, birth defects, and physical and mental retardation (Overall 1992: 153). There are also economic and social costs to the parents and the wider community. The expense and effort of caring for multiple-birth children can be enormous. The mother of one set of quintuplets produced by IVF "changes diapers 50 times a day and goes through 12 liters of milk a day and 150 jars of baby food a week" (Overall 1992: 154). Depending on existing social arrangements, some of these costs may be borne by the community – and this raises questions about community interest in people's reproductive decisions. Other community costs arise when scarce resources

must be apportioned. Multiple-birth infants often need such extensive neonatal care that they affect the care available for other infants.

> When sextuplets were born prematurely in Cambridge in England, they effectively closed a special-care baby unit for three months. Dr Cliff Robertson said: 'From May to July we had to turn away more than thirty pre-term babies. God knows where they went. In order to have six babies we put thirty at risk'.
>
> (RD 66)

A decision to apportion scarce resources for one purpose amounts to a decision not to use them for competing purposes. To a degree, these are variations on familiar problems, but the way they are generated introduces new considerations. Once we bring technology, along with a variety of doctors, nurses, and technicians, into human reproduction, we have already ceased treating it as a private matter. Moreover, since the new procedures can have significant social impacts, there is expanding justification for the view that society should have a say in these decisions. For present purposes the key question is whether such decisions involve *conceptual* innovations. The answer depends on how we think about concepts.

Some aspects of these procedures clearly take us into contested conceptual territory. One response to the problems of multiple births is "selective reduction" – a technological fix for a problem generated by a technological solution to another problem (Overall 1992: 150); it consists of aborting some of the fetuses. This is an invasive procedure that has risks for the woman, and also risks termination of all the pregnancies she has gone to great lengths to achieve. (See Overall 1992 for discussion of methods of selective reduction, risks involved, and some of the social issues.) In addition, it takes us directly into the tangle of issues – including conceptual issues – surrounding abortion, which I will not pursue here. However, one point is worth noting:

> Abortion is used in the United States to terminate unintended (accidental) and unwanted pregnancies. Multifetal pregnancy reduction is used most often to reduce intentional pregnancies that result from the use of ovulation drugs or assisted conception. For many people, it is morally offensive to use technology to create fetuses with the intent to destroy some of them later if the technology works too well.
>
> (Kearney 1998: 182)

I noted above that excess fertilized eggs can be frozen and saved for later use. This has many advantages: if pregnancy does not occur frozen eggs can be used for subsequent attempts without another round of egg harvesting; women with cancer can store fertilized embryos before undergoing chemotherapy; there is a better chance of implantation if it is done in a cycle that did not involve stimulation; frozen eggs allow people to postpone deci-

sions on whether to have a family and how large it should be (Holmes 1992: 196–97). Another consequence of embryo freezing may be considered positive by some and negative by others: it gives more time for genetic diagnosis. Problems also arise because some fertilized eggs do not survive the freeze/thaw cycle, and because freezing equipment may fail or the company doing the freezing may go out of business. How we think about these possibilities depends on our views of the status of the frozen embryos.

From the perspective of the egg-donor, one advantage of frozen embryos is the increased set of options it makes available. A larger set of options is provided by freezing unfertilized ova. Early attempts at freezing ova failed, but the technical problems have been overcome to the extent that companies advertise this service on the World Wide Web. At a time when this technical success had not yet been achieved, one commentator emphasized some negative aspects: "I believe that success in freezing eggs would be disastrous for women. It would be another tooth in the saw that dismembers women into body parts, another spoke in the wheel that requires reproduction as a validation of true womanhood" (Holmes 1992: 197). I will comment only on the first of these remarks. There is a strong tendency among those who advocate and sell all of these procedures to dehumanize the people involved. For example, a woman who provides eggs for IVF is described as an "egg-donor," and a surrogate mother as a "host uterus." Whatever else is involved in these descriptions, they involve issues of conceptualization: Interested parties promote a specific terminology as part of an attempt to get us to think about some individuals or situations in a particular way. We will see in Chs 4 and 5 that along with descriptive concepts, with which we seek to capture features of some item, there are also prescriptive concepts that have an essential tie to action, and that some concepts have both a descriptive and a prescriptive dimension. Often a set of concepts is promoted as a description, but with aim of affecting action. The feminist objection to describing body parts while leaving out the women who possess them is an objection to the kind of attitudes and behaviors that advocates of these descriptions seek to promote.

The existence of extra fertilized eggs raises another contentious issue: the appropriateness of using these eggs for research. I am not going to enter into this ethical minefield in any detail, but I do want to note one aspect that is relevant to my concerns in this book. Several countries have formed commissions to investigate and make recommendations. This involves consideration of the various stages in the development of the fertilized egg, along with consideration of when it should be counted as the beginning of a distinct living being.[60] Several of these committees recommend that research is ethical up until 14 days when "the primitive streak, the first indicator of the embryo's body axis" appears (S 6, cf. Shevory 1992: 232–33). A description of the stages of development requires new concepts, and variations on older concepts that are of recent vintage.

I have offered only a sample of the legal and social issues raised by reproductive technologies; others issues have arisen and more are likely to arise while I write, and while you read. As Rowland wrote in book that appeared in 1992:

> In the time that passes between the writing of this book and its publication, technology will have continued to move us further and further away from the ideas and understanding of reproduction which were the basis of society before the intervention of reproductive and genetic engineering. We need a framework into which to place these changes, a framework based on understanding of how power works and who gains from this technology, a framework to help us form sound judgments about its usefulness and morality.
>
> (RD 201)

Note Rowland's inclusion of further issues that I have not discussed, issues that derive from feminist concerns and suggest a further enlargement of the range of concepts that are considered relevant for thinking about these topics.

Issues of comparable novelty and complexity are posed by other medical technologies, such as organ transplantation and the ability to extend life. They raise, among others, the question of appropriate criteria for death. Brain death – the absence of brain waves – is a widely used criterion in some countries. Accepting this criterion allows us to classify a person as dead while vital parts of the body are kept functioning in order to permit harvesting organs or the delivery of a fetus. This criterion was long resisted in Japan where the connection between life and a beating heart is a deeply held tradition. A 1997 law established brain-death as a legal criterion and led to the first Japanese heart transplant since 1968 (*Chicago Tribune* March 1, 1999, Sec. 1: 4). This criterion depends on our understanding of electromagnetism, viewing the brain as a source of electromagnetic radiation, and development of technologies that allow us to detect this radiation. All of this would have been quite unthinkable in, say, the early nineteenth century. In addition, brain-death is an especially tricky criterion in the case of anencephalic babies who do not exhibit brain waves, but who have a brain stem that controls heartbeat and respiration. Such babies do not live very long, but the question became a legal issue when the US parents of such a baby wanted to donate its organs for transplantation (FK 27–33).

Conceptual issues arise from other new technologies as well, and have been arising for substantial periods of time. In the case of computer technology, for example, issues have included whether a program embodied in ROM chips falls under the laws of copyright or patent protection, the ownership of documents and music published on the web, the ease with which photographs can be altered by widely available hardware and software, new

privacy issues, and much more. Many of these issues require new ways of thinking that, in turn, require the development of new concepts and creative adaptations of older concepts.[61]

The discussion in this section brings us into a realm of current debates, and differs in this respect from the historical studies above. Still, reflection on those studies should underline the point that the confusions and tensions I have been considering are quite normal as we face new situations in which existing conceptual frameworks become inadequate. We need a theory of concepts in order to understand how these changes have taken place, and to understand the range of possibilities that are currently on offer.

## 2.4 Philosophical Concepts

> Philosophy may perhaps be the chaste muse of clarity, but it is also the mother of hypotheses.
>
> (SM 12)

By *philosophical concepts* I mean concepts that play a theoretical role in philosophy. Familiar examples include standard distinctions such as those between analytic and synthetic propositions, norms and descriptions, and teleological and deontological systems of ethics. Other examples include the concepts of a final cause, a transcendental argument, and a self-justifying proposition. Since philosophical studies often overlap with other disciplines, many concepts become integral to philosophical theorizing in particular contexts. Examples include the concept of a scientific theory, an observation, and a political system. The entire array of logical concepts – deduction, validity, necessary condition, evidence, and so on – may also be considered philosophical concepts in some contexts. My concern in this section is to illustrate how developments within philosophy, along with developments in fields that become subjects of philosophical scrutiny, can lead to changes in our philosophical concepts.

As a first example consider EMPIRICAL EVIDENCE as used in philosophy of science. Two themes are central to this concept: this is the crucial kind of evidence for the epistemic evaluation of non-analytic propositions, and it has something to do with the use of our senses. But our understanding of what counts as empirical evidence and its exact epistemic import has varied with different philosophical positions, and with the development of science. Consider some of these variations, beginning with alternatives that are (relatively) internal to philosophy.

There are three main types of philosophical theories of perception, with many varieties of each: direct realist, indirect realist (also known as representationalist), and phenomenalist. Perhaps the oldest version of direct realism is due to Aristotle who held that our senses provide direct knowledge of properties of physical objects exactly as those properties exist apart from

our awareness of them. The detailed development of this view depends on a metaphysic that distinguishes between an object's form and matter, and holds that the same form can occur in different bits of matter. The view also requires a particular theory of mind. In *De Anima* Aristotle argues that our minds are forms of our bodies, but they are forms in which other forms can be instantiated. When I perceive a physical object, the form of that object is instantiated in my mind. Moreover, mind has no internal structure that in any way distorts the forms that come to exist in it, and all knowable properties of an object are embodied in its form. Thus the immediate object of perception is an instance of the object's form in my own mind, and perception provides undistorted access to the properties of physical objects. Theories which hold that the mind has an internal structure that is implicated in what we perceive provide an important contrast. Kant was the key figure in developing this kind of theory, along with the consequence that perception provides knowledge only of things as they appear, not as they are in themselves. I will not pursue the details of Kant's theory of perception here, but I note that it does not fit neatly into the trichotomy of theories that I am discussing.

Many contemporary philosophers are direct realists of a different sort. Their main concern is to deny that there is any entity that stands between the perceiver and the physical object perceived. In one common version, physical objects cause our perception and we perceive those physical causes. Proponents of this view recognize the existence of illusions and other forms of misperception – a subject that is sorely neglected in the Aristotelian tradition – and thus do not hold that we always perceive physical objects exactly as they are. Rather, the emphasis in this form of direct realism is on the thesis that perception is a two-term relation involving only a perceiver and a physical object. The procedures by which we come to learn the nature of that object may be quite indirect (Brown 1992a).

Indirect realism shares the realist aspect of direct realism in holding that perception is caused by the action of physical objects on our senses. The characteristic difference between the two views is that indirect realists consider perception to be a triadic relation between a physical object, a perceiver, and some intermediate item. A typical version holds that a physical object acting on an organism causes an intermediate item that we perceive directly. On one classic version the immediate objects of perception are *ideas*, which are mental entities caused by the physical interaction between external objects and our senses. The properties of a sensory idea depend on properties of both the physical object that initiates the causal process and the organism on which that object acts. As a result, an idea may or may not mirror properties of that physical object. Locke called those ideas that mirror external properties "ideas of primary qualities," and those that do not "ideas of secondary qualities." Since physical objects cause ideas of both types, both are a potential source of information about those objects, but any resulting knowledge is indirect in two respects. First, it requires

argument to determine which type of idea we are perceiving. Second, if we are dealing with an idea of a secondary quality, it takes further argument to extract information about the physical object that initiated the process. (See Brown 1987, Ch. 6 for a more detailed account.) Seventeenth-century thinkers as diverse as Boyle, Descartes, and Galileo held versions of this view. Indirect realism has been quite unpopular among twentieth century philosophers although some versions have been put forward – e.g., Russell (1948); Sellars (P); Wright (1977, 1985), and others (see Wright 1993). Historically, the most important objection to indirect realism focused on the third item that supposedly stands between the perceiver and the object perceived, which has been viewed as an impediment to knowledge of the physical world. In twentieth-century-English-language philosophy much emphasis has been placed on the claim that our everyday perceptual concepts support direct realism. Many indirect realists can happily concede this claim since the arguments for their view are empirical. They seek to replace common perceptual concepts with concepts that, they hold, embody a more accurate account.

The development of phenomenalism began with Berkeley, and versions of phenomenalism dominated English-language philosophy of perception in the first half of the twentieth century. Contemporary versions of direct realism were largely developed in opposition to phenomenalism. Typically the central thesis of phenomenalism is that the immediate objects of perception are mental entities – ideas in Berkeley's version, sense data in twentieth-century versions – which provide all of our information about physical objects. Phenomenalists agree with direct realists in holding that perception is a binary relation, but disagree about the immediate objects of perception. Phenomenalists agree with indirect realists that the immediate objects of perception are internal to the perceiver, but reject any notion of a transcendent cause of perception. All intelligible talk about physical objects and their causes must be reducible to talk about what we will perceive under various conditions. Question about the ultimate source of ideas or sense data are rejected as confused. In twentieth-century terminology, physical objects are *logical constructs* out of sense data. In philosophy of science phenomenalism leads directly to instrumentalism (although this is not the only path to this view) – the thesis that the sole aim of science is to provide means of predicting what we will perceive under various circumstances. Science does not aim to discover the nature of a world that transcends all perception. Berkeley held that no such world exists; other phenomenalists take an agnostic position on the existence of such a world, contending only that if it does exists, we cannot know anything about it; others argue that claims about such a world are, strictly speaking, meaningless. (Berkeley gives arguments for all three views; cf. Brown 2000a). The last version was typical of twentieth century phenomenalism and was associated with a particular theory of meaning – and thus of concepts; I will return to this topic in Ch. 3.

Now consider some reasons for thinking that various philosophers are associating different concepts with perception words. We have seen that "I see *x*" has different implications for different views. For some philosophers this sentence implies that *x* exists apart from anyone's awareness of it, for others it has no such implication. On some views the sentence implies that *x* has exactly the properties it appears to have, on other views it does not imply this. Some versions require a particular metaphysic to understand the implications of our sentence, other versions invoke a different metaphysic, and proponents of yet other versions claim that no metaphysic is involved. But one standard question that arises in debates over concepts is the extent to which differing implications indicate different concepts. On some theories of concepts, these debates are (in part) debates over the conceptual content associated with perception language. In addition, seeing is a dyadic relation on some views, a triadic relation on others. Moreover, our interpretation of this sentence has implications for the nature and aims of scientific research, and thus for the import of empirical evidence.

Whatever account of perception we adopt, it is widely agreed that perception provides the empirical evidence on which all scientific beliefs are ultimately evaluated. Yet our understanding of the exact role of perception in providing this evidence has been altered by scientific developments. Consider, in particular, the process of extending the range of our senses that began with the introduction of the telescope. I will carry on this discussion in terminology that scientists would typically use, recognizing that various philosophical considerations – such as a commitment to phenomenalism – would require a major rewriting of this account.

The telescope places an instrument into the physical process that intervenes between distant objects and our eyes, with the result that this process is altered. In this regard the telescope differs from instruments such as a meter stick or an astrolabe, which just add an item to our visual field. This new feature yields two key results: it allows us to study items that we cannot detect with our unaided senses, and it provides new information about items that are detectable by our senses. Galileo's discovery of four moons of Jupiter illustrates the first situation; examples of the second type – such as the phases of Venus and the resolution of the Milky Way into distinct stars – had a deep epistemological significance at the time because they were in direct conflict with naked eye observations. Since the prevailing view of perception in Galileo's day was Aristotle's, this conflict raised the question of why we should accept telescopic results, which might be caused by distortions introduced by the telescope.[62] Galileo's response was to argue that there are intrinsic defects in our eyes that distort what we see when we look at small, bright, distant objects – and that the telescope corrects these defects (Brown 1985). This involves a new view of the role perception plays as a source of empirical evidence.

Telescopic observation was the first step along a path that has been extremely fruitful: recognition that the world is full of items that we cannot

detect with our senses, but that we can study by interposing instruments between those items and our senses. These instruments interact with the items we would study, and yield outputs that we can sense. The innovation is particularly striking when we use instruments to study items to which our senses do not respond at all. The magnetic compass, an early instrument of this kind, allows us to see the direction of the earth's magnetic field even though we have no senses that respond to magnetism. Probably the next major steps in this direction occurred at the beginning of the nineteenth century with the discovery of infrared radiation by Herschel, followed by Ritter's discovery of ultraviolet radiation. The story opens with Herschel's study of the spectrum of light from the sun. Herschel was working with different colored filters and he noticed that heat and light sometimes occurred together, but that he sometimes felt a sensation of heat with little light, and sometimes light with little heat. This led him to explore the association of heat with light in some detail. In one set of experiments he used a prism to break sunlight into its spectral colors, and thermometers to measure the temperature in different colors – and at measured distances beyond the edges of the visible spectrum. Herschel found that the temperature was greater towards the red end of the spectrum, continued to rise for a distance beyond the red end, reached a peak, and dropped off. Measurements at the violet end of the spectrum showed that "the power of heating is extended to the utmost limits of the visible violet rays, but not beyond them; and that it is gradually impaired, as the rays grow more refrangible" (Herschel 1800: 291). Herschel concluded that light and radiant heat are the same; that what we call "light" is just that part of spectrum that our eyes detect; and that "the invisible rays of the sun probably far exceed the visible ones in number" (1800: 291–92). However, later studies of the transmission of heat and light led him to doubt this conclusion (Hacking 1983: 177–78).

Ritter became interested in possible rays beyond the violet edge of the spectrum after reading Herschel's paper.[63] Ritter knew that "hornsilver" (silver chloride) darkened in the presence of light, and darkened more intensively in light towards the violet end of the spectrum. So he dampened a strip of paper with hornsilver and placed the strip in the spectrum from sunlight projected in an otherwise darkened room. The strip quickly darkened, especially in the violet and beyond, allowing Ritter to conclude that the radiation continues in this direction too (Guiot 1985; Wetzels 1990).

Empirical studies that take us beyond the limits of our senses have become standard in late twentieth century science. One indicator of the range of these developments is the need to attach adjectives to "telescope." We now have radio telescopes, X-ray telescopes, and neutrino telescopes which do not operate on the electromagnetic spectrum. To these we can add electron microscopes, Geiger counters, the much more complex systems of detectors at high-energy physics laboratories, and many others.[64] Indeed, the results we examine are often – and to a growing degree – processed by

computers before any person examines them. Our senses remain central to this process in that the output from our instruments must pass through our senses in order to become epistemically relevant to us. But the specific qualities we experience are irrelevant to the content of this information. Just as we can sometimes study a single object using different senses, so a computer can produce a visual, auditory, or Braille output. The information carried by this output is independent of the sensory modality used. This stands in stark contrast to the common empiricist thesis that different sensory modalities provide different information. In general, the requirement that any information we can use must pass through our senses is a pragmatic constraint on our instruments. But it is the information embedded in these outputs that provides the empirical evidence we use to evaluate scientific theories; the sensory modality in which this information appears is epistemologically irrelevant.[65]

For the moment I want to draw one key point out of this discussion. The thesis that claims about the physical world must be evaluated on the basis of empirical evidence is a constant feature of science, but we find quite different accounts of the nature of this evidence among those who accept the general principle. For Aristotelians such evidence is just the information about the world that we gather with normally functioning senses. For the seventeenth-century philosophers and scientists who broke with the Aristotelian tradition, our senses do not always accurately show us what is in the world, and argumentation is required to distinguish which of our percepts are reliable. As philosophical empiricism developed into phenomenalism, specific details of our ideas or sense data became central. As a result, for Berkeley and many later empiricists we cannot literally perceive the same things by means of two different senses. But as science discovered that there is much more in the world than we can sense, and developed ways of gathering information about these items, it provided challenges to some of these philosophical views. This, in turn, required elaboration of our understanding of what counts as empirical evidence. Many older empiricists took it as basic that our senses provide indubitable evidence, but once we introduce instrumentation into the evidence-gathering process, our evidence becomes less certain. One way of seeing the point is to ask what can be more certain than seeing the numeral 5 on a digital read-out. Yet even if we agree that we are not likely to get this wrong, the story changes once we report "5 ohms," or "5 miles per hour," or "5 neutrinos in the last 24 hours." Now the accuracy of our report depends on the accuracy of our understanding of the instruments we are using, and it becomes possible to challenge an evidence report by challenging this understanding. Meanwhile, at least since Galileo, we have been learning about how our senses operate, where they are reliable, and how to improve their reliability (Brown 1985, 1987). As we take these factors into account, we shift the concept associated with the phrase "empirical evidence" and thereby change our understanding of the epistemological significance of such evidence. A theory of concepts should provide the tools for a more detailed under-

standing of the nature of these conceptual changes, and of the relations between successive concepts in this philosophical domain.

Consider next a related concept, KNOWLEDGE. I will discuss this concept in greater detail in Ch. 8. For the moment I want to focus on an aspect of this concept that has been central to epistemology at least since Plato. Unger (1975) describes "knowledge" as an "absolute term": KNOWLEDGE, like FLAT, does not admit of degrees. Just as a surface must meet a definite set of standards to be flat, and anything that fails to meet that standard just is not flat, so a belief must meet a set of standards to count as knowledge. These are very demanding standards and Unger concludes that we have little knowledge. He also urges the introduction of new concepts that will free us from having to accept such radical skepticism (317–18). I want to consider this suggestion, but it will be useful to approach the matter with some historical perspective.

In *Theaetetus* Plato posited infallibility as one of the conditions for knowledge, and Descartes' version (captured in the demon) has dominated modern philosophy. Since we are seeking infallibility, any consideration that shows it possible that we may be mistaken is sufficient to require that we put a claim aside as not known – subject to later reconsideration once we become clear on the means we have for achieving knowledge. When Descartes introduces the demon into his *Meditations* he has already used familiar illusions to argue that perceptual beliefs are fallible. The function of the demon is to cast doubt on beliefs arrived at by pure reflection, including beliefs in simple truths such as that equals added to equals yield equals. The demon works on our minds and causes us to believe that true propositions are false, and that false propositions are true. But to understand what Descartes was up to we should keep in mind that he did not consider it particularly difficult to defeat the demon. He believed he had accomplished this task by the end of his third meditation, and done so in a way that raises a problem about how we can ever fall into error – a subject he addresses in the fourth meditation. Once the demon has been defeated, Descartes proceeds to reconstruct the body of knowledge, establish a range of claims that meet his tough demands, and show the epistemic limits of beliefs in other domains. If Descartes had been right, and we can achieve knowledge in this very strong sense (even in the limited realms in which he thought it possible), then knowledge, so conceived, would be well worth pursuing. But it is worth noting that Descartes did not consider infallible beliefs to be the only worthwhile cognitive goal. He also held that there are cognitively defensible beliefs in realms in which infallibility is not possible – and pursued such results at length in his scientific writings. In other words, Descartes recognized that our choice is not simply between knowledge and ignorance, but that between these poles there is a great deal of room for better or more poorly founded results.[66]

This intermediate realm becomes especially important if we hold that knowledge – in the strong Cartesian sense – is, at best, a rare phenomenon.

Then one key concern of epistemology is to understand the grounds for reasonable, although fallible, acceptance of claims. Pursuit of this project may require the replacement of some established epistemic concepts and the introduction of new concepts – as Unger suggests. But there are reasons for thinking that this process has been ongoing for some time. Consider, for example, inductive justification. Many philosophers have held that if inductive justification is legitimate, it is just as certain as deductive justification. Thus Hume was able to cast doubt on the rationality of inductive justification just by noting that in such cases it is always possible for the premises to be true and the conclusion false. Mill, thinking along the same lines, but from a more optimistic perspective, held that there are "certain and universal inductions; and it is because there are such, that a Logic of Induction is possible" (1868: 359). Of course we make mistakes in inductive inferences – as we do in the deductive case – but the occurrence of mistakes does not prove that no certain inductions are possible. Thus, commenting on a famous case, Mill wrote: "That all swans are white, cannot have been a good induction, since the conclusion has turned out erroneous" (184). Others, however, take the characteristic feature of induction to be exactly the fact that it can provide good reasons for accepting a conclusion even though it is possible that all of our evidence statements are correct while the conclusion is false. Understanding the nature of this support is a central research project in inductive logic, and there is no reason to think we can pursue this project by analyzing concepts we already have.

Reflection on the nature of justification raises another issue. It is widely held that justification is one necessary condition for knowledge. But it is also widely recognized that justification comes in degrees, and this suggests that we may want to think of knowledge as being susceptible of degrees, not as an absolute concept after all. That there already is some basis for such an approach is suggested when we note that it is not obviously outrageous to claim that I know A better than I know B.

How we respond to examples of this sort depends on our views on a number of issues concerning concepts. Analytic philosophers commonly assume that there is a single concept of knowledge that "we" all share, and that disputes among philosophers about the analysis of this concept arise because it is difficult to formulate its necessary and sufficient conditions. But other explanations are possible for the failure to arrive at an agreed analysis. One is to hold that we do not all associate the same concept with the word "knowledge." On this account philosophers who engage in arm-chair analysis of their own concepts are actually much better at the task than the debates in the literature suggest – but they are not all analyzing the same concept (Brown 1999). Another possibility is that rather than just analyzing a given concept, we are attempting to forge a concept that is appropriate for our epistemic situation. This will lead to an ongoing project because (as in the case of scientific research), when we learn more about our epistemic capabilities and limits, we may well find that previously available concepts

are not adequate, and must be replaced. In addition, as we learn more and develop new techniques, older epistemic concepts may become inadequate because our actual epistemic situation changes.[67] On this view it is unimportant whether we retain the word "knowledge" for some particular epistemic concept. If it turns out that the term generates confusion because it is associated with many different concepts, or carries baggage that we wish to discard, then we would have pragmatic grounds for dropping the term. There is no guarantee that any of the concepts that have been associated with this term will continue to play a significant role in epistemological theorizing – just as there was no guarantee for such concepts as natural place, phlogiston, or telegony. I will return to these questions in Chs 7 and 8, after my preferred theory of concepts is in place.

I want to introduce one more example – Quine's (1953) attack on the analytic-synthetic distinction – although I will postpone detailed discussion until Sec. 3.7. For the moment, note the fundamental nature of this attack. On the prevailing view analytic propositions express meanings of terms, synthetic proposition use those terms to express factual claims. Analytic philosophers typically hold that the special domain of philosophy is a priori knowledge, and most practitioners of analytic philosophy are empiricists who hold that all a priori knowledge is expressed in analytic propositions. Thus if the concept of analyticity is incoherent, as Quine maintains, the set of concepts that provides analytic philosophers with their customary understanding of their discipline is undermined. So Quine is attacking the self-conception of analytic philosophy, and doing so by challenging the coherence of the system of concepts in which this self-conception is expressed. Those who accept this challenge can either abandon the practice of philosophical analysis, or reconstruct it on the basis of a modified set of philosophical concepts.

## 2.5 Some Forms and Generators of Conceptual Change

> Terms in scientific theories do not have static meanings, but are defined and redefined within the context of their evolving usage.
>
> (Cushing 1990: 35)

The immediate lesson I want to draw from these case studies is that conceptual change is a common feature of our cognitive history. I have considered only a small selection of fields in which conceptual change can be studied; those who feel that their fields of interest have been neglected are invited to add further studies. In this section I want to draw together several strands of this discussion by highlighting some of the situations that motivate conceptual change and some of the kinds of conceptual changes that occur. My remarks on these two issues will overlap, although there will not be any direct correspondence between generators and forms of change, nor will I attempt to provide a comprehensive account.[68]

Often conceptual innovation is a response to an empirical discovery. The case of sperm is fairly straightforward: microscopic study reveals entities of an unanticipated type, and a concept is introduced with those entities as instances. The cases of X-rays and radioactivity were also empirically driven, but the observed phenomena are not instances of the new concept. Rather, the concept refers to the cause of these phenomena. Isotopes illustrate another response to unanticipated empirical evidence: introduction of a concept into an explanatory scheme that accounts for the evidence. These cases also illustrate how the attempt to assimilate new data sometimes involves a period of uncertainty and further research before scientists work out the appropriate concepts. For example, when Becquerel discovered radioactivity he thought he was studying photoluminescence; the Curies continued to think of radioactivity on this model for some time. When Rutherford and Soddy introduced METABOLON, a concept that did not survive, they believed they had discovered a new form of matter. These cases were later classified as short-lived isotopes, but this could occur only after the concept of an isotope was introduced about a decade later. Moreover, our current understanding of isotopes was not achieved until two decades after the initial version of this concept was proposed.

Introduction of a new concept does not always require a new empirical discovery. Sometimes a new concept is introduced in the course of constructing a theory that provides a better account of existing data than is provided by older theories. Newton's gravitational theory is one such example, and his concept of mass (as distinct from weight) was introduced because he needed a concept to do a job that was not seen as necessary by earlier theorists studying the same phenomena. I will discuss the notion of the job that a concept does in Ch. 4, and mass in Sec. 9.4.

Another powerful motivator of conceptual change is the discovery of an internal inconsistency in a theory. It is far from obvious that the classical concept of a set is inconsistent; it took a subtle argument by Russell to demonstrate the problem. One consequence of this discovery is the contemporary distinction between *classes* and *sets*, where a class is the more general notion and a set is a mathematically well-behaved class. By way of contrast, Bohr's theory of the atom was known to be inconsistent from its inception. The inconsistency arose because Bohr introduced the thesis that certain electron orbits are stable into a framework based on classical electrodynamics, which entails that no such orbits are stable. The inconsistency was eventually removed when the entire theory was replaced by quantum theory, which involves several conceptual innovations.

Sometimes an inconsistency arises because of external developments. We saw that California law recognized only one natural mother per child, but considered both blood tests and the fact of giving birth as sufficient grounds for establishing natural motherhood. IVF created situations in which the two tests give conflicting results. Our exact account of this case will depend on the theory of concepts we adopt. For example, if we distinguish between

conceptual content and criteria of application, this case may not involve *conceptual* change; if we include criteria of application in the content of a concept, it does. On an operationist view of concepts, as originally conceived, there never was a single concept of a natural mother, but two concepts and an ambiguous term. Before the development of IVF this ambiguity could be ignored in practice, but the new technology changed the pragmatic situation. It forced us to attend to this disparity, but had no impact on the concepts involved. I will return to these topics in the chapters that follow.

Consider another case whose exact treatment will depend on the theory of concepts we adopt. Sometimes we encounter a situation in which we have both an explicit analysis of a concept and a set of paradigm instances of that concept, and we are equally confident of both. On some theories of concepts, both contribute to conceptual content. But new developments may result in a clash between the analysis and the paradigm instances. The concept science will illustrate the point. For Kant it was equally central that a science is a subject that has achieved permanently established foundations, and that Newtonian physics is a science. Yet work in physics since Kant's time has shown that we cannot have both of these: we must either recognize that the foundations of a science are subject to re-examination, or that Newtonian physics fails to be a science. At present most philosophers are more confident in the scientific status of Newtonian physics than in any analysis of what counts as a science. Indeed, the latter subject is now in flux exactly because attempts to formulate criteria for scientific status often have consequences for specific cases that are considered unacceptable. For example, falsificationist accounts of science reject Kant's demand for established foundations, but there are different forms of falsificationism with different consequences for the scientific status of various subjects. If we read falsificationism as demanding that any theory that faces an anomaly must be rejected, then there are no sciences. If we relax this rigid demand it becomes much harder to agree on which subjects count as science.[69]

I will consider other generators of conceptual change as we proceed through this section, but I want to begin looking at some of the forms that conceptual change takes. Concepts provide criteria of classification (this is not their only function) that allow us to organize items into classes whose members are treated as identical in certain contexts and for certain purposes. Conceptual change may lead to the reorganization of items and of systems of classification. Consider a concept that can apply only to a single item (a *unit concept*), such as the Aristotelian concept THE CENTER OF THE UNIVERSE. One kind of revision occurs if we change our view of the item that meets this description. Thus Kepler maintained that the sun, not the earth, is the center of the universe, and Newton held the center to be the center of mass of the planetary system. (Again, whether this involves *conceptual* change will depend on one's theory of concepts.) A more drastic change occurred when scientists concluded that there is no center of the

universe and the concept ceased to play any role in physical science. Changes of the latter sort also occur for concepts that admit multiple instances, such as phlogiston, telegony, and radioactive induction.

Consider now a different kind of situation: a classification is rejected, but the existence of the item or items that fall under that classification is not in doubt. In pre-Copernican astronomy EARTH and SUN were unit concepts that were eliminated in later astronomy, but no doubts arose about the existence of the items that had constituted these classes. The earth was moved into the same class as the planets, and the sun into the same class as the stars. Since the Ptolemaic concept of a planet explicitly excluded the earth from its extension, reclassification of the earth as a planet altered the concept of a planet. The early concepts SUN and STAR also treated these as distinct classes; the Copernican reclassification resulted in a new class whose members have characteristics derived from each of the predecessor classes. As is common, we retain the words "sun" and "star" because of their clear referents. Thus we can say that the sun is a star; this is nonsense in an Aristotelian or Ptolemaic framework.

The sun/star case illustrates one common way of generating a new class: via the set theoretical union of two classes that were previously treated as mutually exclusive. Whether this regrouping involves conceptual change – and if so, the degree of conceptual change – varies among cases. At one extreme, we can disjoin any two classes – say tables and planets – and associate a term, "T," with this new class. "T" will have a simple analysis as "either a table or a planet," and might yield a useful abbreviation, but have no theoretical interest. But theoretical import can arise because such groupings often involve a decision to ignore – in certain contexts – some of the features that distinguish members of the predecessor classes. SIBLING and SPOUSE may seem to be trivial re-groupings of this sort, but this is not quite right because describing someone as a sibling or spouse treats gender differences as irrelevant. Sometimes gender is important. Historical cases include royal succession and control of property; contemporary cases include diagnoses of endometriosis and prostate cancer. Whether the appropriate concept to use in a given case is wife or spouse may be a highly contested issue, and proposals to deploy one of these concepts rather than the other may be part of a substantive social agenda. A hereditary monarchy in which succession passes to the eldest sibling is structured differently than those in which it passes to the eldest brother. Similar points apply to PERSON. The claim that every person counts equally for moral or legal purposes is a substantive claim that has been resisted in societies that countenance slavery or a variety of forms of gender discrimination. Those who advocate such distinctions can adopt either of two options in attempting to press their view: They can reject a general principle such as "All persons are equal before the law," or they can accept the verbal formulation while excluding specific individuals or groups from the class of persons. Advocates of each strategy associate a different concept with the word "person."

Now consider cases in which items that were once considered *the same* are moved into different classes because differences are found that are relevant in a particular context. As Putnam noted, an instance of this type occurred when chemists recognized that "jade" refers to two different chemical compounds (1975: 241): jadeite is a silicate of sodium and aluminum; nephrite is a silicate of calcium and magnesium. Since jade is an ancient concept, a considerable body of chemical concepts had to be developed before this distinction could be recognized. The discovery of isotopes was another development of this sort with considerably greater theoretical significance. Most elements occur as multiple isotopes, beginning with three isotopes of hydrogen, and variations in behavior can be considerable. For example, $H_2O$ in which the H is deuterium is toxic to humans. (While the same chemical reactions occur, the rates at which they occur are different.) Thorium provides a subtler example: it has more than twenty-five known isotopes, all radioactive, with half-lives varying from microseconds to billions of years.

In the cases we have just considered there is a kind of uniformity in the results of the subdivision. Jade, which was once considered a single mineral, is now recognized as two minerals. Isotopes involve greater variety, but we (now) have a uniform account of what constitutes an isotope, and in this sense the various isotopes are all of the same kind. Other cases yield more heterogeneous results. The ancient Greek concept of an element encompassed five instances; each considered a single, uniform type. As a result of subsequent research the elementary status of all of these was rejected, but different members of this class were treated in different ways. Ether (as understood by the Greeks) was rejected as non-existent; fire exists but has no place in twentieth century notions of elementary matter; water is now viewed as a compound and is homogenous in chemical contexts where isotopes are not important; earth and air are heterogeneous mixtures of several components.

Another type of conceptual change occurs when established boundaries between classes break down. The reclassification of the earth, planets, sun, and stars provide familiar examples of this sort; consider some variations on this theme. One important case that is empirically motivated occurs when items are found that overlap classes previously assumed to be mutually exclusive. Viruses and euglena (which are both mobile and photosynthesize) are well-known examples. The Australian animals that came to European attention in the late eighteenth century had a similar effect. This is especially true for the monotremes – the platypus and echidna:

> The fact that these animals juxtaposed incontestably mammalian characteristics like hair and warm blood with others previously identified only with birds and reptiles forced naturalists to consider whether some quadrupeds were intrinsically more mammalian than others. And the systematic oddity of the Australian fauna was, in a sense, contagious. Redrafting the boundaries of a previously well-defined category was not

> necessarily a matter of simple expansion; new proximity to external classes potentially shifted all internal relations too.
>
> (RI 12)

The diversity of the biological world seems to regularly thwart our attempts at neat classification. For example, shiner perch have live births (Judson 2002: 45) while mangrove fish can survive more than two months out of water and move across land (182). (See also 187–93 for discussion of sex cells – such as ova and sperm – in species that have considerably more than two kinds.)

In discussing disjunctive concepts we encountered cases that involve forming a new concept by abstracting from some features of the original disjuncts; this results in the introduction of a more general concept than either of those with which we began. In Sec. 2.2 we encountered three forms of generalization mathematics that involve conceptual change. I want to review and extend that discussion.

One type of mathematical generalization involves introducing a new concept such that a subset of its instances is isomorphic to the instances of its conceptual predecessor; number systems and the gamma function provide examples. In the latter case Euler's generalization uses concepts from calculus that were not available a few decades earlier, and that are not now in the repertoire of many people who are quite capable of understanding factorials. Exponents introduced a second form of mathematical generalization that yields the same relation between instances of the new and older concepts as the previous case, but proceeds differently: An established mathematical structure is reinterpreted as involving implicit limitations that may not have been apparent to those who used that structure in the past. Generalization then proceeds by relaxing those limitations. Extensions of the realm of geometry from Euclidean geometry of two and three dimensions to both Euclidean and non-Euclidean geometries of any number of dimensions provide a further illustration worth exploring.[70]

Recall how the distance between two points is calculated from a set of Cartesian coordinates. I will use $x$, for the distance between the $x$-coordinates of the two points, and similarly for the other coordinates. I will also consider the square of the total distance to avoid square-root signs. Applying the Pythagorean theory for the 2D and 3D cases, respectively, we get:

$$d^2 = x^2 + y^2, \tag{G1}$$

$$d^2 = x^2 + y^2 + z^2. \tag{G2}$$

One way of thinking about the relation between these expressions is to view G1 as a special case of G2. This can be implemented by considering each term on the right hand side of G2 as having a coefficient that is limited to the values zero and one. G1 is then the special case in which the coefficients

of $x^2$ and $y^2$ are one and the coefficient of $z^2$ is zero. The formal extension to any number of dimensions is now straightforward: G2 is a special case of a much longer formula in which the coefficients of $x^2$, $y^2$, and $z^2$ are one, and all other coefficients are zero.

We take a step toward non-Euclidean geometry by removing the requirement that coefficients be only zero or one; an additional generalization takes us to a much richer array of geometries. Consider the concept of a *quadratic form*: the most general quadratic expression that can be built out of a set of parameters. In addition to the square of each parameter, we include the products of all possible pairs of parameters. The generalizations of G1 and G2, including coefficients, are:

$$d^2 = Ax^2 + By^2 + Cxy, \tag{G3}$$

$$d^2 = Ax^2 + By^2 + Cz^2 + Dxy + Exz + Fyz.^{71} \tag{G4}$$

The values of the coefficients are sufficient to characterize a geometry, and the extension to higher dimensions is straightforward. The Minkowski geometry of special relativity has, in these terms, the following distance rule (after converting the time dimension to spatial units):

$$d^2 = x^2 + y^2 + z^2 - t^2. \tag{G5}$$

In other words, it is a 4D geometry in which the coefficients of $x$, $y$, and $z$ are one, the coefficient of $t$ is minus one, and the coefficients of the three mixed terms are zero.[72] This is a non-Euclidean geometry because of the negative sign on $t$, although it is a flat space.[73]

In discussing logarithms I noted a third kind of generalization that I want to develop further: We take a property of a mathematical structure as a defining feature of a more general structure, and the original structure becomes a special case. Consider two further examples. First, Euclid took the concept of a straight line to be clear, and the thesis that a straight line is the shortest distance between two points to be an intuitively correct postulate. With the introduction of non-Euclidean geometries the shortest distance between two points became a feature of the specific geometry. The concept of the shortest distance was generalized to the concept of a *geodesic*, and a straight line became a special case for a particular geometry.

Second, we begin with the concept of a vector as it is usually learned in elementary mathematics and physics. One useful property is the *scalar product of a pair of vectors*, defined as the product of the lengths of the two vectors multiplied by the cosine of the angle between them. The scalar product of two orthogonal vectors is zero. For our purposes the key feature of the scalar product is that it maps a pair of vectors onto a scalar. Mathematicians have generalized the concept of a vector into that of a vector space. This more general structure is specified by a set of axioms,

includes elementary vectors as a special case, but also includes a wide array of structures that do not look like vectors from an elementary perspective.[74] As part of this generalization the concept of a scalar product is generalized to the concept of an *inner product* which is *any function* that maps a pair of vectors onto a scalar. A vector space need not include such a function, so the notion of an inner product becomes the defining feature that distinguishes two classes of vector spaces. Moreover, when there is an inner product, having an inner product of zero is taken as the defining feature of orthogonal vectors; the original orthogonal vectors are now a special case. These discussions of geometry and vector spaces follow both the historical process and a common pedagogical sequence. As a result, contemporary students go through the same process of conceptual change that the mathematical community encountered at an earlier time.

## 2.6 Some Philosophical Issues

I want to end this chapter by highlighting some issues that will have occurred particularly to philosophers. These issues will have to be addressed by any theory of concepts and are noted here for future reference.

First, consider the relation between conceptual change and change of belief. It seems eminently reasonable to maintain that competing beliefs can be formulated in terms of the same concepts. People can disagree on the breed of a particular dog, or on the number of dogs in the next room, while using the same concepts of dog, room, number, etc. Moreover, a single individual can be unsure which of two (or more) incompatible propositions to adopt while working in a single conceptual system. Thagard argues that the distinction between change of concept and change of belief is a matter of degree: they shade into each other so that "It would be futile to try to offer criteria for identity of concepts that attempt to specify when a concept ceases to be the concept that it was" (1992: 34). But, as van Fraassen noted in a discussion of the theory/observation dichotomy, the fact that instances of two concepts occur on a continuum with a gray area where they meet is compatible with the existence of clear instances of each (1980: 13–14). The existence of a gray area where belief change and conceptual change merge does not eliminate the need to address the distinction in clear cases – if there are such cases.

However, while a theory of concepts must come to terms with this issue, there are different ways in which it might do this. One approach is to take the distinction as a test case for theories of concepts, and reject any theory that fails to get the distinction right. But suppose we have a theory of concepts that fails this test while passing other tests and that is, on balance, the best available theory. Two other responses besides outright rejection are available. We can consider the specific failure to be an anomaly for our theory, and thus a reason for seeking a better theory, but still take the best available theory as a basis for thinking and research

about concepts until a better theory is developed. Or, we can consider the overall success of a theory to be a reason for rejecting an isolated claim that clashes with it. That is, an otherwise successful theory of concepts might provide a reason for rejecting the view that the distinction between change of concept and change of belief is fundamental. Some philosophers will respond that this distinction is intuitively correct, but such intuitions are not sacrosanct. The intuition may be an artifact of a prevalent theory of concepts that draws the distinction, so that the intuition may vanish when an alternative theory is adopted. In any case, the issue must be addressed by any theory of concepts proposed in the current philosophical environment.

Another issue is raised by the discussion in the previous section where I noted that there are different kinds of conceptual change. A theory of concepts should provide the cognitive tools needed to distinguish kinds of conceptual change and the relations between them. Different theories may divide up kinds of change in different ways, and even allow or require kinds of conceptual change not allowed or required by other theories. As a result, consideration of different kinds of conceptual change plays a familiar double role: a theory of concepts should provide some insight into the kinds of change that occur, while the ability to account for recognized forms of change will provide a test for such a theory.

Two further complexities will have to be addressed. First, a single item often falls under multiple concepts, so it will be considered "the same as" different items in different contexts. We will have to consider how changes in some of these classifications affect other classifications. Moreover, the various concepts that we use to characterize a single item may undergo different kinds of conceptual change, and we will have to explore how these changes interact.

Second, I have been using a common mode of expression in talking about varying concepts of, say, earth or water. But this may be misleading. Talk about "varying concepts of *x*" is most clearly appropriate when we can pick out the specific item in question, or instances of the type of item in question, by some referential procedure. But we have already encountered important cases in which we cannot do this. Such cases include items that are not easily available to observation, such as isotopes, but also include items that seem much closer to observation, such as natural mother as once understood in California. In addition, a comprehensive theory of concepts must deal with a variety of contested concepts. These include logical concepts such as negation and logical consequence, normative concepts such as ought, and other concepts that have concerned philosopher such as causal relation and truth. Some writers approach this topic by distinguishing between *concepts* and *conceptions*, where the former are more fundamental and the latter are particular elaborations of these. But this distinction rests on the view that there is an important body of concepts that is widely, perhaps universally, shared, but hard to analyze. Yet

many of the concepts we have been examining – such as isotope and derivative – are poor candidates for such universal status. The advent of reproductive technologies suggests that the universality thesis is questionable even in more familiar cases. An alternative approach is that in the cases under discussion we use the same word but associate it with different concepts.

Cases in which we replace one concept by a different concept are not all equally drastic; a concept and its successor may be systematically similar, and the degree of similarity may differ in different cases. Such similarities will help us understand why we may keep a word even as the concepts associated with it vary, and why linguistic change tends to occur at a slower pace than conceptual change. For the moment my concern is to note that a theory of concepts will have to give an account of just what is being contested when we encounter disagreement over the analysis of a concept, and what is involved in the claim that the same concept is being analyzed. In particular, one contested concept – CONCEPT – is central to this study. Philosophers, psychologists, and others disagree on the nature of concepts, and a theory of concepts that provides a general account of disagreements over concepts will have to apply to this case as well.

I want to end this chapter by considering one more issue. Many philosophers hold – sometimes explicitly, sometimes implicitly – that conceptual change does not occur. The examples I have been discussing make such a view dubious. At the very least, these examples provide a much richer variety of cases than are typically found in the literature, and those who deny that the introduction of new concepts and the rejection of older concepts is a central part of human cognitive development have some work to do. Still, Davidson has offered a general argument against the possibility of alternative conceptual systems that many philosophers find plausible. The most focused statement of this argument occurs in his paper "On The Very Idea of a Conceptual Scheme" where he proceeds in two stages. First, he considers the possibility of two conceptual schemes that have nothing in common, so that nothing we can express using one scheme is expressible using the other scheme, and argues that this notion is unintelligible. Then he extends his thesis to cases of partial conceptual variance:

> We must conclude, I think, that the attempt to give a solid meaning to the idea of conceptual relativism, and hence to the idea of a conceptual scheme, fares no better when based on partial failure of translation than when based on total failure.
>
> (1984: 197)

It is this second thesis that is of interest here. I can easily concede Davidson's first claim – to which he devotes most of his essay, as well as discussions in other papers. But Davidson does not actually offer an *argument* against the second possibility. In spite of the strong claim in the passage just quoted, all

Davidson has to say on this topic is that cases of apparent conceptual disparity *may* just be cases in which the same concepts are being expressed in different words. The case studies presented in this chapter make a strong case for the conclusion that more serious changes take place in human cognitive history. In later chapters I will provide an account of how such changes take place without loss of intelligibility.

# 3 Some Theories of Concepts

> It would be odd if the only qualitative dimensions of the world were those which are tied to the sensory centers of the human brain.
>
> (TE 149)

The work of Wilfrid Sellars will be the starting point for my own theory of concepts. In the present chapter I examine some major theories of concepts that predate and overlap Sellars' work – work done with an eye on the history of philosophy. Although theories of concepts are found in all philosophical traditions, I will focus on the empiricist tradition which provides the most explicit and sustained discussions of theories of meaning – which merge into theories of concepts. For these philosophers the nature of meaning is a central philosophic topic, and theories of meaning are a central tool in their approach to a variety of issues. I will offer both expositions and critiques of the philosophers I discuss, with the critiques mainly aimed at raising issues I will have to address in constructing my own account.

## 3.1 Locke[1]

For Locke, as for philosophers generally in the early modern period, the items we are directly aware of are all *ideas*. Following Descartes, the term "idea" is used to emphasize that these are mental entities that exist only insofar as someone is conscious of them. The point is reasonably clear in cases of imagination. If I imagine a green flying horse there is some item before my mind. Presumably I am not in cognitive contact with an actual green flying horse, and the specific item before my mind exists only as long as I am aware of it. A similar point was taken to hold for memory: the item I am aware of when I remember a past event is not the event itself, which no longer exists, but an idea that exists only as long as I am conscious of it. This view extends directly to reasoning: when I reason about circles there is some item I am thinking about – another idea. Perception provides the most difficult (and controversial) case. For Locke, I perceive an object in the physical world as a result of a causal interaction between that object and my

sense organs. This interaction initiates a process in my body that (in some unknown manner) produces an idea; this idea is the object of my direct awareness. I perceive the physical object that initiated the process indirectly, because of its role in causing the idea that I perceive directly. It is beside my purpose here to elaborate or evaluate this approach to perception, although I note that it generates questions of how we know that physical objects exist, and even whether they exist. Responses to these questions provide a major line of philosophical research; some of these responses will be germane to our discussion.

For Locke, ideas provide the content of all cognitive activity. To the extent that Locke has a theory of concepts, concepts are ideas – although, we will see, there are different kinds of ideas, and thus different kinds of concepts. Ideas also provide the basis for a theory of word meaning. Each word in our vocabulary is associated with an idea which *is* the meaning of that word. A sound that is not associated with an idea is just a meaningless sound, not an actual word. (I will focus on spoken language but parallel points hold for written language.) It is important to keep in mind that for philosophers in this period language is not the fundamental medium of thought. Thinking consists of manipulating and comparing ideas. Language is a superstructure that is convenient for communication and as an aid to memory, but is also a source of confusion because different people often associate different ideas with a particular sound, and sometimes use sounds without any associated idea. I want to examine the central claims of Locke's version of the doctrine of ideas, beginning with the different kinds and sources of ideas.

It is a characteristic empiricist thesis that, in a sense to be made more precise shortly, all ideas derive from experience. Two types of experience were generally recognized; Locke calls these *sensation* and *reflection*. Sensation includes all experience associated with our external senses: it provides direct awareness of ideas of colors, sounds, feels, and such, without any discursive activity. Reflection covers cases in which we attend to what is occurring in our minds – what we would call "introspection" and Kant calls "inner sense." Through reflection we become aware of the contents and activities of our own minds, such as our emotional states, and whether we are currently perceiving or imagining.

Now consider the difference between a case in which I am perceiving an item – say, a table – and one in which I am thinking about that table while not actually perceiving it. In both cases I am aware of ideas, although the ideas that occur in cases of perception are particularly vivid. The ideas that occur in when I remember, imagine, or think about a table are less vivid. It will be useful to have some terminology to distinguish the vivid ideas that constitute objects of current experience from the rest. Appropriate terminology is due to Hume, but I will introduce it now and use it henceforth. Hume calls the vivid items that occur in perception *impressions* and reserves the term *ideas* for the dimmer items that occur in the absence of sensory

experience. For Hume, as for his predecessors, both impressions and ideas are mental entities – Hume calls them "perceptions of the mind" (2001: 7). Note especially that the *content* of the idea before my mind when I am seeing a table is identical to the visual content of the idea before my mind when I am remembering that table. The only difference between the impression and the idea is a difference in how vivid they are. A parallel account holds for reflection. If I experience joy I am aware of a specific impression; if I remember joy, or imagine it, I am aware of an idea that is qualitatively the same as the impression, although less vivid. In a similar way, if I am now engaged in an act of imagination I am aware of this through reflection; I am aware of a comparatively dim idea of imagination when I remember having imagined.[2]

Consider another fundamental distinction, that between *simple* and *complex* ideas. A simple idea is "in itself uncompounded, contains in it nothing but *one uniform Appearance*, or Conception of the mind, and is not distinguishable into different *Ideas*" (II.ii.1: 119). Examples of simple ideas are most easily found in sensation: a specific shade of red, or the feel of solidity that I experience when I press on the table in front of me. The clearest examples of complex ideas are those in which several simple ideas are combined into a new idea; my idea of the table is complex since it involves color, size, shape, solidity, and more. Locke treats the idea of the table as a single complex idea, although we are actually aware of several distinct simple ideas that occur together. We can recognize that two ideas are distinct when we find that they can be independently altered. For example, I may see a red square surface, but it is clear that red can occur in conjunction with different shapes, and that square can occur in conjunction with different colors.

Now consider Locke's version of a central empiricist thesis: The mind of a newborn infant is initially a "white paper" without any ideas; all ideas are ultimately derived from experience. Two points about this doctrine must be kept in mind. First, it applies only to simple ideas. In the terminology I have adopted from Hume, every impression is simple, and every simple idea is a less vivid copy of a preceding impression. But once I have a stock of simple ideas there is no limit to the ways in which my imagination can rearrange them into complex ideas – such as a green flying horse that smells like a rose. The same point applies to simple ideas of reflection: I first experience, say, pleasure, and this produces the idea of pleasure about which I can then think, and which I can combine in imagination with other ideas.

Second, ideas are the *objects* of mental operations. Locke takes it for granted that numerous abilities are built into the mind; these include the abilities to perceive, imagine, remember, reason, abstract, and others. My ideas of these activities come from introspection, but the ability to exercise these activities is part of the original equipment of my mind. Thus I first perceive and, as a result, acquire the idea of perceiving. Locke usually classifies each idea as an idea of sensation or reflection on the basis of its source. Thus my

idea of a shade of red is an idea of sensation, even when I recall it with my eyes closed.

Locke's account of the source of our ideas suggests that every simple idea is a copy of an impression, but one feature of Locke's discussion indicates that this is not his view. Locke maintains that some of our simple ideas are derived from more than one sense (II.v) and that some are derived from both sensation and reflection (II.vii). Shape provides one example of the former sort: we have a single idea of a circle, derived from both vision and touch. We also have a simple idea of pleasure derived from both sensation and reflection. Yet if we consider the differences in the sensory qualities of visual and tactile ideas, it is clear that this idea of a circle cannot be a direct copy of either. By way of contrast, Berkeley, Hume, and most later empiricists hold that two different ideas are associated with the word "circle," one visual and one tactile, and they have nothing in common; but this is not Locke's view. At this point it is less than clear what counts as a simple idea for Locke.

Locke's discussion of Molyneux's problem (II.ix.8) raises a further question about the exact nature of Lockean simple ideas. Molyneux asked whether a man who was born blind and learned to distinguish a sphere from a cube (of the same material) by touch would, upon being given sight, be able to tell which is which by sight alone. Locke responds that he would not, and this is surprising given Locke's claim that we have a *single idea* of a geometric shape that can be derived from either vision or touch. Presumably, our subject has already acquired this idea from touch and Locke's account of the workings of the mind (developed mainly in Book IV) suggests that he would be able to compare the current impressions with the idea already stored in memory. Berkeley and Hume can give the same answer as Locke to Molyneux's question without inconsistency – but we will eventually want to consider the possibility that Locke is on to something, although the point cannot be adequately developed within the framework of an empiricist account of concepts. I will leave aside ideas that have multiple sources in the remainder of my discussion of Locke.

The nature of complex ideas is also more complex than might appear at first glance. Locke claims that all complex ideas are made by the mind (e.g., II.xii), but does not stick consistently to this claim – which is fortunate, since the claim is implausible in the context of Locke's own discussions. When I see a table in front of me I become aware of a complex idea without *constructing* it out of the component simple ideas. The converse claim – that all ideas constructed by the mind are complex – is more plausible. Locke distinguishes several kinds of complex ideas; I want to examine some of these.

Consider, first, a special class of ideas, *abstract ideas*, which Locke introduced to deal with a problem that will concern all the classical empiricists.[3] Each impression is fully determinate. We see a specific shade of red, a circle of definite radius, and so forth. Since simple ideas are copies of impressions, each simple idea, whether it occurs in isolation or as part of a complex, is

also fully determinate. As a result, those complex ideas that are combinations of simple ideas are also fully determinate. But we are also able to think general thoughts. We can, for example, think of red without limiting ourselves to a specific shade; we can think of color in general; and we can go to higher levels of abstraction, such as thinking of properties in general. In a similar way, we can think of tables apart from any specific table, and proceed to furniture, manufactured objects, and so on. One especially important case concerns shapes. Mathematicians prove theorems about triangles that are not tied to a triangle of any specific size or shape; they also prove more general theorems about polygons. There must be some difference in my mind when I am thinking about triangles in general than when I am thinking about a specific triangle. For Locke this translates into the requirement that there must be different ideas before my mind in the two cases. Locke deals with this issue by maintaining that the innate endowment of the mind includes the ability to form abstract ideas: to compare a number of ideas that we already have and create a new idea that contains features that all the instances share, while leaving out those features that distinguish one from the other. This new idea, like all ideas, is a mental particular, and serves as the idea before my mind when I think general thoughts. It is not altogether clear whether Locke considered abstract ideas to be complex ideas or a completely different type. At II.xii.1 Locke seems to distinguish between complex and abstract ideas. But in this passage he also seems to distinguish complex ideas from ideas of relations, yet at II.xiii.3 he describes relations as a species of complex ideas, and proceeds to treat them as such. In at least one later passage Locke writes of "abstract complex *Ideas*" (II.xxxii.6: 385). Later empiricists, beginning with Berkeley, reject abstract ideas – which will require that they provide an alternative account of general thoughts.[4]

Many of our most common and important concepts are relational; thus a theory of concepts must include an account of relational concepts. Since Locke considers these to be complex ideas, he must explain their origin. He must also provide an account of the content of specific relational concepts, and this account must be in terms of simple ideas.

Locke begins his discussion of relations by distinguishing those "*Ideas*, whether simple or complex, that the Mind has of things, as they are in themselves," from those that "it gets from their comparison one with another" (II.xxv.1: 319). Although Locke insists that these relational ideas "*all terminate in*, and are concerned about those *simple Ideas*, either of Sensation or Reflection; which I think to be the whole Materials of all our Knowledge" (II.xxv.9: 323), he is quite clear that relational ideas are not to be identified with the ideas of the relata. This point is important from two directions. First, Locke notes that two individuals can arrive at the same relational idea as a result of comparing different relata:

> the *Ideas* of relation may be the same in Men, who have far different *Ideas* of the Things that are related, or that are thus compared, *v.g.*

> Those who have far different *Ideas* of a *Man*, may yet agree in the notion of a *Father*: Which is a notion Superinduced to the Substance, or Man, and refers only to an act of that thing called Man, whereby he contributed to the Generation of one of his own kind, let Man be what it will.
>
> (II.xxv.4: 320)

Second, two items may be compared in terms of many different relations:

> there is *no one thing*, whether simple *Idea*, Substance, Mode, or Relation, or Name of either of them, *which is not capable of almost an infinite number of* Considerations, in reference to other things: and therefore this makes no small part of Men's Thoughts and Words. *v.g.* One single Man may at once be concerned in, and sustain all these following *Relations*, and many more, *viz.* Father, Brother, Son, Grandfather, Grandson, Father-in-Law, Son-in-Law, Husband, Friend, Enemy, Subject, General, Judge, Patron, Client, Professor, European, English-man, Islander, Servant, Master, Possessor, Captain, Superior, Inferior, Bigger, Less, Older, Younger, Contemporary, Like, Unlike, *etc.* to an almost infinite number: He being capable of as many Relations, as there can be occasions of comparing him to other things, in any manner of agreement, disagreement, or respect whatsoever.
>
> (II.xxv.7: 321–22)

Note especially that in many cases the relata will be the same, although the comparisons differ.

Locke also holds that a relational idea may be clearer than the ideas of its relate:

> This farther may be considered concerning *Relation*, That though it be not contained in the real existence of Things, but something extraneous and superinduced: yet the *Ideas* which relative Words stand for, are often clearer, and more distinct, than of those Substances to which they do belong. The Notion we have of a Father, or Brother, is a great deal clearer, and more distinct, than that we have of a Man; or, if you will, Paternity is a thing whereof it is easier to have a clear *Idea*, than of *Humanity*. . . .
>
> (II.xxv.8: 322)

Yet these remarks fall short of providing a full account of how relational ideas arise, although providing an account of the genesis of ideas is one of Locke's central projects. He seems to think that, in at least some cases, there is no great problem in explaining how relational ideas arise. After mentioning a few examples, Locke writes:

> These and the like *Relations, expressed by relative terms, that have others answering them, with a reciprocal intimation*, as Father and Son; Bigger

> and Less; Cause and Effect, *are very obvious* to every one, and every Body, at first sight, perceives the Relation.
>
> (II.xxv.2: 319–20)

Now this claim might be appropriate if Locke were discussing cases in which we recognize an instance of a relation given that we already have the relevant concept, but it provides no insight into how the concept is acquired. Suppose I notice that *A* is larger than and to the left of *B*. Two different relational ideas pertaining to the same relata are now before my mind. That they are distinct from each other and from the ideas *A* and *B* is clear enough if I already have the concepts LARGER THAN and TO THE LEFT OF, but suppose I do not have these concepts. Why should these concepts – and myriad others – leap to my mind? Consider, moreover, Mary and Jane: Mary is older than Jane, richer than Jane, smarter than Jane, and Jane's sister-in-law. These relations are not all obvious at first sight.

In addition, Locke's examples do not provide adequate accounts of the *content* of relational ideas. Locke does nod in this direction. For example, he tells us "That a *Cause* is that which makes any other thing, either simple *Idea*, Substance, or Mode, begin to be; and an *Effect* is that, which had its Beginning from some other thing" (II.xxvi.2: 325). But this is hardly an account of the ideas of cause and effect in terms of simple ideas. It is far from clear that the task can be carried out; we will see how other empiricists fare.

Next consider SUBSTRATUM, a concept that comes in for much criticism by later empiricists (and others). Locke introduces this concept because he thinks it is required by our ideas of specific substances, such as my idea of a particular table or a particular mind. He tells us that we conceive of a substance as "a certain number of simple *Ideas* [that] go constantly together; which being presumed to belong to one thing . . . are called so united in one subject, by one name . . . " (II.xxii.1: 295). Moreover, he adds, "not imagining how these simple *Ideas* can subsist by themselves, we accustom our selves, to suppose some *Substratum*, wherein they do subsist, and from which they do result, which therefore we call *Substance*." Locke provides several other, more or less metaphorical, accounts of this substratum. Material objects require material substratum, which he describes as *supporting* qualities, as that in which qualities *inhere*, as *standing under*, and *upholding* these qualities; similar suggestions hold for the spiritual substratum that binds together the ideas of a single mind. Locke also points out that he has no *idea* of substratum. But this implies that he has no means of thinking about this item. Indeed, given his view that the meaning of any word is an idea, the word "substratum" would seem to be a meaningless noise.

Locke's successors in the empiricist tradition generally reject the concept of substratum – although Berkeley attempts to save spiritual substratum. I want to note another possibility: Perhaps we do have a concept of substratum (which is not the same as believing that the concept has instances), and failure of the doctrine of ideas to make sense of this concept

should be taken as an argument against that doctrine. Reasons for thinking that we have this concept appear if we reflect on Locke's own remarks when he introduces substratum. He provides a good deal of information about why he thinks this concept is required and why he thinks it is instantiated. In other words, Locke tells us what he means by the word "substratum," and I submit that we understand what he is getting at. We may not agree that the concept is required, and we may give reasons for dropping it from our metaphysics, but such reasons must be based on an understanding of the concept at issue (cf. Weitz 1988: 114, 117). In this regard substratum is on a par with phlogiston, radioactive induction, and telegony.

Let me make the point another way. Suppose Locke were to pick out a specific idea that we associate with material objects, such as solidity, and declare that this is what he means by "substratum." If we literally have no concept of substratum – no notion of what he is talking about – then we have no grounds for rejecting this identification. The obvious objection to this move is that SOLIDITY does not do the job that SUBSTRATUM is supposed to do – which indicates that we have some concept of substratum. An adequate theory of concepts must give an account of the content of that concept.

I want to note some additional problems with Locke's account of ideas, not primarily for the sake of criticizing Locke, but with an eye towards highlighting issues that will have to be dealt with by any theory of concepts. Consider, first, some of the central concepts that occur in Locke's epistemology. These include PRIMARY QUALITY, SIMPLE IDEA, IDEA OF A SECONDARY QUALITY, and ABSTRACT IDEA. A theory of concepts must be capable of providing an account of the content of each of these concepts. Let us see how far we can go from a Lockean perspective. Each of these concepts requires a specific idea, and since these are all general concepts, the relevant ideas must be abstract ideas.[5] Presumably, the idea of a simple idea will be formed by examining simple ideas, retaining what they have in common, and leaving out the features in which they differ. In a similar way, the idea of an abstract idea will be formed by examining abstract ideas, etc. But it is doubtful that this process will provide a distinct abstract idea for each of the concepts that Locke deploys. One example will make the point sufficiently clear. It is important for Locke to distinguish PRIMARY QUALITIES from IDEAS OF PRIMARY QUALITIES, and this requires different abstract ideas corresponding to these different concepts. How will these abstract ideas differ? What do we have before the mind when thinking of a primary quality, and how does this differ from whatever we have before the mind when thinking of the idea of a primary quality? If, as Berkeley maintains, we cannot make this distinction, then Locke is unable even to think of his materialism and indirect-realist theory of perception – let alone argue on their behalf. Nor is it clear how we are to distinguish PRIMARY QUALITY from SECONDARY QUALITY. It will not do to say that once we have lists of primary qualities and secondary qualities we can survey these lists and form the

appropriate abstract ideas. Locke's strategy is to first introduce the distinction and then seek means of determining which of our ideas of qualities fall into each class. This requires that the distinction be made before we can arrive at any examples – and thus cannot be achieved by abstracting from a set of established examples. Locke's version of the theory of ideas does not seem to have the resources needed to introduce these concepts.

There is another aspect of these examples that will concern us: classification of ideas into different types requires that we deploy predicates that take ideas as their subjects. In contemporary terminology these are *second-order* concepts. So are EXISTENCE and UNITY, which Locke includes among the simple ideas. Locke does not address the distinction between first-order and second-order concepts, and this is not surprising since the distinction was not generally available in Locke's milieu. We will see this distinction begin to emerge in Hume, although it does not become fully explicit until Kant. For now, the important point is that Locke's theory of concepts also lacks the resources needed to make this distinction.

I want to return to the doctrine of abstract ideas in order to raise another issue. Locke holds that I form the abstract idea of red by surveying several different shades of red and abstracting features they have in common while leaving out features in which they differ. But this does not seem possible if each simple idea "contains in it nothing but *one uniform Appearance* . . . " (II.ii.1: 119). Moreover, our abstract idea of red can itself become part of the basis for an ascending series of abstractions yielding, say, color, secondary quality, and quality. Ideas at each of the higher levels are supposed to have *less content* than the ideas from which it was formed. Thus the abstract idea of red still has a good deal of content, although less than the simple ideas of specific shades of red. Simple ideas begin to seem rather complex.

It seems that Locke does not provide a coherent account of simple ideas, and this is a particularly significant failing. One task of a theory of concepts is to provide an account of the nature of conceptual analysis. The account we find in Locke – and the overwhelming majority of later empiricists – treats conceptual analysis as an analogue of chemical analysis: There are unanalyzable basic concepts and conceptual analysis consists of resolving complex concepts into their basic constituents. For this view of analysis to be sustainable, it is crucial that we have a clear account of what constitutes a basic concept. Since simple ideas are Locke's basic concepts, he fails at this central task. We will see that this problem is not peculiar to Locke; it recurs in other empiricist versions of the theory of ideas. Examination of such failings helps point the direction to a more adequate theory of concepts.

There is a further question about Locke's theory of concepts that will arise in other philosophers. Locke holds that the mind includes a bundle of abilities that are exercised on ideas. All thinking deals with ideas, which occur only as objects of thought (in the wide sense in which "thought" is regularly used in this period). But one function of concepts is to organize the

items we think about into sets that have relevant similarities. In this respect concepts are tools for thinking about various subjects. Thus while concepts occur as objects of thought (e.g., when we engage in conceptual analysis) concepts cannot be *just* objects of thought.

## 3.2 Berkeley[6]

Berkeley also treats ideas as the only objects of consciousness and distinguishes cases in which we are actually perceiving from those in which we call up an idea in imagination or memory; perception is characterized by more vivid ideas whose occurrence we cannot control. I continue to use Hume's impressions/ideas terminology to express this difference. Along with other empiricists, Berkeley takes it as given that the range of ideas we can contemplate is limited by the range of impressions we experience. Still, Berkeley's version of the doctrine of ideas differs significantly from Locke's. I will focus on those features of Berkeley's account that yield differences in his theory of concepts.

Consider simple ideas. Berkeley holds that the simplest ideas we can form are copies of the simplest impression we experience, but these are not Lockean simple ideas. We cannot, for example, see an object that has extension but no color, so we cannot form an idea that has extension without color. Thus Berkeley rejects Locke's view that simple ideas have a single uniform appearance:

> I can imagine a man with two heads or the upper parts of a man joined to the body of a horse. I can consider the hand, the eye, the nose, each by itself abstracted or separated from the rest of the body. But then whatever hand or eye I imagine, it must have some particular shape and colour. Likewise the idea of man that I frame to myself, must be either of a white, or a black, or a tawny, a straight, or a crooked, a tall, or a low, or a middle-sized man.
>
> (PHKI*10*)

In other words, Berkeley holds that all impressions and all ideas are complete particulars. Each of Locke's simple ideas is, from Berkeley's perspective, an abstraction, and Berkeley denies that the mind has the ability to form abstract ideas. Indeed, the rejection of abstract ideas is central to Berkeley's philosophy.

Berkeley begins his attack on abstraction in the passage just cited. Having noted that he cannot form the idea of a man that has no specific size, shape, and color, he adds some further examples, and concludes:

> I own myself able to abstract in one sense, as when I consider some particular parts or qualities separated from others, with which though they are united in some object, yet, it is possible they may really exist without them. But I deny that I can abstract one from another, or

> conceive separately, those qualities which it is impossible should exist so separated; or that I can frame a general notion by abstracting from particulars in the manner aforesaid.
>
> (PHKI*10*)

This denial of abstract ideas raises the question of how we think general thoughts; Berkeley replies by providing a "selective attention" account. Although an idea is complex, we may attend only to some of its features in a particular situation. When we do this, the idea currently in mind can represent all ideas which share these features: "an idea, which considered in it self is particular, becomes general, by being made to represent or stand for all other particular ideas of the same sort" (PHKI*12*).

Berkeley's account is particularly clear in mathematical proofs. When we prove a theorem, such as that the interior angles of every triangle sum to 180°, we think about a specific triangle, but only a few properties of that triangle enter into the proof. Thus the proof holds for all figures that share these properties – in the case at hand, all triangles (PHKI*16*). In the second edition of PHK Berkeley adds another kind of example: "we may consider Peter so far forth as man, or so far forth as animal, without framing the forementioned abstract idea, either of man or of animal, inasmuch as all that is perceived is not considered" (PHKI*16*). The same method provides the basis for the new mathematical physics by allowing us to prove general results about extension, motion, force, and so forth (PHKI*11*; ALC*7*: 293–95).

The core of this account is Berkeley's understanding of what is required for one idea to *represent* other ideas. "Represent" has a special meaning for Berkeley, which can be brought out by considering a distinction between *representing* and *signifying* that runs through his texts. (See Winkler 1989, Sec. 1.4 for a useful discussion.) Berkeley holds that *A* can represent *B* to the extent that *A* and *B* share properties; *C* signifies *D* whenever a firm association has been established between *C* and *D* so that even though *C* and *D* have nothing in common, thinking of *C* leads automatically to thinking of *D*. Berkeley uses the relation between a word and the idea that provides its meaning as his model of the signifying relation. Let us focus on written language, although parallel points apply to spoken language, sign language, and Braille. When I am reading I become aware of a visual idea that has been conventionally associated with some other idea. Understanding the meaning of the word requires having established this association.[7] A key feature of signification is that these associations are arbitrary. There is nothing in the graphemes "green" or "vert" that provides a reason for associating these visual ideas with a specific color, or something else, or nothing at all. When someone who does not read English sees "green" for the first time, nothing in this idea provides a clue as to its meaning – just as there is no reason why someone who reads only French should think of "green" on seeing "vert" (e.g., EVI*64*, *143*; PHK*43*).

Berkeley holds that our dealings with the empirical world are based on the signifying relation. We learn from experience that visual and olfactory ideas of smoke are associated with the visual idea of fire and that, under appropriate circumstances, the visual idea of fire is associated with ideas of warmth and pain. Nothing in any of these ideas indicates its association with another idea. In the same way, we learn from experience that we cannot walk through solid walls: given the set of ideas we associate with being-up-against-a-wall, we cannot, by any act of will, elicit the experiences that would be associated with passing-through-the-wall. All of the connections we normally think of as cause–effect relations in the empirical world are of this sort (PHK*65–6*). Indeed, the signifying relations between the ideas that guide our dealings with the world constitute a language, although it is a language created by the deity, not by us. This language contains the arbitrariness of human languages in that the deity might have created different associations among ideas. In that case we would have learned those associations instead of the ones we have learned. If, for example, we found that we could walk through green walls, we would learn the appropriate associations as easily as readers of English and French learn to associate the grapheme "pain" with different ideas. To put the point in the starkest terms, Berkeley holds that our impressions are caused in us directly by the deity, and that there are no causal relations between impressions – there are only relations of signification.

This account of signification applies to all cases in which we associate ideas that have no common content. An especially important case is the connection between visual and tactile ideas of shape, since Berkeley holds that they do not share any content:

> we can no more argue a visible and tangible square to be of the same species from their being called by the same name than we can that a tangible square and the monosyllable consisting of six letters whereby it is marked are of the same species because they are both called by the same name.
>
> (EVI*140*)

Visible shapes, he holds, are of no importance except when they indicate tangible shapes, "which by nature they are ordained to signify" (EVI*140*). This theme is picked up in PHK; for example, "The ideas of sight and touch make two species, entirely distinct and heterogeneous. The former are marks and prognostics of the latter" (PHK*44*). Thus, contra Locke, there are no ideas derived from more than one sense. The ideas of each sense are distinct from those of the others, although we learn from experience that certain ideas from one sense signify ideas from another sense.

We arrive here at a view that is accepted in much later empiricist philosophy, although it is sometimes considered problematic. In terms of the theory of concepts it means that we do not have, for example, a single concept of a sphere. Rather, we have two distinct concepts, TACTILE SPHERE

and VISUAL SPHERE. Whatever correlations we find between instances of the two concepts, there is no deeper connection. The relation between a visual and a tactile sphere is of the same kind as the relation between the shape of a visual sphere and its color; there is nothing we can learn about one of these ideas by studying the other. When this is combined with an empiricist theory of meaning, it follows that the term "sphere" is quite as ambiguous as "bank" or "wound."

The distinction between signification and representation is related to another key Berkeleyian distinction, that between ideas and *notions*. Notions are the items before our mind when we think about spirits and their activities. Berkeley says little about notions, but it is clear that he uses "notion" as a technical term for a set of items that must be distinguished from ideas. The following passage is a particularly clear statement of this point:

> our souls are not to be known in the same manner as senseless inactive objects, or by way of *idea*. *Spirits* and *ideas* are things so wholly different, that when we say *they exist*, *they are known*, or the like, these words must not be thought to signify any thing common to both natures. . . . We may not I think strictly be said to have an idea of an active being, or of an action, although we may be said to have a notion of them. I have some knowledge or notion of my mind, and its acts about ideas, inasmuch as I know or understand what is meant by those words. What I know, that I have some notion of. I will not say, that the terms *idea* and *notion* may not be used convertibly, if the world will have it so. But yet it conduceth to clearness and propriety, that we distinguish things very different by different names.
>
> (PHK*142*, see also PHK*27*)

Let us examine why Berkeley requires this distinction.

According to Berkeley, each of us is a spirit and each spirit is conscious of itself and of its current activities; thus far Locke would agree. But consider cases in which we remember or imagine spirits and their activities. Suppose, for example, I am thinking about seeing while my eyes are closed, or about some previous or possible future state of my spirit. Locke would say that in these cases we are thinking of ideas of reflection, but Berkeley denies that there are any ideas of reflection. He does so for a reason that might, at first, seem odd: Ideas are passive while spirits and their activities are active – and no passive item can *represent* an active item. To understand what Berkeley is up to, we must examine a doctrine that Berkeley alludes to on several occasions, although he never picks it out for explicit discussion.

In effect, Berkeley holds that ideas have two essential properties: Ideas exist only when perceived – their *esse* is *percipi* – and ideas are passive. Spirits have two contrary essential features: Spirits are perceivers – their existence does not depend on being objects of consciousness – and spirits are active.

These two features of spirits are emphasized in PHK*137–139*, and Berkeley insists that if an idea does not represent a spirit "in those mentioned, it is impossible it should represent it in any other thing" (PHK*138*). In TD*3* Berkeley insists that he has no idea of God or of any spirit, but adds: "I have therefore, though not an inactive idea, yet in my self some sort of an active thinking image of the Deity" (TD*3*: 232). In Berkeley's view, I submit, the essential features of an item must be included in anything that is capable of *representing* that item. Thus only an active entity can represent an active being, and no idea can represent a spirit.[8] Two tasks now present themselves: first, to develop Berkeley's doctrine of essential properties; second, to consider why Berkeley thinks we must be able to *represent* spirits.

Given the paucity of remarks about notions in Berkeley's text, I will return to his discussions of ideas to tackle the first task. Consider some of Berkeley's reasons for rejecting Locke's distinction between ideas and qualities, particularly primary qualities and their ideas. Qualities, for Locke, exist unperceived; yet if I am going to think of them, I must do so by means of ideas. In the case of primary qualities the procedure seems straightforward because primary qualities are supposed to be exactly like their corresponding ideas except that the qualities exist unperceived. But, Berkeley maintains, this will not work exactly because ideas exist only when perceived. He maintains that we cannot remove this feature of ideas even in thought (PHK*5*): "To be convinced of which, the reader need only reflect and try to separate in his own thoughts the being of a sensible thing from its being perceived" (PHK*6*). Somewhat later Berkeley adds that he will rest his entire case on one consideration:

> It is but looking into your own thoughts, and so trying whether you can conceive it possible for a sound, or figure, or motion, or colour, to exist without the mind, or unperceived. This easy trial may make you see, that what you contend for, is a downright contradiction. Insomuch that I am content to put the whole upon this issue; if you can but conceive it possible for one extended moveable substance, or in general, for any one idea or any thing like an idea, to exist otherwise than in a mind perceiving it, I shall readily give up the cause . . . the bare possibility of your opinion's being true, shall pass for an argument that it is so.
>
> (PHK*22*, cf. TD*1*: 200)

Note Berkeley's claim that an unperceived idea is a *contradiction*; I want to pin down how this supposed contradiction arises.

Berkeley, like his contemporaries, does not distinguish between first-order and second-order properties. With this in mind, let us ask what he means when he says that we cannot conceive of an idea that is unperceived. I suggest he can mean only one thing. When I examine the content of any idea, *perceived is included in that content* along with whatever color, size,

shape, odor, taste, or feel is included. Suppose, then, that we want to think of something that is like an idea, but unperceived. How would we go about this task? One option is to begin with an idea and add *unperceived* to its content, but this will create a contradiction. Another option is to remove *perceived* from its content. To see what this might mean consider the idea of a warm, round, red solid. In imagination I can remove warm from this complex at will. Suppose, then, that in order to conceive of a quality I begin with an idea and remove *perceived* from its content. This, Berkeley maintains, is impossible. If I were to remove perceived from the content of an idea I would no longer have an idea before my mind.

But there is a third possibility. Recall that, for Berkeley, I cannot have an idea of a shape that has no color, although I can selectively ignore the color in order to establish results that hold for extension alone. Can I apply this approach to think of qualities? Suppose I bring the idea of a round, red, solid before my mind and then, in order to focus on the primary quality solidity, selectively ignore its color, shape, and its being perceived so that I can establish results about solidity. I think Berkeley would respond that this is not enough. In order to think of the *quality* solidity I must also add *unperceived* into the content of the idea, and this is where the attempt breaks down. For while I am ignoring part of the content of the idea – that it is perceived – this is still present in the idea so that adding unperceived still generates a contradiction.

A parallel analysis applies to *passive*. Berkeley argues that ideas cannot cause other ideas because all ideas are passive (e.g., PHK*25*).[9] But, for Berkeley, the only way that all ideas can be passive is if passivity is included in the content of every idea. Now suppose we try to form an idea of a spirit. Since spirits are active beings, activity must be included in the idea, and this will generate a contradiction for the same reasons that we found in the case of unperceived. Berkeley underlines the point in two successive passages that occur towards the end of PHK. First, he insists that "it ought not to be looked on as a defect in a human understanding, that it does not perceive the idea of *spirit*, if it is manifestly impossible there should be any such *idea*" (PHK*135*). Berkeley then clarifies the nature of the manifest impossibility:

> it is not more reasonable to think our faculties defective, in that they do not furnish us with an idea of spirit or active thinking substance, than it would be if we should blame them for not being able to comprehend a *round square*.
>
> (PHK*136*)

The upshot, then, is that Berkeley assumes two fundamental dichotomies: perceived/unperceived and passive/active. The first member of each dichotomy is included in the content of every idea and is ineliminable. Up to a point, this parallels our inability to have the visual idea of a shape without

some color. Still, that analogy is too weak: we can keep a shape the same while varying its color. That is, we can replace the determinable *color* with one member of its set of mutually incompatible determinates. But we cannot treat the dichotomies perceived/unperceived and passive/active in a parallel fashion: if we replace perceived by unperceived, or passive by active, the result is no longer an idea. Moreover, the analogy breaks down in another way. There are non-visual ideas that do not include color, but passivity and being perceived are included in the content of every idea. If we wish to bring an active item before our minds, we require something other than ideas – thus notions.

However, the need to think about spirits via an active item arises only if we wish to represent them. Thus we must consider why it is important for Berkeley that we be able to represent spirits. Recall that signification relations can occur only between items we have directly experienced. But, Berkeley contends, there are two kinds spirits that we know exist, and thus can think about, but that we can never directly experience: God and the spirits of other people. For Berkeley I know that God and other minds exist because these can be established by arguments, although I will not develop these arguments here since they would take us too far beyond our concern with concepts (see Brown 2000a for details). The important point is that to think about these items at all we need some mental surrogate for them, and this mental surrogate must include their essential properties. Although to my knowledge Berkeley never states the point, his implicit position is that we need cognitive representatives in order to think about items we cannot experience. In the cases he develops in detail, we must be able to represent triangles in order to prove theorems that hold for all triangles exactly because we cannot perceive all triangles. *A fortiori*, we need cognitive representatives to think about other spirits. Thus notions must be included, along with ideas, in our cognitive repertoire.

Given that we have notions, we can establish signification relations between notions and ideas. Thus once I have the notion of another spirit, I can establish a signification relation between my idea of a red face and my notion of an embarrassed person. Moreover, Berkeley argues, once I recognize that all my impressions are caused by God, then I should establish a signification relation between every impression and the notion of God, so that all experience leads the mind to the deity.

I want to summarize some lessons we can learn from Berkeley for the theory of concepts. Note, first, that Berkeley's argument to show that we cannot conceive of qualities, and thus of material objects, would fail if we treated being perceived and being passive as second-order properties – as properties of ideas, not as items included in their content. Then we could maintain the same content but change the second-order properties without generating a contradiction. Second, as I argued in considering Locke's discussion of substratum, we do understand exactly what Berkeley is rejecting when he argues against materialism, and this implies that we have

the concepts needed to think of that view. Any attempt to argue that we lack the concepts needed to formulate a specific theory would seem to be self-defeating as long as it is clear what theory is being attacked. One can go on, as Berkeley does, to argue that materialism is false, and that even if it were true we could never have grounds for believing it. But these arguments also assume that we have the conceptual resources to grasp the theory being attacked.

Third, Berkeley's selective-attention account of abstraction is a genuine step beyond Locke – although for reasons that Berkeley does not highlight. On Berkeley's approach an idea such as that of a particular triangle becomes a general concept because of *the way we use it*. For Locke we create the abstract idea of a triangle; once we have done this the abstract idea exists as an object we can contemplate and which serves as the concept. For Berkeley there is no such object. Rather, I can have exactly the same idea before my mind – say, an isosceles right triangle – whether I am thinking of that triangle, right triangles in general, isosceles triangles in general, triangles in general, plane figures, and more. The triangle represents all of these, and what I can learn from this triangle in a specific case depends on the features to which I direct my attention. There is an item – an idea or a notion – that is a necessary part of any concept, but in at least some cases, part of what individuates a concept is the way that item is used by a cognitively active being.

There is a further, undeveloped, hint of this theme in Berkeley's texts. Berkeley says little about relational concepts, but when he does mention them he maintains that we have *notions*, not ideas, of relations, since relations involve not just ideas, but also some comparison carried out by the mind: "all relations including an act of the mind, we cannot so properly be said to have an idea, but rather a notion of the relations or habitudes between things" (PHK*142*). Notions are objects of thought, and one of their functions is to represent mental acts. So Berkeley's suggestion seems to be that if we want to think about a relation, we must do so via the kind of cognitive object that is appropriate for considering mental acts. This might provide an approach to the problems about relations that we found in Locke, but Berkeley does not develop the suggestion.

## 3.3 Hume[10]

Hume, like Locke, distinguishes simple from complex ideas, and takes the former to be the ultimate basis for all knowledge.

> Simple perceptions or impressions and ideas are such as admit of no distinction nor separation. The complex are the contrary to these, and may be distinguish'd into parts. Tho' a particular colour, taste, and smell are qualities all united together in this apple, 'tis easy to perceive they are not the same, but are at least distinguishable from each other.
>
> (I.1: 7–8)

Taken in conjunction with Hume's dictum that "whatever objects are separable are also distinguishable, and that whatever objects are distinguishable are also different" (I.7: 17), we arrive at an initial account of simple ideas of sensation that is in general accord with Locke's "single uniform appearance." Hume's initial examples include scarlet, orange, sweet, the taste of a pineapple, heat, thirst, hunger, and pain, among others. (I.1–2). In some cases we must treat Hume as being somewhat casual in order to sustain this reading; for example when he includes red – rather than a specific shade of red – among the simple ideas (I.1). Later, in his brief initial discussion of ideas of substances, Hume says that the simple ideas involved in the idea of gold include "a yellow colour, weight, malleableness, fusibility . . . " (I.6: 16). Again we find yellow, rather than a specific shade of yellow, while the last two are dispositions, which would seem to be examples of complex ideas.[11] These examples suggest that Hume's account of simple ideas is not all that clear.

Serious complications arise when we introduce a point on which Hume appears to agree with Berkeley: we cannot have an idea of anything that cannot actually exist.[12] But, Hume tells us, it is "a principle generally receiv'd in philosophy, that every thing in nature is individual, and that 'tis utterly absurd to suppose a triangle really existent, which has no precise proportion of sides and angles" (I.7: 18). It follows that any idea of a triangle we can form has sides and angles of determinate sizes. In general, "*the mind cannot form any notion of quantity or quality without forming a precise notion of degrees of each*" (I.7: 17). Other examples include the impossibility of forming an idea of a line apart from its length, or the idea of warmth apart from a specific degree of warmth. Berkeley would approve of these conclusions, although he would not accept Hume's argument from what occurs in nature to features of our ideas. In effect, Hume adopts one theme from Locke and another from Berkeley. Hume agrees with Berkeley about the simplest ideas that can actually appear before the mind – we may describe these as *psychologically simple*; but he agrees with Locke about which ideas are *epistemically simple*.[13]

Hume's most detailed account of the interplay between these two conceptions of simplicity occurs when he introduces "distinctions of reason." He is concerned here to reconcile the thesis that we cannot form an idea of the color of a body apart from its form with the fact that we do distinguish shape from color, and can think about specific shapes apart from their associated colors and conversely. We arrive at this distinction only as a result of comparisons.

> Thus when a globe of white marble is presented, we receive only the impression of a white colour dispos'd in a certain form, nor are we able to separate and distinguish the colour from the form. But observing afterwards a globe of black marble and a cube of white, and comparing them with our former object, we find two separate resemblances, in what formerly seem'd, and really is, perfectly inseparable.
>
> (I.7: 21–22)

This reads rather like Berkeley's account of selective attention; it would be the end of the matter except for the passages already noted in which Hume describes a specific shade of color, a particular taste, and so forth as simple ideas. Simple ideas so conceived play a central role in Hume's epistemology.

Hume elaborates on this interplay between Lockean simple ideas and the complexity of the simplest ideas we experience in the "Appendix" to the *Treatise*, which contains his later reflections on several themes. Here he maintains that simple ideas may resemble each other in multiple ways.

> *Blue* and *green* are different simple ideas, but are more resembling than *blue* and *scarlet*; tho' their perfect simplicity excludes all possibility of separation or distinction. 'Tis the same case with particular sounds, and tastes and smells. These admit of infinite resemblances upon the general appearance and comparison, without having any common circumstance the same. And of this we may be certain, even from the very abstract terms *simple idea*. They comprehend all simple ideas under them. These resemble each other in their simplicity.
>
> (I.7: 18–19)

Two points about this passage bear special consideration. First, Hume still holds that there are simple ideas within which no distinctions are possible. These ideas are simple in the epistemic sense; I will use "simple idea" only for these in the remainder of my discussion of Hume. My idea of a white sphere is not simple in this sense since it admits of distinctions, albeit distinctions of reason.

Second, one of Hume's aims in this passage is to cut off a counter-argument to his view that there are simple ideas by acknowledging that a simple idea may resemble other simple ideas in various and different ways. Thus *A* may resemble *B* in some respects but not in others, and *A* may resemble *C* in ways that *B* does not resemble *C*. But, Hume contends, it does not follow that these ideas are complex; varying resemblances can occur without complexity. Still, unlimited comparisons of similarity could be taken as a challenge to the notion that some of our ideas are simple. At the very least, it suggests that SIMPLE IDEA is not a simple concept. Indeed, SIMPLE IDEA is a second-order concept, and we are reminded that a theory of concepts should include an account of higher-order concepts. Moreover, if simplicity is not second-order, simplicity is part of the content of every simple concept, making every such concept complex. (Compare the discussion in Russow 1980: 345–47.) I will argue shortly that Hume has an inkling of the notion of a second-order property, although he does not provide an account of this notion.

I turn next to Hume's account of general thoughts. Hume says that he agrees with Berkeley's rejection of abstract ideas and account of how we think general thoughts, but the view Hume claims to share with Berkeley is not an accurate account of Berkeley's position. Hume writes, "all general ideas are nothing but particular ones, annex'd to a certain term, which gives

them a more extensive signification, and makes them recall upon occasion other individuals, which are similar to them" (I.7: 17). Language thus plays a role in Hume's account that it does not play for Berkeley. Berkeley views language as a source of error; in the final section of PHKI he admonishes the reader to attend directly to ideas, not to words. Berkeley does discuss how language becomes general, but he treats general words and general thoughts as distinct, although related, topics. While Berkeley introduces the topic of general words first, he immediately shifts the discussion to ideas because, "By observing how ideas become general, we may the better judge how words are made so" (PHKI*12*: 31). For Hume, language and resemblances among ideas are equal partners in our ability to think general thoughts. In addition, a third element – *habit* – plays a central role. To produce a general thought we first notice a resemblance among several ideas, then we generate a habit of applying the same word to all items that resemble in this way. Once we have established this habit, encountering the word will elicit one of the resembling items in all its particularity; but the mind changes the specific idea if some of its properties become problematic. "The word raises up an individual idea, along with a certain custom; and that custom produces any other individual one, for which we may have occasion" (I.7:19). Hume describes this last point as "one of the most extraordinary circumstances in the present affair . . . " (I.7:19). Suppose I am using a specific idea of a triangle as a basis for reasoning about all triangles, and mistakenly generalize a feature of this triangle. Hume maintains that the mind will automatically show me my mistake:

> Thus shou'd we mention the word, *triangle*, and form the idea of a particular equilateral one to correspond to it, and shou'd we afterwards assert, *that the three angles of a triangle are equal to each other*, the other individuals of a scalenum and isosceles, which we over-look'd at first, immediately crowd in upon us, and make us perceive the falsehood of this proposition, tho' it be true with relation to that idea, which we had form'd.
>
> (I.7: 19)

While many commentators agree that the proposal is extraordinary, it is in accord with Hume's relatively mechanical view of the workings of our minds. It is difficult to see what other account Hume could give.

Now consider Hume's missing shade of blue, which challenges the key thesis that every idea is copied from a prior impresson: "There is however one contradictory phænomenon, which may prove, that 'tis not absolutely impossible for ideas to go before their correspondent impressions" (I.1: 9, cf. Hume 1975: 20–21). Hume imagines an adult who has experienced a wide variety of colors, but has never seen a particular shade of blue.

> Let all the different shades of that colour, except that single one, be plac'd before him, descending gradually from the deepest to the lightest;

> 'tis plain, that he will perceive a blank, where that shade is wanting, and will be sensible, that there is a greater distance in that place betwixt the contiguous colours, than in any other. Now I ask, whether 'tis possible for him, from his own imagination, to supply this deficiency, and raise up to himself the idea of that particular shade, tho' it had never been conveyed to him by his senses? I believe there are few but will be of opinion that he can; and this may serve as a proof, that the simple ideas are not always deriv'd from the correspondent impressions; tho' the instance is so particular and singular, that 'tis scarce worth our observing, and does not merit that for it alone we should alter our general maxim.
>
> (I.1: 10)

Williams (1992: 86) notes that this case is not nearly as singular as Hume suggests. Examples arise wherever we can arrange putatively simple ideas in an ordered sequence. Cases include ideas of specific degrees of heat or cold, sounds of a specific loudness, and shades of every color. In addition, some tastes are more bitter than others, some are sweeter than others, some surfaces are rougher than others, and so forth. Morreall (1982: 408–9) notes that there need not be only one shade of blue between two given shades, and that we are also able to extrapolate shades of color, and other sensibles that occur in an ordered series, beyond the set of examples we have experienced. Examples discussed in Ch. 2 suggest that our ability to construct new concepts has an even greater range than these examples indicate.

Limits on our ability to construct new concepts constitute a perennial theme in the history of philosophy. Philosophers have frequently identified some class of especially important concepts and argued that these must either be innate or have been acquired by copying them directly from experience. The view is at least as old as Plato (cf. Buchdahl 1969: 110–14) and it is worth noting an overlap on this topic between Hume and Plato, however great their differences on other topics. In *Phaedo* Socrates considers the source of our concept of equality: "not the equality of stick to stick and stone to stone, and so on, but something beyond all that and distinct from it – absolute equality" (74a, 1961: 57; all page references are to Plato 1961). We have this concept but, Socrates argues, we could not acquire it from experience because we do not experience absolute equality. Two physical objects may appear equal to one person and unequal to another, but we have never "thought that things which were absolutely equal were unequal, or that equality was inequality" (74c: 57). The less-than-absolutely equal items we encounter suggest this concept to us, and Socrates maintains that such experiences provide an occasion to remember the concept of absolute equality, which we must already have. Moreover, this result holds in all cases in which we judge that something is "like something else, but it falls short and cannot be really like it, only a poor imitation . . . anyone who receives that impression must in fact have previous knowledge of that thing which he

says that the other resembles, but inadequately" (74d–e: 57). This, in turn, shows that:

> before we began to see and hear and use our other senses we must somewhere have acquired the knowledge that there is such a thing as absolute equality. Otherwise we could never have realized, by using it as a standard for comparison, that all equal objects of sense are desirous of being like it, but are only imperfect copies.
>
> (75b: 58)

This conclusion also holds for "absolute beauty, goodness, uprightness, holiness, and, as I maintain, all those characteristics which we designate in our discussions by the term 'absolute'" (74c–d: 58).

The concepts that concern Plato all involve differences of degree, and Plato takes it for granted that we could not construct the limiting cases of these sequences for ourselves. Thus he holds that these concepts are either innate or directly copied from experience.[14] If we eliminate one of these sources, we may conclude that the other must be operative. Hume takes the same dichotomy for granted in most of his discussions of simple ideas, and argues that none of these ideas are innate (see especially III.14: 157–58, but also I.1; 1975: 19–20). It follows that simple ideas are copied directly from experience, and Hume sometimes argues that we do not have an idea that we claim to have by arguing that there is no impression from which it could have been derived. But this means that the missing shade of blue, and the many other examples noted above, really contradict his central claim – as Hume says. I suggest that these cases, as well as those cited by Plato, indicate a third option: an ability to create new concepts. While such an ability has not often been recognized in the history of philosophy, we should recall that all three of our classical empiricists agree that the human mind has a large number of cognitive abilities built into it; we are dependent on experience only for the material on which these abilities act. The scope of our ability to operate on available conceptual contents to produce new contents is one of the central issues I will address in subsequent chapters.

A central feature of the way Hume develops his account of simple ideas requires that he move with care in arguing from the absence of an impression to the conclusion that we lack a particular idea. Hume introduces the claim that every simple idea is copied from an impression as an *empirical claim*. In the *Treatise* he tells us:

> Every one may satisfy himself in this point by running over as many as he pleases. But if any one shou'd deny this universal resemblance, I know no way of convincing him, but by desiring him to shew a simple impression, that has not a correspondent idea, or a simple idea, that has not a correspondent impression.
>
> (I.1: 8)

And even more forcefully in the first *Enquiry*:

> Those who would assert that this position is not universally true nor without exception, have only one, and that an easy method of refuting it; by producing that idea, which, in their opinion, is not derived from this source. It will then be incumbent on us, if we would maintain our doctrine, to produce the impression, or lively perception, which corresponds to it.
>
> (1975: 19–20)

This leaves Hume in a delicate situation: any case in which he asserts that we do not have a simple idea because we lack the requisite impression could be offered as a counter-instance.[15] Thus Hume uses this principle selectively, arguing that we lack the ideas of material and spiritual substance, but do have the idea of a causal connection – although it is an idea of reflection (III.14, 1975: 75–76).

I suggested in the two previous sections of this chapter that we have the concepts of material and spiritual substances – which does not imply that these concepts have instances. But in the empiricist tradition the distinction between having a concept and deciding if that concept has instances vanishes for simple ideas. Given the thesis that every simple idea is copied from an impression – along with the view (shared by Berkeley and Hume) that having an impression of $x$ is just what it means to say that $x$ is instantiated – it is clear that we cannot have an uninstantiated simple idea. It is worth pondering whether this is a virtue or a vice in a theory of concepts; I will return to this topic in this and later chapters. Hume also assumes, without discussion, that if we have an idea of substance it must be a simple idea. In the case of the soul this assumption seems in accord with the traditional view that the soul is a simple entity, but we can still ask whether our *concept* of a simple entity must be a simple concept. Hume's response is found in his discussion of this idea when he asks: "For how can an impression represent a substance, otherwise than by resembling it?" (IV.5: 153). This is a familiar theme: it seems that Hume shares Berkeley's view of *representation*.[16] Recall also Hume's insistence that simple ideas resemble the impressions from which they are derived. I submit that this resemblance is necessary if these ideas are to serve as cognitive proxies for impressions.

Hume also assumes that any idea of a causal connection must be a simple idea copied from an impression. But every impression is an independent entity that has no intrinsic connections to any other impressions. As a result, any impression of a causal relation would be just an additional item that occurs between the impressions of a cause and an effect. There is no sense in which such an impression, or the derived idea, can be said to establish a *relation* between a cause and an effect. In one respect this should come as no surprise: early in the *Treatise* Hume includes relations among the complex ideas (I.4). But it is surprising to find him later treating the idea of a causal relation as a simple idea and seeking the impression from which it is copied.

In fact, Hume has not done any better than Locke or Berkeley in developing an account of relational concepts, in spite of the central role that relations play in Hume's philosophy.

Hume does, however, take an important step towards recognizing the distinction between first-order and higher-order concepts. For Hume, the belief that an entity exists is a property of an idea, not part of its content. Hume offers several arguments on behalf of this claim, but the most important for present purposes consists of pointing out that if belief is included in the content of a concept, then it is impossible for two people to hold contrary beliefs about the existence of a single type of entity since they would be thinking of different ideas (III.7). In a footnote to this discussion Hume questions whether existence is an idea, and thus questions the thesis, generally received at the time, that every proposition – even those of the form "*x* exists" – must involve at least two ideas. In his "Appendix" Hume adds that he was mistaken in holding that "two ideas of the same object can only be different by their different degrees of force and vivacity" (400–1), and asserts that there are other ways in which they can "feel" different, but does not elaborate. The upshot is that Hume recognizes the distinction between the content of an idea and a property of that idea, and thus that there can be significant differences between ideas that have identical content, although he does not provide a developed account of higher-order concepts.

## 3.4 Early Twentieth Century Empiricism

Two major forms of empiricism emerged early in the twentieth century. One of these, logical positivism, developed in Austria and Germany, and was brought to the attention of English-speaking philosophers by Ayer (1936). Recent scholarship has shown that the aims and views of the original positivists were more complex than the image that has long prevailed. While the positivists were, in important respects, radical empiricists, they worked in a neo-Kantian framework that made their approach rather different from that of the British tradition (cf. Friedman 1999; Giere and Richardson 1996; Tsou 2003). However, these differences were submerged when most positivists moved to the English-speaking world after the rise of Naziism. The result was a synthesis of the positivists' original themes with those prevalent in their new environment. Logical empiricist philosophy of science was an important outcome of this synthesis. In Sec. 3.5 I will consider a topic from the literature of logical empiricism that has major significance for our concerns here. In the present section I will examine a variation on classical empiricism that developed in the English-speaking world.

A central feature of the new version of empiricism is its total rejection of the *psychological* basis of earlier empiricist epistemology. The new empiricists drew a sharp distinction between epistemology and psychology, holding that psychology is an empirical science, while epistemology is an autonomous a priori discipline; thus psychological results have no relevance to epistemology.

Epistemology is concerned with the foundations of knowledge – including the foundations of empirical knowledge – and must, as a matter of logic, precede any empirical investigations. It is the task of epistemology to establish norms for scientific research, so epistemological results have implications for the methodology of science and thus, indirectly, for its content. I will explore this normative claim as we proceed, but will focus mainly on one central theme of this new empiricist epistemology: a theory of concepts that is distinct from any psychological theory. One common approach was to proceed by reworking Hume. Price's (1940) essay, "The Permanent Significance of Hume's Philosophy," provides a clear example of this approach; I will begin my account by examining that discussion.

Price begins with Hume's claim that all ideas are copies of impression, but notes that "idea" can mean either a *mental image* or a *concept*. It is only in the latter sense, Price insists, that ideas have any relevance to philosophy. Whether mental images are derived from impressions "is not of the faintest philosophical interest. It is a psychological doctrine, not a philosophical one, and it has nothing whatever to do with Empiricism. Empiricism is a theory about concepts, not about images" (10). Price next drops the term "concept" because it has a "subjectivist flavor"; instead, he will talk of *universals*.

We are now ready for Price's first formulation of the Empiricist Principle: "Every universal which we are aware of has *either* been abstracted from experienced instances *or* is wholly definable in terms of universals so abstracted" (1940: 10). However, this formulation is not satisfactory because it amounts to an inductive generalization stating how we become aware of universals. It is an empirical claim, not a philosophical proposition. In order to eliminate all empirical aspects from the principle, it must be restated in semantical terms – that is, in terms of linguistic symbols and their meanings. This will require some new terminology.

Let us distinguish between primary and secondary symbols: secondary symbols can be defined in terms of other symbols, primary symbols cannot be so defined. Given this distinction, Price restates the "Empiricist Principle" as the claim that primary symbols can be defined only ostensively:

> the meaning of a primary symbol is given, and can only be given, by pointing to a particular which we are acquainted with in sense or introspection and saying, "*That* is an instance of what I mean by the symbol 'so-and-so.'" . . . So, we may say that since secondary symbols are reducible to primary ones, *all* our understanding of general symbols, according to Empiricists, rests ultimately on ostensive definition.
>
> (1940: 11)

Price then reverts briefly to what he calls a more "vulgar" language – the language of impressions and ideas – along with the rather vague notion that all ideas must be "*cashed* by means of impressions" (11). In this terminology, the empiricist thesis claims that any supposed idea that cannot be

cashed in terms of impressions is a pseudo-idea, analogous to a forged check. This yields the "Empiricist programme: to show that all ideas are ultimately cashable by impressions, that is by data which we are acquainted with in sense or introspection, or in any other sort of acquaintance there may be" (1940: 11–12).[17]

Three features of this new empiricism require further elaboration. First, Price points out that two distinct questions may arise with respect to each term in our secondary vocabulary: What is its meaning? and Is it instantiated? Answering the first question requires analysis – providing definitions of terms on the basis of other terms, and ultimately in the primary vocabulary. Without an analysis we may not be clear on the exact meaning of the term in question. Moreover, if we cannot provide an analysis for a term, it may be because we have been using a meaningless term. (Of course, it may just be a failure of analysis.) Meaning analysis is an a priori discipline: it requires nothing more than reflection on the meanings of terms already available to the analyst.

Price's second question arises because of our ability to recombine terms in our primary vocabulary in new ways – a process that is enhanced by the use of previously defined secondary terms. The process of introducing new terms can proceed by reflection alone, and may result in the production of a meaningful term describing an item that does not actually exist. I can, for example, define "framis" as "a green flying horse that smells like a rose." If each term in this description is meaningful, then "framis" is a meaningful term, but this does not guarantee that a framis exists in the actual world. Whether such an entity exists is an empirical matter. Note carefully that this second question arises only for terms in the secondary vocabulary; given that terms in the primary vocabulary can be defined only by ostension, every term in the primary vocabulary is instantiated.

Second, definitions of terms from the secondary vocabulary must be expressed in analytic propositions. This guarantees their a priori status, and thus their proper place within philosophy as conceived by these philosophers. Moreover, since empiricists standardly hold that formal logic (including pure mathematics) and analytic propositions constitute the only domains of a priori knowledge, we are led directly to the view that logic and meaning analysis are the only proper domains for philosophy – a conception of philosophy that dominated the English-speaking world throughout the twentieth century. For the remainder of this chapter I will include propositions of logic and mathematics under the rubric "analytic."

Third, the thesis that all meaningful empirical statements are, in principle, reducible to statements in the primary vocabulary has important consequences for the resolution of disagreements. The primary vocabulary constitutes a language that is available to all normal human beings; it can be established independently of any beliefs that individuals hold. To be sure, different natural languages exist, but terms in two primary vocabularies that have the same ostensive definitions are straightforwardly interchangeable.

The secondary vocabulary does not have this kind of belief-independence since people with strikingly different beliefs will have different terms in their secondary vocabularies. But claims involving these terms can be translated into the primary vocabulary, and statements of empirical evidence can also be formulated in the primary vocabulary. This provides the basis for testing empirical claims, and for comparing claims made from different frameworks. Those familiar with the literature in philosophy of science from the second half of the twentieth century will recognize Price's primary vocabulary as the *observation language* that has been so widely debated.

I turn now to matters which are less clear than the three just noted. Consider the nature of the items we are acquainted with by sensation. These were labeled "sense data," and I will use that convenient terminology. One central question concerns the ontological category in which sense data belong: whether they are mental, physical, or something else. Most philosophers who discuss sense data – both critics and defenders – treat them as mental entities. This view is tempting given the historical roots of sense data in the doctrine of ideas and the traditional thesis that we are directly acquainted only with the contents of our own minds. Nevertheless, some proponents of sense data did not accept this view. Russell, for example, initially maintained that sense data are physical (1957: 137–38), but later moved to the view that they are neither mental nor physical, and that both the mental and physical realms are "logical constructs" out of this neutral material (1921, 1960). (To claim that *A* is a logical construct out of *B* is to claim that all meaningful statements about *A* can be reduced to statements about *B* without loss of meaning.) Price (1964: 137–38) and Ayer (1936: 123) agreed, although Ayer later moved to the view that sense data are not entities at all, but that "sense data" is a terminological innovation that helps provide a clear formulation of philosophical issues concerning perception (1961, 1965). Moore generally avoided this question, although he eventually decided that it is likely that sense data are mental, but that he wasn't really sure (1962: 58).

Yet the issue is vital since our answer will play a central role in determining which terms can be introduced by ostension, and which must be defined. If we are acquainted only with mental entities, then language about physical colors and shapes will have to be introduced by definition. If we are acquainted with physical colors and shapes, then the terms referring to these will be included in the primary vocabulary. If the items we are acquainted with are neither mental nor physical, every term that refers to a mental or physical entity will have to be defined. How do we decide this issue? As Ayer points out (1936: 122), if the issue is decidable at all, it will have to be decided a priori: no empirical test will determine the ontological status of sense data. Yet it is not clear what resources we have for making such an a priori determination within the empiricist framework. Presumably, only analytic propositions are knowable a priori. But (with the possible exception of propositions of formal logic and mathematics, which are not relevant

here), analytic propositions state analyses of expressions from the secondary vocabulary. Such analyses cannot decide the scope of the primary vocabulary. Nor will it help to hold that "sense data" is a term from the primary vocabulary, to be introduced by ostension. According to this account every item we experience is a sense datum; the question we are concerned with arises only after one has accepted this claim.

This problem also points to a deeper issue: "sense datum" does not seem to be a term of either the primary or secondary vocabulary. Since every item we can point to is supposed to be a sense datum, any attempt to introduce the term by ostension will leave "sense datum" without distinctive content. But this very generality also prevents us from introducing "sense datum" by definition. All terms from the primary vocabulary are equally appropriate for the definition, so "sense datum" is again left without any particular content. Sense-datum semantics is offered as a general theory of meaning, but it is not able to account for the language needed to express that theory. This will serve to introduce a major constraint on any general theory of meaning: *reflexive consistency*. The theory must be capable of accounting for the meanings of the terms used to state that theory.[18]

Next, while it was generally agreed that sense data are particulars and that instances of acquaintance with these particulars are brief, there was controversy about whether sense data are private or public entities, whether they continue to exist unperceived, and whether they can have properties that they do not initially appear to have. These issues are somewhat tangential to the role of sense data in a theory of meaning, but we cannot avoid them completely because sense data played a double role as both semantic and epistemic foundations: All meaning analysis ends with sense data, and all evaluations of the truth or falsity of non-analytic propositions reduces to claims about sense data. Moreover, to a large degree epistemic concerns dominated discussions: the epistemic role of sense data in an empiricist-foundational epistemology led to the requirement that our awareness of sense data be indubitable, and that sense datum reports be infallible. This demand for infallibility then served as a criterion for deciding issues about sense data, which in turn had an impact on the theory of meaning. Note especially that the indubitability in question must be considerably stronger than just a psychological inability to doubt that I am aware of a particular datum. The indubitability of the data must underwrite the infallibility of sense datum reports, and the anti-psychologism of the philosophers in question required that there must be some sense in which this certainty is *logical*. Yet this cannot be the familiar sense in which the negation of a logically true proposition is inconsistent; specific sense datum reports are empirical propositions with consistent negations. The logical certainty of sense datum reports must be delivered by a different route.

Let us approach the issue from the reverse direction: the fallibility of ordinary empirical statements is guaranteed their having consequences that

can be checked at later times or by other observers, and that may result in their retrospective refutation. Sense datum reports will be insulated from such refutation if they have no testable consequences. One way to achieve this is to hold that sense data are private, momentary entities – entities that exist only for the individual who is perceiving them, and only for the brief instant in which they are perceived. But it is far from clear that items of acquaintance invoked to provide the *meanings* of terms in our primary vocabulary must have the features invoked to insure indubitability. So far we have encountered no reason why the terms of the primary vocabulary could not be associated with public, enduring entities. Indeed, Wittgenstein (1953) and others eventually argued that private, momentary entities are the wrong kind of item to provide the basis for semantics.

The above remarks illustrate one way in which the assumption that semantics and empirical knowledge have a single foundation played an important role in the development of the sense-datum view. We find a similar unified foundation among the classical empiricists, but their case for this unified approach rested firmly on empirical psychology. Another feature of classical empiricism is echoed in the early twentieth-century view that both meaning and empirical knowledge ultimately rest on the simplest ideas we experience. Sense data were taken to be colored patches, specific sounds, and such, so that terms in our primary vocabulary that deal with the sensory world refer to qualities; terms that refer to material objects must be among the terms of the secondary vocabulary.[19] Part of the reason for this view of sense data seems to come from examination of our own perceptual consciousness. Strictly speaking, it was claimed, I see colors and shapes, hear sounds of a particular loudness and pitch, and so forth. Yet this seems to be an empirical fact about human perception, and twentieth-century empiricists could not take this as a basis for selecting the sense data. Instead, we find the appeal to indubitability as the basis for selecting the sense data. Price, for example, writes:

> When I see a tomato there is much that I can doubt. I can doubt whether it is a tomato that I am seeing, and not a cleverly painted piece of wax. I can doubt whether there is any material thing there at all. Perhaps what I took for a tomato was really a reflection; perhaps I am even a victim of some hallucination. One thing I cannot doubt: that there exists a red patch of a round and somewhat bulgy shape, standing out from the background of other colour-patches, and having a certain visual depth, and that this whole field of colour is directly present to my consciousness.
>
> (1964: 3)

But while the thesis that my awareness of a bulgy red patch has a kind of certainty that does not accrue to material object claims may be important in the construction of a foundational epistemology, it does not follow that the items of which I can be most certain are, *ipso facto*, the appropriate items for conferring meaning on the primary vocabulary. *Certainty* is not a semantic

category; expressions occurring in fallible propositions may be semantically more basic than, or on a par with, terms that occur in infallible propositions (if such exist).

This leads to my next concern: attempts to construct material-object language out of a quality language. The guiding principle of this research was Russell's maxim: "Wherever possible, logical constructions are to be substituted for inferred entities" (1957: 150). If we begin with sense data and introduce material objects by definition, we will achieve a more parsimonious ontology and epistemology than if we introduce the latter as independent, postulated entities. (A parallel point holds for the theoretical entities of physics.) But even if we carried out these constructions, their significance for questions of semantic priority is unclear. Suppose we take one set of terms, *P*, as primary; introduce a second set of terms, *S*, as secondary; and then define items in *S* in terms of items in *P*. This does not preclude the possibility that we may have been equally successful if we worked in the reverse direction. The general point is clear from elementary logic, which provides a paradigm of this kind of construction. We can begin with one or two propositional connectives and introduce the rest by definition, but there is considerable flexibility in which connectives we take as basic. We can also introduce either of the two standard quantifiers as an undefined item, and the other by definition. If there are reasons for preferring one choice over the other, they will derive from considerations other than definability. Russell recognizes this in the case of material objects and their properties:

> In physics as commonly set forth, sense-dataum appear as functions of physical objects: when such-and-such waves impinge upon the eye, we see such-and-such colours, and so on. But the waves are in fact inferred from the colours, not vice versa. Physics cannot be regarded as validly based upon empirical data until the waves have been expressed as functions of the colours and other sense data.
>
> Thus if physics is to be verifiable we are faced with the following problem: Physics exhibits sense-data as functions of physical objects, but verification is only possible if physical objects can be exhibited as functions of sense-data. We have therefore to solve the equations giving sense-data in terms of physical objects, so as to make them instead give physical objects in terms of sense-data.
>
> (1957: 141)

Yet even if Russell's claim about verification is correct, it remains unclear why we should take the items that are epistemically fundamental as semantically fundamental. Perhaps terms referring to material objects are semantically basic, and we are in a position to discover the special role that sense data play in verification only after we establish a meaningful language.

We are left, then, with the question of how we decide which (if any) terms in our language constitute the primary vocabulary – and thus in which direction

we ought to pursue the process of logical construction. Seventeenth- and eighteenth-century empiricists had a clear answer to this question, but it rested on their view of human psychology and their program stands or falls with that psychology. This basis is clear in Hume's challenge to those who would deny that all simple ideas are copies of impressions: Give me an example of a simple idea that is not copied from an impression. Early-twentieth-century empiricists explicitly rejected a psychological basis for their claims of semantic priority, and have not provided an acceptable non-psychological basis. This leaves us with no reason to believe that Price's "Empiricist Principle" is true or that his semantic program is worth pursuing.

I turn now to another criticism of the empiricist program that is central in the literature. Attempts to define four types of terms have been especially important. One type, the so-called "theoretical terms" of science, will be discussed in Sec. 3.5. Normative terms – including, but not limited to, terms from ethics and aesthetics – constitute a second type. The empiricist program is most plausible for descriptive concepts, and runs into serious difficulties for normative concepts. I will not explore this topic in detail here, but I note that one important response was to deny that there are genuine normative concepts – that is, to deny that normative terms have "cognitive significance." There are, however, two ways of looking at this result. We might consider it a triumph of empiricist philosophy, much as Hume's elimination of any concept of substance was considered a triumph by many empiricists – although Hume had doubts about this which he expressed in the "Appendix" to the *Treatise* (398–401). *Or*, we might consider this a clear failure and thus a counter-instance to the empiricist principle. I will consider this second approach in the next chapter.

The third problematic type of term consists of logical constants and other non-descriptive terms, such as prepositions, that play a central role in language. Locke held that these terms, which he called "particles," do not get their meanings from ideas, but from acts of the mind. He devoted a chapter of the *Essay* (III.vii) to them, and attempted a detailed analysis of "but." On Locke's account particles are introduced by ostensive definitions; his account may not impress more recent empiricists, but it is worth noting that Locke recognized the need for a special account of these terms and tried to provide one. Twentieth-century empiricists limited their discussion of these terms to the logical constants, and these were generally treated as deriving their meaning from a set of linguistic rules, rather than by association with sense data or reduction to terms that are so associated. An important consequence of this move is that we actually have two theories of meaning at work – one for empirical terms and one for logical constants. This raises the question whether an alternative theory of meaning might include these terms in a single unified account. We will consider such a theory in Ch. 4.

Finally, there are the material-object terms of everyday discourse. Sense-datum theorists never succeeded in carrying out this reduction, but propo-

nents of the approach took this to be a failure of analysis, maintaining that the reduction is possible in principle, although very difficult in practice. Eventually, some philosophers began to suspect that the difficulties had a deeper root, and that there are principled reasons why the reduction cannot be carried out. I want to examine two arguments for this conclusion. Both turn on a point that sense datum theorists recognized from the beginning: everyday material-object statements are statements about public objects that exist through time. Thus these statements have implications that go beyond what I perceive at a given moment. These include implications about what I would have experienced had I looked from a different direction, or touched an object that I did not (in fact) touch, and so forth. They also include implications about what other people would have perceived if they had examined this object. Since the aim is to *analyze* material-object statements in terms of sense data, the analysis must include accounts of statements about the sense data that would have been perceived under various counter-factual circumstances.

The first challenge comes from Sellars (P 76–84), who argues that introduction of a sense-datum language requires that we already have a fully meaningful material-object language. One common approach to the required counter-factual hypotheticals is to analyze them in terms of "possible sense data." Sellars focuses on this notion and proceeds in two main stages. He begins the *first stage* by asking how we are to understand this notion and noting that there are two different senses of "possible," an epistemic and an ontological sense. These can be illustrated by considering what we mean by a "possible skid." The epistemic sense is used when we are not sure whether a particular car skidded at a particular time and place, although we have evidence that is consistent with a skid having occurred. Thus we say that a skid was possible and cite the evidence. We are using "possible" in the ontological sense when we are sure that no skid occurred, but recognize that one would have occurred under the prevailing circumstances if the driver had acted in certain ways. It is the ontological sense that is relevant for the analysis of material-object statements since the "possible sense data" referred to in the analysis did not actually occur, but would have occurred if the perceiver had carried out certain actions. Now our beliefs about what would have occurred had the driver taken certain actions depend on our believing certain generalizations about the current conditions. In a similar way, claims about possible sense data require generalizations about the situation at hand. Sellars provides the following example: Suppose I am facing a fireplace in which a fire is burning, my eyes are closed, I am not blind, and I can open my eyes at will. Translation of talk about the fire into sense-datum language must include claims about the sense data that I would see if I were to open my eyes: a toothy, orange-yellow sense-datum. However, I believe this claim because I believe certain generalizations about what would occur if I opened my eyes under these circumstances. But these generalizations appear in a peculiar way in the translation project. The point

of the translation is to give the meaning of a material-object statement, and part of this meaning, we are told, is that if I opened my eyes I would see a toothy, orange-yellow sense-datum. But for this to be part of a correct account of the *meaning* of the material-object claim, the presupposed generalization must be *true*. Yet the generalization involves material-object expressions such as "eye" and "fireplace" in the description of my physical state and the relevant circumstances. The translation into the sense-datum language will not be complete until these expressions have also received the appropriate translations.

At this point a critic such as Sellars can maintain that no complete translation of material-object language into sense-datum language is possible, while sense-datum theorists can still maintain that the task is difficult, and that it has not been shown impossible in principle. This takes us to the *second stage* of Sellars' argument. In order to carry out the translation, the required generalizations must be completely expressed in sense-datum language. Sellars acknowledges that I can formulate generalizations that report no more than correlations among my own sense data, but this is not the kind of generalizations we require because such generalizations are essentially autobiographical: they have no significance for anything beyond my own experience. The language we are trying to translate is language about public objects, and this requires generalizations about what *anyone* would experience under these circumstances. Thus I cannot rest content with correlations in my own experience. Rather, I must use these correlations as evidence for generalizations about what others would experience – generalizations that refer to public physical objects. There is no way in which I can avoid such reference as I attempt to eliminate it. Every attempt to eliminate all reference to material-object terms from my analysis of these terms requires generalizations that make use of material-object language. (Recall the passage quoted above in which Price begins a discussion of sense data by telling us what we are indubitably aware of when we see a tomato.) In essence, Sellars' point is that sense-datum terms are constructed so as to have no implications beyond momentary experience, and thus do not have enough content to provide definitions of terms that have such implications. The partial definitions that philosophers have offered are plausible only to the extent that they still incorporate material-object language. Parallel objections will not occur if we begin with public-object language and use it to introduce sense-datum language – which is what advocates of sense data actually do.

The second argument is due to Berlin (1965) and also depends on the role of hypotheticals in sense-datum translations of material-object statements. Berlin focuses on statements that assert the existence of objects not currently present to my senses – e.g., "There is a table in the next room." Claims about absent objects will have to be completely translated into hypothetical propositions, and Berlin argues that this project is fundamentally misconceived because categorical existential statements have a different logical form, and thus a different meaning, than any set of hypothetical statements. Berlin agrees

that dispositional properties can be completely expressed in terms of hypotheticals, and that physical objects have dispositional properties. He also agrees that surface grammar is not always an accurate indicator of depth grammar, so that many statements that appear to be categorical should be analyzed as conditionals. However, assertions of continued existence – whether observed or unobserved – are not among these because on the standard modern analysis hypothetical statements do not have existential import. Thus the statement that there is a table in the next room makes an existence claim that cannot be captured by any set of hypothetical statements. Moreover, my inability to see the table in the next room is the result of a causal condition – a wall is blocking my sensory access to the table. Suppose the wall were to become transparent. I will now be able to see the table and will no longer be limited to saying what I would observe under certain counter-factual conditions. But this change in causal conditions is not a change in the meaning of the statement "There is a table in the next room." The *meaning* of this claim remains the same whether I can see the table or not.

Berlin's point is that this is not just a temporary failure of analysis. Rather, he has displayed an important class of expressions that cannot be given a complete sense-datum analysis. Whatever may the case with respect to verification, there are at least some material-object expressions for which the attempt to reduce their *meaning* to sense-datum language is misconceived in principle.

I turn next to an issue that has appeared in each of the previous sections of this chapter: relations. Relations play a central role in modern logic; the failure of Aristotelian logic to deal with relations was one of the main issues that led nineteenth-century logicians to seek a richer logic. In modern logic relations can be dealt with in a straightforward way: properties can be treated as sets and relations can be treated as sets of ordered sequences. For example, binary relations can be treated as sets of ordered pairs. Thus the relation of being older than is just the set of all pairs $<a,b>$ such that $a$ is older than $b$. The pair $<b,a>$ is a different item and in the present example only one of these two ordered pairs can belong to this set; if $a$ and $b$ are the same age, neither pair belongs to this set. If $a$ is older than $b$, but $b$ is taller than $a$, then $<a,b>$ will belong to the sets constituting older, but $<b,a>$ will belong to the sets constituting taller. Triadic relations are treated as ordered triples, and so on.

However, while this construction is adequate for purposes of formal logic, it does not address the question that concerns me here: the analysis of relational terms into sense-datum language. Two approaches to this problem would seem to be available. To develop the first approach we must recall that expressions in the base language refer to qualities. This approach would treat all relational expressions as part of the secondary vocabulary and attempt to analyze their meanings using only expressions in the primary vocabulary. I know of no systematic attempt to carry out this analysis, and the project of completely capturing the meaning of

relational expressions in non-relational terms does not seem promising. In this context it is not enough to specify sets of ordered sequences. As the examples in the previous paragraph indicate, our understanding of the meanings of the relational terms provides the basis for including items in particular sets. In addition, two relations (or properties, for that matter) may be coextensive even though the terms that describe them have different meanings.

The second approach is to include some relational terms in the basic vocabulary. This is highly plausible since this vocabulary is supposed to refer to easily observable items, and many relations fall into this category. Relations such as that *a* is to the left of *b*, larger than *b*, and heavier than *b* will often be as easily recognizable at a glance as *a*'s color, shape, or smell. But, as noted in Sec. 3.1, not all relations are of this sort. For example, we do not recognize at a glance the relations spouse, sister-in-law, original and copy, or co-effect of a single cause. Thus we will, again, have to distinguish between basic relational concepts and those that are defined, and then undertake the project of showing that the required definitions can be constructed. To my knowledge, this work too has not been done.

I want to end this section by considering one respect in which even advocates of an empiricist theory of concepts must come to terms with the kinds of conceptual change discussed in Ch. 2. Presumably all the concepts considered there would be found among the secondary concepts. An unanticipated new phenomenon, or a decision to reorganize a subject matter, or modify a traditional way of thinking in a domain, would require modifications of our stock of secondary concepts. Ideally, on an empiricist model, we would do this by retreating to the basic concepts and introducing new constructs to replace the old ones. If we could do this, we would avoid some epistemological problems and some difficulties of communication. But the actual problem of finding an appropriate construct and showing that it does the job would still involve the kind of trial-and-error procedure we encountered in Ch. 2, and would sometimes lead to a major restructuring of our classifications of sense data. Yet it seems clear that this restructuring would have to be guided by some grasp of the concepts in question; the question of conceptual content cannot be deferred until after the classifications are in place. In addition, advocates of the empiricist approach acknowledged that actual reductions are very difficult and had few, if any, examples to display. As a result, all the cognitive action takes place on the level of secondary concepts, and the in-principle reducibility of these concepts would do nothing to help us understand the actual development of concepts in our cognitive history.

## 3.5 Theoretical Terms

The introduction of theoretical terms in science provides a particularly interesting case study. Terms such as "infrared radiation," "electron," "isotope,"

and "gene" were introduced to describe items we cannot sense; indeed, there is much in the world that escapes our unaided senses. The large-scale postulation of such items became a pervasive feature of science in the nineteenth century, and was thus not a problem that the classical empiricists had to face. It became a pressing problem for twentieth-century philosophers of science working in the empiricist tradition. Discussion of their attempts to solve this problem will throw further light on the notion of a primary vocabulary that is particularly close to sensory experience, and on some further issues that arise when one attempts to reduce all language to such a vocabulary. In addition, the discussion will bring theoretical concepts into focus as a topic that a theory of concepts must address.

Postulation of items that we cannot sense raises many questions, but I am concerned here with the meanings of terms introduced to refer to these items. I am not primarily concerned with evidence for their existence – although our discussion of twentieth century empiricism indicates that the two issues will be intertwined. Note especially that THEORETICAL is an epistemological concept; it marks our reasons for postulating an item and believing that it exists. It does not imply that the item has some kind of second-class existence. Our inability to detect an item with our senses says nothing about its metaphysical status.

Introduction of language to refer to items we cannot sense does not automatically generate the problem that will concern us. Whether there is a special problem about the meanings of these terms depends on the theory of meaning we adopt. From an empiricist perspective the problem is clear: Theoretical terms must occur in the secondary vocabulary and thus require analysis in terms of the primary vocabulary if they are to stand as meaningful expressions. Moreover, the philosophers who addressed this problem had a second option open to them. Given their view of the normative role of philosophy, they could have declared these expressions meaningless, and thereby put themselves in the position of criticizing this scientific practice. But logical empiricists did not take this option, at least where physics was concerned.[20] Instead, they exercised considerable ingenuity in attempting to provide the required analyses, with Carnap playing the leading role. Moreover, when they saw that a particular analytical approach would not work, they typically concluded that the failure lay with them, not with the scientists. As a result, logical empiricists made several modifications in their own program as their attempts to provide an analysis of theoretical terms developed. I will trace the main steps in these attempts.

Initially the logical empiricist theory of meaning included two central theses:

> Every descriptive term that is not in the primary vocabulary must be completely defined in that vocabulary – where to define a term is to show how to eliminate it. Hempel calls this the "Requirement of univocal eliminability of defined expressions" (1952: 17–18). (LE1)

> All definitions are analytic propositions. (LE2)

Thus the earliest approach was to seek explicit definitions of theoretical terms as logical constructs out of observation terms. The passage from Russell (1957: 141) quoted above describes this requirement in the case of material-object language, and the approach transfers directly to the case of theoretical terms. Hempel describes the parallel project:

> Any term in the vocabulary of empirical science is definable by means of observation terms; i.e., it is possible to carry out a rational reconstruction of the language of science in such a way that all primitive terms are observation terms and all other terms are defined by means of them. This view is characteristic of the earlier forms of positivism and empiricism, and we shall call it the *narrower thesis of empiricism*.
>
> (1952: 23–24)

Braithwaite provides a specific example of this approach:

> Electrons, on this view, are logical constructions out of the observed events and objects by which their presence can be detected; this is equivalent to saying that the word "electron" can be explicitly defined in terms of such observations. Every sentence containing the word "electron" can, on this view, be translated without loss of meaning into a sentence in which there occur only words which denote entities (events, objects, properties) which are directly observable.
>
> (1953: 52–53)

However, this approach was subjected to three major criticisms from within the logical-empiricist camp.

The *first* was developed most fully by Braithwaite (1953). The approach we are examining requires that all types of evidence for the presence of an item be included in the definition. But this means that we cannot discover a new means of detecting the same unobservable entity. Rather, if we wish to associate new empirical evidence with a theoretical term we must redefine that term. In effect, we reject the originally postulated non-observable and postulate a different non-observable. But, Braithwaite argued, this is not how theoretical terms are used in science. Rather, it is considered a major triumph if entities postulated to explain some observable phenomena can provide explanations of other phenomena – especially phenomena that were unknown when the theory was constructed. Yet "if the theoretical terms of a theory are logically constructed out of observable entities, the theory will be incapable of being modified to explain new sorts of facts . . . " (1953: 53). Indeed:

> the hypotheses of the theory will be logically deducible from the empirical generalizations which they were put forward to explain. Since the

> empirical generalizations are, of course, logically deducible from the hypotheses, such a definition of theoretical terms would make the set of hypotheses logically equivalent to the set of empirical generalizations.
>
> (1953: 67)

Thus "the theory becomes merely an alternative way of stating these facts" (1953: 68). The result is a direct clash between LE1 and scientific practice:

> A definition of the theoretical terms would thus sacrifice one of our principal objects in constructing a scientific theory, that of being able to extend it in the future, if way opens, to explain facts about new things by incorporating the theory in a more general theory having a wider field of application.
>
> (1953: 68)

This line of criticism was generally accepted by logical empiricists; I want to underline two features of the critique. First, we have a clear illustration of philosophers who – in practice – are not treating the theory of meaning as either a priori or normative. At least in the case of theoretical terms of physics, the project these philosophers are pursuing is to develop a theory of meaning that is in accord with central features of scientific practice. If a proposed theory of meaning clashes with established scientific practice, this may be taken as empirical evidence against the theory and lead to its modification.

Second, there is an additional assumption implicit in Braithwaite's assessment of LE1: that as science develops and new phenomena are associated with postulated entities, the *meanings* of the theoretical terms that denote these entities remain stable. We will encounter this view at several points in the course of this book. It is captured, for example, in the view that while an expression may occur in both analytic and synthetic propositions, only the analytic propositions express the term's *meaning*; re-evaluations of synthetic propositions are changes of belief, not changes of meaning. A philosopher who adopts this view might argue that once the meaning of a theoretical term is established, new phenomena can be explained by the theory without being incorporated into the meaning of the theoretical term. But this view was not adopted in the discussions we are now considering.

There is another possible response to Braithwaite's argument: accept pervasive meaning change. We would then treat cases in which new phenomena are incorporated into an existing theory as cases in which one body of theoretical language is replaced by a new body of theoretical language that is found to be more empirically adequate. Such replacement would not be an arbitrary change of meaning; in typical cases it would involve a relatively small change in the sense that the new meaning of the new theoretical term will be quite similar to that of the term it replaces. Hempel notes this possibility:

> the procedure of expanding a theory at the cost of changing the definitions of some theoretical terms is not logically faulty; nor can it even be said to be difficult or inconvenient for the scientist, for the problem at hand is rather one for the methodologist or the logician, who seeks to give a clear "explication" or "logical reconstruction" of the changes involved in expanding a given theory. In the type of case discussed by Braithwaite, for example, this can be done in alternative ways – either in terms of additions to the original partial interpretation [more on "partial interpretations" shortly], or in terms of a total change of definition for some theoretical expressions. And if it is held that this latter method constitutes, not an expansion of the original theory, but a transition to a new one, this would raise more a terminological question than a methodological objection.
>
> (1965: 205)

Hempel's suggestion notwithstanding, the view that the meanings of terms are fixed and usually do not change when we revise our beliefs about a given item is widely held by philosophers. It is often invoked in criticisms of Feyerabend, Kuhn, Sellars and others who believe that meaning change plays a pervasive role in science.

The *second* major criticism of the demand for explicit definitions comes from Hempel; it concerns theoretical terms that take quantitative values. The range of permissible values for such terms is usually the set of real numbers. But "in view of the limits of discrimination in direct observation, there will be only a finite, though large, number of observable characteristics . . . " (1952: 30). Given this finite basic vocabulary, we will be able to achieve only a denumerably infinite set of defined terms. The problem, then, is that there are not enough terms in the observation language to provide explicit definitions for every possible value of typical quantitative terms.

Again we face a situation in which we can either rule out these common scientific terms because they clash with our theory of meaning, or acknowledge the importance of these quantitative terms in science and seek a modified theory of meaning. Hempel opts for the latter course: "rather than exclude those fruitful concepts on the ground that they are not experientially definable, we will have to inquire what non-definitional methods might be suited for their introduction and experiential interpretation" (1952: 31). Before considering a non-definitional alternative I want to discuss a *third* highly influential objection to the original empiricist program.

Disposition terms such as "soluble" provide an important part of the scientific vocabulary, and dispositions are not observables. It seems straightforward that they should be defined in terms of conditionals – e.g., an item is soluble if it dissolves when put in water. This definition can be formalized using "$Sx$" for "$x$ is soluble," "$Wxt$" for "$x$ is put in water at time $t$," and "$Dxt$" for "$x$ dissolves at time $t$":

$$(x)[Sx \equiv (t)(Wxt \supset Dxt)].^{21} \qquad \text{(D)}$$

However, there is a problem with this proposal since the conditional being used is a material conditional. If $x$ is not put in water $Wxt$ is false and $Wxt \supset Dxt$ true. Thus, on this analysis, any item that is never put in water is soluble. Clearly this is not acceptable. Note the role that the material conditional plays in generating this problem. The use of this conditional was challenged in the extensive literature on contrary-to-fact conditionals. But at this stage in the discussion Carnap – who makes the next important move – was so strongly committed to the material conditional that he respond to this problem by rejecting LE1.[22] This led to a new phase of the discussion of theoretical terms that Hempel describes as "the *liberalized thesis of empiricism*" (1952: 31).

In a paper originally published in 1936–37, Carnap fiddled with the formalism (see 1996: 214–25) and wrote down:

$$(x)(t)[Wxt \supset (Sx \equiv Dxt)]. \tag{R1}$$

This is a *bilateral reduction sentence*. R1 says that if an item is put in water, then if it dissolves it is soluble, and if it does not dissolve it is not soluble. This specifies an empirical test for solubility, and provides an empirical basis for introducing "soluble" into the language. R1 avoids the problem we encountered with D, but does so at a price: It is only a *partial definition* of the disposition term since it says nothing about cases in which an item is not put into water. We can, however, further specify this term by considering other tests. For example, if we were to discover that a specific X-ray pattern is associated with all and only those items that dissolve in water, we could introduce a second bilateral reduction sentence:

$$(x)(t)[Xxt \supset (Sx \equiv Pxt)]. \tag{R2}$$

Thus our partial definition receives further elaboration as we learn more about the property in question.

There is variation on this situation that we should also consider. We might have reasons for introducing a disposition term $A$ whenever test $B$ yields outcome $C$, and consider this outcome sufficient, but not necessary, for $A$. And we might discover a different test that is necessary but not sufficient for $A$. This would lead to a pair of *reduction sentences* instead of a single bilateral reduction sentence:

$$(x)(t)[Bxt \supset (Cxt \supset Ax)], \tag{R3}$$

$$(x)(t)[Ext \supset (Fxt \supset \sim Ax)]. \tag{R4}$$

Here too we can discover further tests that will allow us to introduce $A$ in other circumstances.

Reduction sentences of both types provide empirical grounds for introducing a theoretical term, but they do not provide means for eliminating

that term from the language of science. In this respect they are not *definitions* as traditionally conceived. Still, Carnap, Hempel, and others considered this to be a major forward step since it allows for a sense in which theoretical terms have "open texture": their meanings can be extended to deal with new empirical situations. Thus reduction sentences solve Braithwaite's problem about explicit definitions. However, a different problem soon emerged.

Introduction of reduction as a means of specifying meaning requires that we reject LE1, but this new device generates an unintended conflict with LE2. As Carnap pointed out (1996: 217–18), from a reduction pair such as R3 and R4 we can deduce:

$$(x)(t)\sim(Bxt\&Cxt\&Ext\&Fxt),$$

which says that nothing has all four properties $B$, $C$, $E$, and $F$ at the same time. Since the capital letters stand for distinct observable properties, this expression is not analytic. Given that analytic propositions entail only analytic propositions, the reduction pair does not consist only of analytic propositions. This is surprising because $A$, the disposition term we are introducing, supposedly has no antecedent meaning. Thus reduction sentences seem to express conventions, which are prototypical analytic propositions. Carnap concluded that the reduction pair R3 and R4 constitute both conventions and factual statements (1996: 218). Yet it seems that either R3 or R4 alone is just a convention, and there is no reason for considering one of them to be conventional and the other not conventional.

Carnap thought that this problem did not occur for bilateral reduction sentences, so that these are pure conventions, but Hempel soon showed that once we introduce multiple bilateral reduction sentences for a property, we can also derive clearly synthetic propositions from them (1965: 114–15). In his paper "Meaning Postulates" (1952) Carnap proposed another formal maneuver which sought to separate the factual from the conventional aspects of reductions sentences, and restore the role of analyticity. But this proposal was lost in the midst of another development.

When Carnap introduced reduction sentences he believed that all theoretical terms could be treated as disposition terms, but later rejected this view and adopted an approach that had been developing in the literature for some time (1956b).[23] A scientific theory is now divided into two parts. First there is a set of *uninterpreted axioms* in which all the theoretical terms occur; second, there is a set of *correspondence rules* that connect this theoretical system to the observation vocabulary. It is not required that each theoretical term enter by itself into a correspondence rule. Specific theoretical terms may occur only in expressions constructed out of more than one theoretical term; this may even be the only mode in which *any* of the theoretical terms enter into correspondence rules. It is not even required that every theoretical term enters into a correspondence rule. But once the correspondence rules have been estab-

lished, all the theoretical terms acquire empirical significance as a group. (See Feigl 1970 for an especially clear account.) This approach involves another major departure from earlier empiricist views since it no longer requires that meaning be conferred on each individual theoretical term. Although a theory-independent observation language still plays a key role in determining the meanings of theoretical terms, the connection between the theoretical terms and the observation language may be highly indirect.

This new structure generated two different views on the meaning of theoretical terms. One view held that theoretical terms are strictly meaningless in the absence of the correspondence rules; the alternative view held that part of the meaning of theoretical terms derives from the *implicit definition* of these terms by the axiom system – although the terms are not fully meaningful until they have been tied to the observation language by correspondence rules. The proposal also generated considerable debate about the nature of the correspondence rules. If LE2 is to be maintained, correspondence rules must be analytic, but Hempel concluded that these rules are not analytic; he even expressed doubts about the significance of the analytic-synthetic distinction (1963: 703–4; 1965: 133; 1970: 161). Carnap, however, continued to seek a means of separating analytic statements that express meaning from empirical statements, and thus preserve the central role of analyticity in specifying the meanings of terms in the auxiliary vocabulary (e.g., 1963b: 961–66).

The developments we have examined involve a progressive weakening of the original empiricist requirements for meaningful theoretical terms. Commenting on the last of these proposals, Carnap wrote:

> The criterion proposed here is admittedly very weak. But this is a result of the development of empiricism in these last decades. The original formulations of the criterion were found to be too strong and too narrow. Therefore, step by step, more liberal formulations were introduced.
>
> (1956b: 51–52, see also Hempel 1963: 707)

Carnap emphasized that one criterion for the success of the empiricist program is its ability to capture "the way scientists actually use their concepts" (1956b: 40, cf. 66, 68). This is consistent with the position of naturalistic epistemology which holds that a theory of science – including a theory of meaning for theoretical terms – must aim at making sense of actual science, not at establishing a priori criteria for scientific practice.

In general, logical empiricists took the presence of theoretical terms in science as a given, and as a challenge to be met by their theory of meaning. When they failed (by their own lights) to provide acceptable accounts of these terms, they modified their theory of meaning. This procedure is consistent with the following "practical attitude" that, according to Carnap, was found among members of the Vienna Circle:

> We regarded terms of the traditional philosophical language with suspicion or at least with caution and accepted them only when they passed a careful examination; in contrast, we regarded terms of mathematics and physics as innocent and permitted their use in our discussions unless cogent reasons had shown them to be untenable.
>
> (1963a: 65–66)

In this regard, physics was guiding philosophy, rather than philosophy guiding physics. But once we have rejected an a priori status for our theory of meaning, we open up the possibility of even more radical departures from the empiricist program. One alterative appears if we consider the picture of an axiom system and a set of correspondence rules from a different direction. We could then argue that the terms of a theory get their meaning *completely* from relations among the terms in a formal system. Correspondence rules would not be involved in conferring meaning on theoretical terms, but rather in providing interpretations of observables in the language of the theory. Feyerabend and Kuhn made this move around 1962, but C. I. Lewis developed a view of this sort at a considerably earlier date. I will consider Lewis' theory of meaning next.

## 3.6 C. I. Lewis

Lewis developed an epistemology that is in the empiricist tradition, although it also draws heavily on Kant.[24] This epistemology is developed primarily in two books: *Mind and the World Order* (1956, originally published in 1929, henceforth MWO) and *An Analysis of Knowledge and Valuation* (1946, henceforth AKV). The issues that concern me here are mainly treated in MWO and I will focus on that book, but I will include material from AKV when appropriate. One main difference between the two books is that in MWO Lewis treats concepts as mental entities that are independent of any linguistic expressions (MWO 87); in AKV he focuses on language and the theory of meaning. Lewis describes the theory of MWO as "conceptual pragmatism," the word "concept" rarely occurs AKV. In spite of these differences, the epistemologies developed in the two books are substantially the same.

Lewis holds that empirical knowledge involves two independent elements, a *sensory presentation* and a *conceptual interpretation* of that presentation. As in Kant, the sensory and conceptual elements are fully integrated in experience; we do not recognize the presence of these two elements by introspection, but rather by philosophical analysis (MWO 25–26, 53–55, 276). Consider the object in front of me, which I spontaneously identify as a table. The presence of an interpretive element in this experience is indicated by my ability to vary the interpretation: I can identify the item simply as a table, or as an antique, or as an investment, or in many other ways. But all these interpretations are anchored to a sensuous presentation – a *given element* that remains unaltered as my descriptions vary, and that limits the

range of permissible interpretations. As long as I am dealing with actual experience, I cannot identify this item as, say, a fine cognac or the Eiffel Tower. Thus it is not possible to invoke any arbitrary concept as an interpretation of any presentation. This ability to limit interpretations indicates that presentations have features that we distinguish and recognize as similar to features of other presentations we remember (MWO 58–60, 121–22, 131). I will return to Lewis' account of sensory presentations shortly, but first I want to examine his theory of concepts. It should already be clear that Lewis' view of the relation between concepts and sensory experience is fundamentally different from the accounts we have considered thus far.

Lewis holds that we use concepts to interpret presentations, and that the content of those concepts is established independently of any presentations. Concepts occur in systems of interrelated concepts, and the content of each concept is wholly determined by implicational relations to other concepts in that system.[25] All such relations are expressed in analytic propositions, and the task of conceptual analysis is to map out these relations (MWO 80–83, 103–9). This is a major departure from the more common view of analysis as "the process of breaking up a concept, linguistic complex, or fact into simple or ultimate constituents" (Foley 1995). Lewis holds that no concepts are intrinsically simple or fundamental. Every concept has a large set of implicational relations to other concepts, and all concepts are subject to analysis. Because of these relations, conceptual interpretation provides the basis for objective knowledge. Sensory presentations are subjective, and we make the transition from subjective awareness to objective, intersubjective knowledge when we identify a presentation as, say, a table, or a tree, or a planet. Given the relations among concepts, every proposition of the form, "This is a *C*" has consequences that can be checked by further experience. To identify an item as a table implies, for example, that it has a backside and an underside, that it can be seen from many different perspectives, and that it can be touched. Moreover, the main function of conceptualization is to guide our actions in the world, and it is the consequences of our conceptual identifications that provide this guide. While each presentation is private, so that two different individuals may experience qualitatively different presentations in a given situation, their conceptual identifications of these presentations may be the same. It is these identifications that we communicate and act on.[26] From the point of view of action it does not matter if your subjective experience on seeing a stoplight is the same as mine as long as we identify the same items as stoplights and act towards them in accordance with the same rules. It is the subsumption of presentations under concepts that generates an objective public world.

Lewis holds that awareness of a private presentation is infallible, but objectivity brings along fallibility. When we check consequences of a conceptual identification, we may find one that fails. In this case we would normally conclude that we had misidentified the presentation. For example, suppose that at dusk I identify a distant object as a dog but when it neither

moves nor barks I reconsider and identify it as a bush. This new identification has further consequences, which may be confirmed or refuted by future experience. Note especially that many items we encounter are identified as material objects – that is, the concept we use implies that the item is material. It is part of our concept of a material object that such items persist through time and interact with other material objects. As a result, descriptions that use material-object concepts have an unlimited set of consequences for what we may experience in the future. The failure of any of these consequences can lead us to reject the original identification (MWO 279–81). No empirical claim is indubitable; fallibility is the price we pay for objectivity. Moreover, this thesis extends to identifications of qualities. To understand Lewis' view on this point we must return to sensory presentations.

In MWO Lewis is ambivalent about whether we can describe a presentation just as it appears, without making any commitments that are testable by future experience. Sometimes he asserts that the given element in experience is ineffable (e.g., MWO 52–53, 124), but there are also passages in which he says that we can describe presentations by expressions such as "it looks brown," which may capture its appearance without making commitments that open up the possibility of future refutation (MWO 124). In his "Autobiography" Lewis lists his discussion of the given as one of only two parts of MWO that caused him real regret (1968a: 17–18); in AKV he reworks this discussion in terms of a distinction between the *objective* and the *expressive* use of language (AKV 179). We are engaged in an expressive use of language when we use locutions such as "looks like" or "seems like" to describe what is given without opening this description to possible disconfirmation. Lewis maintains that we do not normally use language in this way, but can move to the expressive mode when we wish to cancel the usual implications. Following common practice among empiricists, Lewis usually treats sensory presentations as qualities that we describe "by the use of adjectives of color, shape, size, and so on" (AKV 188). Still, the ordinary use of adjectives is to describe properties that exist through time, so normal descriptions of properties are susceptible to revision. This suggests that there is no reason for restricting expressive language to qualities. I can also say that the item before me "looks like a table," cancelling the implications that could yield a refutation. Lewis does not press this point, but it appears to be his intent since he introduces the expressive use of language with a case in which he seems to see a flight of granite stairs (AKV 179). Indeed, property descriptions are not intrinsically more certain than material object descriptions: under some observational conditions my identification of an object as a table may be more reliable than my identification of its color, size, or shape. In general, expressive language allows us to make irrefutable statements because it uses special phrases to cancel the consequences of normal descriptions that go beyond momentary experience.

The function of expressive language, for Lewis, is solely epistemic, not semantic. Expressive language allows us to describe an indubitable element

in experience, but it plays no role in determining the meanings of terms or the content of concepts. Lewis is quite firm in holding that there is nothing like a sense-datum language whose terms are introduced by ostension. Indeed, no terms get their meaning by ostension. Use of phrases such as "looks like" to describe presentations underlines the point that we must have already mastered public-object language before we can attempt to describe subjective presentations.

For Lewis there need not be a common sensory element in all cases in which I apply a particular concept, nor need I apply the same concept to all cases in which a common sensory element occurs. The description I apply to a presentation can vary with the context. Thus in some contexts I classify an item as round because it looks round, while in other contexts I classify it as round because it looks elliptical. In a similar way, a green appearance may lead me to characterize an object as green in sunlight or as blue in artificial light (MWO 131). Still, whenever I apply a concept I make a transition from subjective experience to ascription of an objective property (MWO 140).

I want to highlight two key differences between Lewis and Kant. First, there are no synthetic a priori propositions in Lewis' philosophy. All a priori knowledge is analytic. For Lewis, synthetic propositions come in two varieties: those which subsume a presentation under a concept, and empirical generalizations; all of these are testable on the basis of future experience. Second, there is nothing in Lewis like Kant's distinction between a priori and empirical concepts. Indeed, the notion of an a priori *concept* is a bit confusing. The primary use of "a priori" is to characterize propositions whose truth-value can be determined by reflection alone, but concepts do not have truth-values. Kant's a priori concepts are concepts that are inherent in the mind, independently of any experience we may have. For Lewis there are no concepts of this sort. All concepts are adopted by human decisions and the set of concepts in our repertoire can be changed by such decisions; I will return to this theme in a moment. There is also a sense in which Lewis' epistemology does not include empirical concepts, as this notion is commonly understood, since none of our concepts are arrived at by abstraction from experience. For Lewis there cannot be any experience until we have a conceptual system in place, and all concepts are constituted in the same way. The thesis that experience requires concepts is the key point of contact between Lewis and Kant, but Lewis' claims that only analytic propositions can be known a priori, and that all synthetic propositions are empirically testable, place him in the empiricist camp. His view that all concepts are constituted in the same way makes him something of a renegade empiricist. Many features of Lewis' account of empirical knowledge are worth exploring, but I will focus only on those that are relevant to his theory of concepts.

It is central to Lewis' philosophy that the concepts each of us wields is a product of human social history and that current concepts can be replaced (MWO 6, 21–22, 110–11; 1970: 250–51). Concepts are tools that guide our

actions in the world. We adopt concepts through a process of trial and error as we attempt to find our way around the world; we abandon concepts when we find that they do not serve our ends. Sometimes a concept is simply dropped from our active repertoire, as occurred with phlogiston and telegony, and may eventually occur with ghosts and witches. Sometimes when we drop a concept we replace it with a different concept. And sometimes, when we follow the latter route, we continue to use the word we associated with the older concept, but now associated with the new concept. Two examples will bring out several key issues.

The concept we currently associate with the word "whale" includes the requirement that whales are mammals, although at an earlier stage in our history this term was associated with a concept which implied that whales are fish.[27] To keep the exposition relatively simple, I will assume that throughout this history our concept of a fish includes, among other features, having gills and lacking lungs, reproducing by means of eggs, and not suckling offspring. The concept of a mammal implies having lungs rather than gills, giving birth to live offspring, and suckling them. For Lewis, in the earlier period "All whales are fish" was a conceptual truth, not an empirical generalization. Now suppose I am living in the earlier period and notice a large sea creature that I identify as a whale, but further examination shows that it lacks gills and is nursing a smaller creature of the same general appearance. Since these features are logically incompatible with fish, the proposition "This creature is a whale" has been empirically refuted. I might also find that I do not have an alternative concept for categorizing this animal, but let this pass for the moment. Suppose, now, that whenever I identify a creature as a whale, further exploration leads me to retract this identification. Once I go beyond the most cursory examination I fail to find any instances of whales – although I do find many animals that have the superficial features of whales along with features characteristic of mammals. I may begin to suspect that my concept of a whale is not instantiated and that it would be useful to adopt a new concept for categorizing these creatures.

It must be emphasized that, for Lewis, I have not discovered that the original concept is wrong or that the proposition "All whales are fish" is false. Since "All whales are fish" is analytic, no experience is relevant to assessing its truth. I may decide to drop this proposition from my active repertoire – which amounts to dropping the concept currently associated with the word "whale" from my conceptual repertoire. But "All whales are fish" has not been refuted. Van Fraassen captures this point in a discussion of Kant's thesis that there is an apodictic basis to our science of matter:

> what we *refer* to as matter may not be an instance of our concept of matter. The pure part of the theory of matter cannot become wrong: in principle it can be propounded in the form of a definition. But although it is apodictic, it can certainly become irrelevant.
>
> (1975: 242)

Lewis holds that we drop a concept when we make a *pragmatic decision* that this concept does not help achieve our goal of making correct predictions in an economical way. Having dropped a concept, we may decide to add a new concept to our repertoire. We may also decide to associate the old word with the new concept – as in our example. After this change "All whales are mammals" expresses a conceptual truth. It will help avoid confusion if we have different symbols for the older and newer concepts. I will use $w_F$ for the concept that includes being a fish among its necessary conditions, and $w_M$ for the concept that includes being a mammal. So my situation is now one in which I have decided to stop describing certain large aquatic creatures as $w_F$ and to begin describing those creatures as $w_M$. "All $w_F$ are fish" and "All $w_M$ are mammals" are both analytic truths; there is no inconsistency because of the difference in the subject terms. But I no longer use $w_F$ in identifying items, and I do not teach "All $w_F$ are fish" to my children – although they might learn about it in a history class. For Lewis, conceptual change is always a matter of either adding a concept to our repertoire or dropping a concept. Lewis holds that each concept has a complete set of necessary and sufficient conditions, so there is no meaningful sense in which we can describe a concept as being changed.

Lewis holds that conceptual change is common in human history – especially in the history of science (MWO 228, 233–35; 1968b: 661–63). In addition to cases in which we find that old modes of classification are inaccurate or superficial, we also encounter situations in which we find it desirable to make classifications never made before, and to think of items not previously considered. Another example will bring out some further features of Lewis' views on conceptual change.

Lewis distinguishes analytic general propositions that express conceptual truths from empirical generalizations that can be refuted by experience (MWO 224). "The dividing line between the *a priori* and the *a posteriori* is that between principles and definitive concepts which *can* be maintained in the face of all experience and those genuinely empirical generalizations which *might* be proven flatly false" (1970: 238–39). But he also holds that when we become sufficiently confident of an empirical generalization we may change its status (MWO 262–64, 375, 393–401). For example, ELECTRICAL RESISTANCE was introduced at particular point in the development of physics. Once this concept was established it became useful to measure and tabulate the resistance of various materials. Over time, having a particular electrical resistance R shifted its status from being one of the properties of material M to being a necessary condition for a sample being identified as M; a sample that failed to exhibit the appropriate resistance was no longer classified as M. At this point we associated a different concept with the word "M" and a different proposition with the sentence "M has electrical resistance R" than we did at an earlier stage. Instead of expressing a proposition subject to empirical test, the sentence now expresses an analytic proposition

that encapsulates a defining characteristic of M. Again, there is no point at which considerations of evidence and logic *require* conceptual change. Rather, conceptual change results from a community *decision* that we arrive at a more effective way of dealing with M by taking R as a criterion for M than by continuing to view the relation between M and R as a generalization subject to further test.

Lewis seems to include all our firm beliefs about an item in the associated concept. For example, he tells us that the concept of a toothache includes "the apprehension of what brought it on and the formula for getting rid of it" (MWO 128), and that the modern concept of water includes "the predictable transformation from liquid to solid at 32° F" (MWO 396).

Given Lewis' account of the role of concepts in experience, if he were writing in the 1950s or later he would be viewed as holding that all observation is theory-dependent. It is clear how Lewis would interpret Hanson's much debated claim that Tycho Brahe and Kepler do not see the same thing when looking at the sun (1958, Ch. 1). Lewis holds that perception requires conceptualization, and the two astronomers associate different concepts with the word "sun." As a result, even when they make verbally identical statements in which the word "sun" occurs, these statements have different meanings. Hanson also notes that there must be a clear sense in which Brahe and Kepler are seeing the same thing if the claim that they are seeing different things is to have any epistemological significance (1958: 5, 7). Lewis can easily capture this point because he recognizes that two different conceptual systems may overlap to a large degree, and that this overlap provides the basis for communication (MWO 84–85). Brahe and Kepler presumably have enough overlap in their conceptual machinery to pick out the specific object they are discussing.

Now consider an innovation that Lewis introduces in AKV, where he distinguishes two aspects of the meaning of a term: *linguistic meaning* and *sense meaning*.[28] A term's linguistic meaning is captured by its relations to other terms; it is the aspect of meaning we find in dictionary definitions. If I could totally master all the connections in a dictionary of some largely unfamiliar language, I would have grasped the linguistic meaning of all terms in that language (AKV 132). In MWO Lewis identified meaning with what he now calls linguistic meaning (MWO 67), but in AKV he argues that something would be seriously lacking in my understanding of the meanings of many terms if I grasped only their linguistic meaning.

A term's sense meaning consists of the criteria for identifying instances of the associated concept. Lewis initially describes sense meaning as a "criterion in mind" (e.g., AKV 37, 131) and considers it to be a cousin of Kant's notion of a *schema* (AKV 134). Like Kant, Lewis considers this schema to be a creature of the imagination, although he is willing to back off from this mentalistic account and consider a behavioral version (AKV 144). Whichever way we go, a term's sense meaning is a rule or procedure that we use to recognize instances; it is determined before we apply the term to actual cases (AKV 143–44).

While Lewis describes linguistic meaning and sense meaning as equally important aspects of a term's meaning, he argues that for purposes of epistemic analysis sense meaning is more fundamental. In particular, we appeal to sense meaning to determine if a proposition is analytic (AKV 151–55). The paradigm case of a true analytic proposition has the form "All *A* are *B*," where the criteria for the application of *A* include the criteria for the application of *B* – as the criteria for a square include the criteria for a rectangle. The paradigm of a false analytic proposition is provided by cases in which the criteria for *A* and *B* are mutually incompatible – as is the case for square and circle. Since these determinations can be explored purely by reflection, the resulting knowledge is a priori – although the concepts we are exploring are part of our repertoire because of pragmatic decisions made in response to experience. The introduction of sense meaning is an important innovation in Lewis' theory of concepts that will reappear in our later discussions.

Now consider some advantages and problems of Lewis' approach to concepts. One dividend of his approach is that relational concepts do not pose a special problem. The specific features of a relational concept will be captured in its implications. It is also clear that problems about the nature of the primary vocabulary will not arise for Lewis. In addition, I think that higher-order concepts can be smoothly integrated into Lewis' approach, although he does not do so.

On the other side, the kind of holism that we find in Lewis introduces a problem we have not yet encountered. For Lewis, each concept C is implicationally linked to a large set of other concepts. All of these implications are part of C's content, and implications involving C are part of the content of these other concepts. As a result, any change in a conceptual system generates changes in many concepts. Lewis recognizes this point and suggests, in some too brief remarks, that our conceptual system is hierarchically structured – like a pyramid – and that the impact of conceptual change is greater at higher points in this structure:

> The decision that there are no such creatures as have been defined as "swans," would be unimportant. The conclusion that there are no such things as Euclidean triangles, would be immensely disturbing. And if we should be forced to realize that nothing in experience possesses any stability – that our principle, "Nothing can both be and not be," was merely a verbalism, applying to nothing more than momentarily – that dénoument would rock our world to its foundations.
>
> (MWO 306)

Thus Lewis does not think that every conceptual change impacts every concept, but he gives no developed account of the scope of such impact, and of how it is constrained. Fodor, a persistent critic of holistic theories of concepts, maintains that no such constraints can be provided in a principled

manner, and considers this to be an overwhelming argument against holistic theories of concepts (e.g., 1995: 76; 1998: 37; Fodor and Lepore 1992: 21, 23–26). Yet a holistic approach to concepts has considerable virtues. This is particularly clear when we consider – as Lewis does – the role of concepts as a guide to action in the world. Each concept encapsulates a set of firm beliefs, so the conceptual identification of an item provides a body of expectations about how that item is likely to behave or respond. But it is highly unrealistic to think that every concept is connected to every other concept so that, say, the change from $w_F$ to $w_M$ alters, to some degree, my concept of a pawn in chess. In subsequent chapters I will attempt to meet Fodor's challenge and construct a *local holism* that captures the virtues of this approach without yielding absurd results.

## 3.7 The Analytic-Synthetic Distinction I

As we have seen, the analytic-synthetic distinction is central to theories of concepts in the empiricist tradition. Whether these theories are atomistic or holistic, twentieth century empiricists explicitly, and classical empiricists implicitly, held that conceptual content is expressed only in analytic statements. As a result, all claims about conceptual content can be evaluated a priori and are logically immune from empirical challenge or support. Since empiricists typically restrict a priori knowledge to knowledge of analytic propositions, and hold that philosophy deals only with a priori knowledge, a challenge to the analytic-synthetic distinction is a challenge to the legitimacy of philosophy itself.[29] In this section I will examine two critiques of the analytic-synthetic distinction as it has been used in the empiricist tradition: Quine's argument that the distinction is incoherent and Putnam's argument that the distinction is not exhaustive.

In "Two Dogmas of Empiricism" (1953) Quine attacks two central theses of modern empiricism: the analytic-synthetic distinction and the doctrine that all meaningful synthetic propositions can be reduced to statements in the observation language. But these theses are intimately related since the reduction must take place by means of analytic propositions. Quine underlines this point when he asserts that the dogmas are, at root, the same (1953: 42).

Quine attacks the distinction by challenging the concept of analyticity, but there are three different ways in which such a challenge could proceed. One could argue that the concept:

A1. is incoherent, or
A2. has no instances, or
A3. does not have the significance that has been accorded to it.

Quine's initial argument in "Two Dogmas" pursues the first approach; if successful it should lead us to drop the term "analytic" from our active language. It is also the hardest of the three to carry through; it is unclear how one could accomplish this task short of finding a contradiction. Quine's

approach is to examine several possible explications of the meaning of "analytic" and argue that each is defective for one of two reasons. Some proposals fail because they require that we already understand some other term – such as "necessarily true" or "synonymous" – that is as much in need of clarification as "analytic." Other proposals amount to introducing "analytic" as an arbitrary label that carries no independent significance.

Critics quickly pointed out that this won't do. The expression "analytic" is not a mere collection of letters that has never been assigned a meaning in our language. Rather, it is a familiar term whose philosophical role is well established and for which there are countless well-known illustrations. The failure of several attempts at analysis does not show that no attempt can succeed, nor does it show that we do not understand how to use the term. Philosophical practice going back at least to Plato indicates that we are often able to recognize instances of terms that we cannot rigorously define. If persistent failures of analysis provide a reason for concluding that a concept is incoherent, just about every concept that philosophers have turned their minds to should have been eliminated long ago. But there are informal ways of explaining the meaning of a term, and there is substantial agreement on which statements are analytic and which synthetic. There are, to be sure, unclear cases, but this does not show that we lack one of the concepts required to draw the distinction.

I think these replies are sufficient for us to conclude that Quine's first argument against the analytic-synthetic distinction fails. But Quine's attempt to show that the concept of analyticity is incoherent amounts to overkill. It would be quite sufficient for his purposes to argue that this concept lacks instances, and should be relegated to the same status as phlogiston, telegony, and $w_F$. Quine's remaining argument in "Two Dogmas" seems to be of this second type. He acknowledges one necessary condition for analyticity – analytic propositions are not subject to empirical challenge – and maintains that no proposition has this property. Rather, all of our beliefs link together in a single seamless web that impinges on experience only at the edges. The goal of science (understood as including all empirical beliefs) is to facilitate the prediction of future experience, so we are concerned to fit this web to our sensory promptings – which are the "edges" of the net metaphor. We pursue this end by making adjustments in the web when we encounter conflicts between web-generated expectations and experience. Any belief can be revised as we accommodate experience – although some beliefs are more central to the web than others. We are more reluctant to revise central beliefs because this will have a greater effect on the overall web than will revision of beliefs closer to the periphery. The holistic character of the web allows us to hold selected beliefs immune to revision and make the changes elsewhere, but no belief is intrinsically immune to reconsideration under pressure from experience. Which beliefs we alter in a specific case is a pragmatic matter: We decide which revisions to make as we pursue the goal of predicting future experience in the simplest, most economical way.

There are striking similarities between Quine's view and Lewis'. Indeed, Quine ends "Two Dogmas" by drawing an explicit comparison between his approach and those of Carnap and Lewis. These philosophers, Quine says, also provide pragmatic accounts of the choice of languages and conceptual schemes, but "their pragmatism leaves off at the imagined boundary between the analytic and the synthetic" (1953: 46).[30] Thus Quine describes himself as espousing "a more thorough pragmatism" (1953: 46) in which every one of our epistemic decisions is made on pragmatic grounds. It will be useful to look more closely at just where Quine and Lewis' disagree.

The key effect of Quine's more thorough pragmatism is to eliminate a distinction in Lewis that is, from the perspective of behavior in the world, wholly artificial. Lewis, we have seen, tries to maintain a distinction between empirical refutations and conceptual change, treating only the latter as resulting from a pragmatic decision. For Lewis there are two different ways in which we reject a statement: If it is synthetic we can reject it as false; if it is analytic we continue to acknowledge its truth but reject it as irrelevant to our concerns. Quine treats this as a distinction without a difference. Recall Lewis' claim that "The dividing line between *a priori* and *a posteriori* is that between principles and definitive concepts which *can* be maintained in the face of all experience and those genuinely empirical generalizations which *might* be proven flatly false" (1970: 238–39). Quine holds that *any proposition* can be maintained in the face of all experience, and that there is no such thing as a proposition being "proven flatly false." But a proposition that is protected at one stage in our cognitive history can be rejected at another stage as we continue accommodating to experience. In every case the decision to retain or reject a proposition is pragmatic.

Quine's account allows for all the options we encountered in Lewis. At any point in time the web is made up of propositions we accept. We may drop propositions from the web or add new propositions. In making these changes we may drop all propositions in which a particular term occurs, or add new propositions containing that term, or alter the web so that this term now enters into different connections than it did previously. We may add propositions containing a new term, and do so at any place in the web. Lewis allows for two kinds of empirical refutation that are also captured in the Quinean scheme. Withdrawing an identification amounts to rejecting a proposition that is close to the periphery and has few connections to other propositions. Rejecting an empirical generalization amounts to rejecting a claim that is a bit farther from the periphery and has relatively few connections to other propositions in the web. The cases that Lewis describes as conceptual change amount, for Quine, to changes even farther from the periphery, that have a greater impact on the web. Both Lewis and Quine extend this account to the truths of logic which they locate, respectively, at the apex of the pyramid, and the center of the web. The outcome of any of these changes is the same for both philosophers: We have a more or less modified scheme with which to generate expectations and guide behavior.

Now consider Putnam's (1962) response to Quine. Putnam thinks that Quine is wrong on two key points: the concept of analyticity is coherent, and there are analytic statements. But Putnam agrees with Quine that the philosophical self-conception of analytic philosophers should be revised because analytic statements are trivial and thus will not sustain the load philosophers have placed on them.

> I think that Quine is wrong. There are analytic statements: "All bachelors are unmarried" is one of them. But in a deeper sense I think that Quine is right; far more right than his critics. I think that there is an analytic-synthetic distinction, but a rather trivial one. And I think that the analytic-synthetic distinction has been so radically overworked that it is less of a philosophic error, although it is an error, to maintain that there is no distinction at all than it is to employ the distinction in the way it has been employed by some of the leading analytic philosophers of our generation.
>
> (1962: 361)

In other words, Putnam takes up the third of the options mentioned at the beginning of this section. Instead of placing analytic propositions at the center of philosophic research, Putnam argues that to understand the nature of scientific knowledge we need to recognize a *third class* of propositions that does not fit the standard empiricist dichotomy between analytic a priori and synthetic a posteriori propositions. Propositions in this third class are not analytic since they are subject to empirical challenge, but they are not ordinary synthetic propositions because we protect them from refutation by isolated experiments. Here is an elegant description of how such propositions work:

> A philosopher, on being asked how much smoke weighs, made reply: "Subtract from the weight of the wood burnt the weight of the ashes which are left over, and you have the weight of the smoke". He thus presupposed as undeniable that even in fire the matter (substance) does not vanish, but only suffers an alteration of form.
>
> (Kant 1963: 215)

As my example suggests, there is an important parallel between Kant's synthetic a priori propositions and propositions in Putnam's third class. It will be helpful, then, to recall that Kant introduced the analytic-synthetic distinction in the context of another distinction between a priori and empirical knowledge. This generates four types of propositions, but only three of these need be considered here: analytic a priori, synthetic empirical, and synthetic a priori. Philosophers in the empiricist tradition have regularly denied the existence of synthetic a priori propositions, identified the analytic with the a priori, and the synthetic with the empirical.

Putnam and Quine are both working in this tradition, which is why Putnam can describe himself as calling attention to a third class of propositions. Putnam notes that the characteristic feature of empirical generalizations is that they are falsifiable by isolated counter-examples (1962, e.g., 363, 372, 374); "All swans are white" is a classic case. No empirical counter-examples are possible for analytic propositions. Propositions that face empirical counter-examples, but that we choose to protect, thus form a third class.[31]

Putnam does not provide a name for propositions in this third class, but I will refer to them as *guiding assumptions* (henceforth, GAs).[32] Members of this class are not synthetic a priori propositions because they are not known a priori. They are adopted on empirical grounds and protected from refutation for substantial periods of time; but they can be overthrown on empirical grounds under appropriate circumstances.[33] As long as a proposition of this sort is accepted, it plays some of the epistemic roles that Kant attributed to synthetic a priori propositions. In particular, GAs limit the number of options available in responding to an empirical challenge, and thereby guide research. Note especially that the ability of these propositions to guide research is directly related to the fact that we understand which empirical results can – from a purely logical perspective – count as empirical challenges.

I want to pursue a further parallel with Kant. Kant admitted two sources of synthetic a priori propositions, the forms of sense and the categories; I am concerned only with the latter. Since Kant held that our knowledge of these propositions derives from our grasp of concepts that are not derived from experience, synthetic a priori propositions have a special tie to a small number of central concepts. While Putnam does not admit such special concepts, he does hold that GAs have a special tie to the central concepts of a scientific discipline, and that changes in the GAs of a discipline involve changes in its conceptual framework. Many readers will note the similarity between this view of Putnam's and some aspects of Kuhn's notion of a paradigm – published in the same year as Putnam's paper.[34] We will see in Ch. 4 that a parallel notion plays a central role in Sellars' theory of concepts.

## 3.8 Conclusion

One goal of this chapter was to present the historical context in which Sellars worked in developing his theory of concepts. Putnam's account of the analytic-synthetic distinction takes us beyond the historical point at which Sellars did his most influential work, but I included that material because it introduces a central Sellarsian theme that will appear in a somewhat different guise. In any case, lives, careers, and the development of ideas often overlap and rarely fit between sharp temporal boundaries. Some will object to the absence of any discussion of Wittgenstein in this chapter. Wittgenstein is a central figure in the development of the theory of meaning, and had a

significant impact on Sellars. I have not included such a discussion because Wittgenstein interpretation is a minefield. Any discussion of Wittgenstein's contributions would require that I defend a particular interpretation, and this would be a major distraction from my concerns in this book. As I proceed I will consider some views that have been attributed to Wittgenstein, but will do so because the views themselves are of interest independently of whether the attribution is historically accurate.

Another goal of this chapter was to introduce some problems that a theory of concepts must address, and that have raised difficulties for earlier theories – especially basic, relational, and higher-order concepts. In addition, a comprehensive theory of concepts must apply to the concepts used to formulate that theory. I will return to these issues as I move towards my own theory.

# 4 Sellars: Exposition, Interpretation, and Critique[1]

> Categorization is not an end in itself but provides access to categorical inferences. Once an entity is categorized, knowledge associated with the category provides predictions about the entity's structure, history, and behavior, and also suggests ways of interacting with it.
>
> (Barsalou 1999: 16)

Sellars, like Lewis, developed a modified Kantian epistemology. Experience, for Sellars, normally occurs in the context of some system of concepts, and we deal with conceptually interpreted experience in our everyday lives and in science. Sellars agrees with Lewis in holding that we change our concepts over time, and that conceptual change is especially clear in the development of science. Sellars also holds that a holistic element plays a central role in determining conceptual content, although we will see that his full account of conceptual content is more complex than Lewis' account. Sellars goes further than Lewis in attacking the doctrine that pure sensory "givens" – of the sort typified by simple ideas and sense data – provide the foundations of semantics or empirical knowledge. But, we will see, sensory inputs play a key role in Sellars' epistemology and theory of concepts. In addition, unlike Lewis and most analytic philosophers, Sellars rejects the view that *only* analytic propositions express conceptual content. This thesis will provide a central theme in our discussion.

In his overall epistemology Sellars breaks sharply with Kant, Lewis, and logical empiricism in his advocacy of scientific realism.[2] Sellars holds that there are items in the world that do not appear in our sensory experience; that such items may have properties that are quite different from any we encounter in ordinary experience; but that these items are knowable through the long-term process of scientific research. The development of concepts that describe such items is a crucial step in achieving this knowledge. These concepts are not available prior to this research, and conceptual innovation is thus a central feature of scientific progress. This research also leads to improved predictions, and increasing our predictive ability is a goal of science, *but not the only goal*. Sellars' account of how research leads to

improved predictions involves another major disagreement with logical empiricism. Logical empiricists took low-level generalizations over experience to be a fixed point for future research, and held that we seek wider generalizations that explain those already discovered. Sellars maintains that as science develops we regularly discover that accepted generalizations are not correct. As we develop better theories – typically involving new concepts – we replace these generalizations with more accurate successors, while also explaining why the older generalizations achieved the degree of accuracy that they did. Sellars backs up this view of scientific innovation with a theory of concepts that (among other things) provides an account of how conceptual change occurs in a coherent manner – which requires that new concepts be anchored in existing concepts – while also introducing genuinely new content. This theory of concepts will be my main concern in this chapter.

Sellars usually follows the common practice of treating "language" and "conceptual system" as interchangeable, and thus treating "theory of concepts" and "theory of meaning" as synonyms. However, Sellars does not actually consider the two domains to be identical. For example, he notes that languages includes meaningful terms that do not function as concepts and offers "alas" as an example (LT 115); I will consider Sellars' specific reason for this claim below. This suggests that the scope of language is wider than the range of our concepts, but Sellars also holds that scope of language is, in different respects, narrower than that of conceptual systems. For example, Sellars is open to the possibility that non-linguistic animals have representational systems that are similar to our own. Thus he notes that he is expanding the scope of "language" when he uses the term "in our broad sense in which 'language' is equivalent to 'conceptual structure'" (SRLG 340). I will return to the relation between concepts and language in Sec. 5.1; in the present chapter I will be concerned only with human concepts where linguistic information is a major source of evidence about conceptual content. Thus, for now, I will follow Sellars' usual practice of treating languages and conceptual systems as the same.

## 4.1 Conceptual Status

One central, and attractive, feature of Sellars' theory of concepts is the explicit recognition that there are different kinds of concepts whose content is determined in different ways. However, he holds that one feature is common to all concepts: every concept derives at least a part of its content from implicational relations to other concepts, so that concepts occur only as members of systems of inter-related concepts. Without such relations we do not have a concept at all, so these relations confer *conceptual status* (SAP 316–17),[3] although "the 'conceptual status' of a predicate does not exhaust its meaning" (SAP 316). Additional features besides conceptual status will play the key role in distinguishing different types of concepts. The context of

the quoted remark makes it clear that Sellars' is discussing what we will come to refer to as *descriptive concepts*, which include most of the familiar concepts of everyday life and science – table, planet, person, electron, and such. I will focus discussion in this section on descriptive concepts until the final two paragraphs.

Sellars' view of conceptual status underlines the holistic element at the heart of his theory of concepts. The key reason for adopting a holistic approach is already present in Lewis (and also in Quine, although Quine prefers not to discuss the issue in terms of concepts and meanings). Let us ask why we have concepts at all; one way of approaching this question is by considering why concepts are useful. Paradigmatic uses of concepts include identifying items and distinguishing items of different kinds, and these activities are useful because the concepts by which we identify items carry information about them. Consider COMPUTER. In order to have this concept we must have some beliefs about computers: perhaps that they are manufactured objects produced by a technologically advanced society, require a source of electricity to operate, are capable of being programmed to do numerical calculations or run a word processor, and much more. To recognize an item as a computer is to apply these beliefs to that item – which allows us to determine what we can do with the object, and how to behave with respect to it. Similarly, we acquire useful information about an item if we can identify it as edible, or a bomb, or a poisonous snake. If I identify an item as a book I may proceed to read it, but not use it to make a telephone call; if I identify an item as an egg, I should know better than to use it as a support under a short leg of a table.

The point of the previous paragraph can be summarized by noting that subsuming an item under a concept brings our current beliefs to bear on that item. This suggestion can be developed further by considering the difference between a *concept* and a *label* (cf. CDCM 306–7). If I encounter a totally unfamiliar object I may label it *L* for ease of reference, but this does not allow me to draw any conclusions about that object.[4] Conceptually competent adults rarely engage in mere labeling. Even to identify an item as a physical object is to subsume it under a concept and thereby license a number of beliefs about that item – such as that it will not suddenly vanish, and that its visual properties will be correlated with tactile properties. When we encounter an unfamiliar item we quickly attempt to move from just putting a tag on it to forming an appropriate concept. Sometimes we engage in risky behavior towards this end, such as poking the item to determine if it is hard, or hot, or carries an electric charge. As we gather information an initial label begins to acquire content, and we begin constructing a concept. When Röntgen encountered an unexpected darkening of a photographic plate in his laboratory he labeled its cause "X" and began to explore the properties of this cause. By the time he was ready to announce the discovery of X-*rays* he had already established many of those properties and conceptualized the cause as a form of radiation.

Let me put the point another way. Descriptive concepts are cognitive tools that we use to think about various subjects and to find our way around various domains. When I subsume an item under a concept I am integrating it into my belief system, and I am thereby primed to infer various features of that item. Which inferences I actually make will depend on contextual factors, but the key point is that identifying an item implies that other concepts also apply to it. Each concept is, from this perspective, the locus of a set of permissible inferences, and it is these inferential ties between concepts that lie at the basis of Sellars' notion of conceptual status. Historically, the most important contrast to this view comes from empiricist theories which hold an atomistic view of our primary concepts. On theories of this type primary concepts are labels. But we saw in Ch. 3 that even on these theories most of our cognitive work is done by secondary concepts that are supposedly built out of these basic concepts and that carry information about the items they describe – information carried by relations to other concepts.

However, in contrast to Lewis, Sellars defends a *local* holism. Sellars does not hold that all concepts link together into a single, massive conceptual system – even a hierarchically structured system. Nor does Sellars hold, in the manner of Quine, that all of our beliefs are members of a single web. Although Sellars does not explicitly describe himself as adopting a local holism, he makes many remarks which indicate that he thinks of us as having multiple, distinct systems of concepts. For example, Sellars tells us that "we can stick to English and yet be said to speak not one language but many" (LRB 312). At the beginning of EPM Sellars says that his goal is to attack "the entire *framework of givenness*" (EPM 128); "framework" is another Sellarsian term for a conceptual system. Sellars also describes modern physics as rejecting the common sense framework of colored physical objects existing in space and time, and replacing it with a different framework (EPM 173). He talks about the distinct frameworks of molar behavior theory and the microtheory of physical objects (EPM 193), and empirical and theoretical frameworks (TE 70). He treats our common notions of space and time as distinct conceptual frameworks (SRII 181), and distinguishes the everyday framework of things from the framework of events which, he says, is a legitimate alternative invented by philosophers (TWO 553–54). Local holism fits well with Sellars' scientific realism: he holds that as a science develops, scientists replace existing systems of theoretical concepts with different systems. In a similar way, Sellars treats traditional empiricism as a conceptual system for thinking about knowledge and meaning, and aims to replace that system.[5]

I view this move to local holism as one of the many virtues of Sellars' theory of concepts, and in developing this theme I will go beyond the hints in Sellars' texts. The key idea is that we should think of individuals (and societies) as deploying multiple, more or less distinct, conceptual systems. Consider some of the systems that individuals deploy in their everyday and

professional lives. We have systems of concepts for describing kinship relations, political systems, and the various games we play. People who have a special interest in a class of items will often have detailed conceptual systems for describing these items, systems that are not shared by those who lack these interests. Furniture designers and antique collectors, for example, will have elaborate systems of furniture concepts that many of us do not share. Stamp collectors will describe properties and make distinctions that are not in the repertoire of many people. Those in specific trades – say, plumbers or astronomers – will have concepts for describing their characteristic tools, materials, and tasks. Depending on one's vocations and avocations, an individual may have a more or less elaborate system of concepts for thinking about musical compositions, grammatical distinctions, stock options, elementary particles, baseball, carpentry, and so forth. These conceptual systems are largely distinct. People can have virtually identical systems of concepts for thinking about baseball quite independently of whether they also share an interest in – and the conceptual systems appropriate to – classical music, bridge, quantum field theory, or futures contracts. Often we can learn a new conceptual system without alerting other systems we have already mastered. I could, for example, learn about cricket without this having any impact on my understanding of propositional logic, or foundational epistemology, or a wide variety of other conceptual systems that I use in my everyday, professional, and recreational activities.

I do not want to understate the complexity of the issues involved. For while we can often treat various conceptual systems as distinct, relations between systems may not always be apparent, and these relations may change in complex ways. Sometimes we discover links between two subjects that had been considered distinct, and this can have profound significance – including generating significant change in the concepts used in each system. The history of physics is full of examples of this sort, from the collapse of the Aristotelian distinction between celestial and terrestrial realms, to the integration of space and time, energy and mass, and geometry and gravitation in relativity theory. In a similar way, modern reproductive technologies raise questions about laws governing inheritance and contracts. Another complexity occurs when two incompatible systems for describing a subject matter coexist in an individual mind and people shift between them, depending on their current interests, without confusion. One example is provided by the different frameworks for thinking about space and time found in special relativity and in everyday life. One may believe that the relativistic spacetime structure is the better candidate for describing the physical world while finding the older system sufficient and more convenient for planning a vacation or making a date. Such cases raise the question of when we should think of alternative systems as describing the same subject matter, as well as what constitutes a single conceptual system. Sellars does not discuss these issues in any detail; I will postpone further discussion until Sec. 5.8.

Given Sellars' view of conceptual status, he distinguishes three types of conceptual systems, and thus three types of concepts.[6] *Formal concepts* are the concepts of logic and pure mathematics; they include such examples as CONJUNCTION, ENTAILMENT, DERIVATIVE, and GROUP. The characteristic feature of formal concepts is that their content is *completely* determined by relations to other concepts in a system. Most of the concepts mentioned thus far in this book are *descriptive concepts*, our second type. These are the concepts we use when we identify an item as a table, a planet, a noun, a democracy, a capital gain, and so forth. While these concepts also occur only as members of systems of concepts, they have two features that distinguish them from formal concepts. First, while relations between concepts in a formal system are mediated by logic alone, Sellars maintains that there are additional relations among descriptive concepts that are mediated by synthetic propositions which are, at least for a time, treated as necessary truths. These synthetic propositions provide the basis for *material rules* which license further inferences between concepts in addition to those licensed by formal logic. Second, Sellars holds that descriptive concepts must be related to their extra-systemic subject matter by *entry transitions*. Roughly, these are non-inferential moves from some extra-systemic item we encounter into the conceptual system we use for describing that item. In effect, we make an entry transition whenever we spontaneously subsume an item under a concept.

Sellars' third class consists of *prescriptive concepts*. These include the moral concept OUGHT along with other evaluative concepts such as LOGICAL VALIDITY and EPISTEMIC JUSTIFICATION. The distinguishing feature of these concepts is that their content is jointly constituted by relations to other concepts and by *departure transitions*. The idea is that while entry transitions are moves from the world into a conceptual system, departure transitions are moves from the system to the world. For example, having decided to sit on a chair, actually sitting would be a departure transition. Many departure transitions are purely voluntary, but prescriptive concepts *require* departure transitions. Exactly what this means, and the exact role of departure transitions in determining the content of prescriptive concepts will be discussed in Sec. 4.4. I turn now to a detailed exposition and, where needed, critique of Sellars' accounts of these different kinds of concepts.

## 4.2 Descriptive Concepts I

Descriptive concepts typically describe items other than themselves.[7] The concepts we use to describe familiar objects and their properties – concepts such as table, horse, red, and hard – are clear examples. As a scientific realist, Sellars includes theoretical concepts such as gene and quark among the descriptive concepts, where the adjective "theoretical" describes our mode of access to these items and our reasons for believing they exist, not a mode of existence (cf. Sec. 3.5). Thus Sellars notes that it is a mistake to

reify "the *methodological* distinction between theoretical and non-theoretical discourse into a *substantive* distinction between theoretical and non-theoretical existence" (EPM 174). GENE and QUARK are introduced to describe items in the world in the same sense in which TABLE and HORSE describe such items.

Of course, the theoretical entities postulated at a given stage in the development of science may not exist, and this possibility underlines an important point: *the term "descriptive" is being used here to indicate the function of a concept*. A descriptive concept is introduced to describe an item that may exist, but it often takes further research to decide whether the concept has a referent. In this usage phlogiston, telegony, and green-flying-horse are descriptive concepts. Phlogiston, for example, was introduced to describe an item believed to be emitted in combustion (and related processes). Further research eventually led to the conclusion that phlogiston does not exist, but this outcome would not have been possible without a determinate concept of phlogiston to guide that research. In other words, *we must distinguish sharply between discussing the content of a descriptive concept and discussing whether that concept is instantiated*; we must have the concept C before we can inquire whether there are any Cs. When we undertake this inquiry several outcomes are possible. We might conclude that Cs exist, or that neither Cs nor anything like them exist, or that while strictly speaking no Cs exist, there are items in the world that are similar to Cs in important respects. This last outcome may lead us to change the concept we are introducing, and if we adopt this course we must then decide whether to associate the word "C" with this new concept.[8] The thesis that we undertake research in a domain with a set of concepts, and that research may lead to alterations in this set, is central to Sellars' view that one major task of research is to *find* the correct concepts for describing various domains. For the remainder of this section I will use the term "concept" only for descriptive concepts unless otherwise stated.

Sellars usually limits his discussions of descriptive concepts to those representing items we can detect with our unaided senses plus the postulated items of natural science; however, the notion of a descriptive concept has considerably wider scope. We have, for example, a variety of grammatical concepts that we use to describe features of languages. We also have a variety of metaphysical and theological concepts, such as soul, angel, and saint. A theory of concepts should allow us to understand how these concepts get their content – as distinct from the question of whether these concepts have instances. It remains a possibility that in some of these cases there is no concept, just an empty word; explaining this difference is another task for a theory of concepts.

Since the function of descriptive concepts is to describe, their content must include some features of the items they describe. Sellars holds that we include all those features we confidently believe to be properties of Cs in the associated concept, and that we exhibit these beliefs when we infer one

concept from another. This is how descriptive concepts guide our behavior and thought with respect to items that concern us. Thus we seek accurate descriptive concepts, and alter the concepts we use to describe items as we learn more about them. This is a highly controversial claim and I want to consider Sellars' reasons for defending it.

Our topic impinges directly on the status of the analytic-synthetic distinction. On the usual accounts, analytic propositions express the content of concepts, while synthetic propositions use these concepts to make additional claims about items that fall under them; assessment of the truth-value of a synthetic proposition has no impact on the content of the concepts that occur in that proposition. Sellars is surprisingly reticent on the analytic-synthetic distinction. He does not reject the distinction, but he may never have arrived at a settled view of which propositions are analytic and which synthetic. For example, he begins SAP by noting, along with Quine, that "analytic" is used in two senses: for truths of logic and for claims that are true by virtue of the meanings of their terms. But the main goal of SAP is to argue that there are propositions that are true *ex vi terminorum*, but not analytic. In this paper Sellars restricts the use of "analytic" to logical truths (SAP 298–99). In a closely related paper Sellars appears to identify analytic truths with formal truths and to suggest replacing the analytic-synthetic distinction with the formal/material distinction (SRLG 331). Elsewhere he asserts that the analytic-synthetic distinction applies only to predicates that have necessary and sufficient conditions of application, and that predicates lacking these are "much more prevalent than logicians have hitherto realized" (EAE 438, n. 10). He also holds that explicit definitions are analytic (SAP 302). One remark (published in 1953) is particularly revealing:

> I am convinced, however, that much of the current nibbling at the distinction between analytic and synthetic propositions is motivated by what I can only interpret as a desire to recognize the existence of synthetic *a priori* propositions while avoiding the contumely which the language traditionally appropriate to such a position would provoke.
>
> (IM 338)

We will see shortly that what Sellars refers to as "synthetic a priori propositions" are identical with those I have described as "guiding assumptions." Sellars consistently maintains that such propositions play a fundamental role in epistemology and semantics. Indeed, it is central to Sellars' account of descriptive concepts that their content cannot be completely captured in analytic propositions or in explicit definitions. Instead, he holds that the content of descriptive concepts is largely determined by implicit definitions that are expressed by synthetic propositions. Thus Sellars rejects the view that *only* analytic propositions express meaning, and this thesis is crucial to his epistemology and philosophy of science, as well as to his theory of concepts. Sellars account of implicit definitions is closely related to his

doctrine of *material rules of inference*; we must consider these doctrines in some detail.

### *4.2.1 Material Rules of Inference*

Suppose I am confident that the synthetic proposition "All *A* are *B*" is true and that an item before me is an *A*. I can infer that this item is also a *B* via the argument:

All *A* are *B*.

*x* is *A*.

So, *x* is *B*. (A1)

This inference is justified by a rule of formal logic, but Sellars notes another way of analyzing the inference. Accepting the proposition "All *A* are *B*" can be viewed as equivalent to accepting a rule that allows us to infer "*x* is *B*" from "*x* is *A*." This leads to an argument with only one premise:

*x* is *A*.

So, *x* is *B*. (A2)

The rule that licenses this argument is a *material rule of inference*.

Treating universal generalizations as material rules of inference is familiar from the work of some logical positivists. Early in their history positivists required that a meaningful non-analytic proposition be capable of conclusive empirical verification or falsification. A watershed in the development of positivism came in the mid-1930s when positivists finally became clear that universal generalizations cannot be conclusively verified. Rather than reject scientific generalizations as meaningless, the majority response, typified by Carnap (1996, originally published in two parts in 1936–37), was to relax the demand for conclusive verification and require only that empirical evidence be *relevant* to the evaluation of synthetic claims. But some positivists (notably Schlick and Waismann) took a different tack: Since rules do not have a truth-value, they are not within the scope of the verification theory of meaning. Thus treating universal generalizations as rules of inference allowed them to save the strict verification criterion; it also supported the instrumentalist interpretation of science that they favored.

Now Sellars is a scientific realist who holds that universal generalizations play a central role in science. Nevertheless, there is a general correspondence between universal generalizations and rules of inference, and Sellars holds that *for certain purposes* firmly accepted universal generalizations should be replaced by the associated rules. As a result, arguments

of type A1 should be replaced by arguments of type A2. To understand Sellars' position we must carefully distinguish between the object-language level, where we are using a set of concepts to think about or act in response to items in some domain, and the metalinguistic level where (among other matters) we are discussing concepts. Let us begin with the object language, focusing on concepts that describe items available to everyday perception, and the use of these concepts in practical situations. I want to make two points.

First, there is a pragmatic element in Sellars' philosophy. He views humans as active beings who must deal with an independent world, and a central function of conceptual systems is to guide action in the world (SRLG 339–41). Thus while Sellars rejects pragmatic analyses of meaning and truth, he maintains that if we reformulate pragmatism as the thesis that the connection between language and conduct "is intrinsic to its structure as language, rather than a 'use' to which it 'happens' to be put, then Pragmatism assumes its proper stature as a revolutionary step in Western philosophy" (SRLG 340). On this view concepts are cognitive tools that guide our responses to items we encounter so that identifying an item as a C amounts to acquiring a license to infer a variety of conclusions about that item. In practice these inferences are of type A2 rather than type A1, and there is a good practical reason for this: Inferences of type A2 are typically faster because they require fewer premises. Indeed, Sellars holds that for practical purposes these inferences should be habitual.

Second, inferences of type A1 are licensed by rules that correspond to analytic propositions. But in practical contexts it is unimportant whether the proposition that licenses an inference is analytic or synthetic. What matters is that a conceptual identification of an item provides a guide for dealing with it. If my identification of an item as a beached whale results in my automatically concluding that it is able to breathe, it is unimportant for my subsequent treatment of this animal whether the rule that licensed this inference was based on an analytic or a synthetic proposition.

This second point provides Sellars' main reason for holding that, whatever function the analytic-synthetic distinction may have, it is not a distinction between propositions that are constitutive of meaning, and those that are not constitutive of meaning. Rather, a descriptive concept is a locus of inferences, and every inference licensed by "*x* is *C*" is part of the content of C. Many of these inferences will be licensed by material rules so that by adopting material rules we build our firm beliefs about items into the concepts we use to describe them. As a result, material rules give part of the content of every descriptive concept.[9] As Sellars notes (IM 317–22; PT 292–93), there is a close parallel between material rules and Carnap's P-rules (1959):

> My only quarrel with Carnap is that he commits himself to the thesis that P-rules are a luxury which a language with factual predicates can take or

> leave alone. I have argued . . . that P-rules, or material rules of inference . . . are as essential to a language as L-rules, or formal rules of inference.
>
> (PT 293)

Now let us move to the metalinguistic level where material rules are stated and evaluated. In the metalanguage we find a one-one correspondence between material rules and synthetic universal generalizations: Every synthetic universal generalization can be matched with a material rule of inference, and conversely.[10] From an epistemic perspective the generalizations are fundamental and the material rules derivative since it is generalizations that we justify or refute. Yet, Sellars holds, when we become confident that a synthetic generalization is true we give it a special place in our thinking. At this point we are no longer testing the generalization, but using it as a basis for dealing with items in its domain. The generalization becomes a fixed point that is not subject to empirical challenge even in the face of evidence that, from a purely logical perspective, could be viewed as a counter-instance. The proposition functions as if it were a necessary truth – indeed, synthetic a priori truth – although under appropriate circumstances it can be reconsidered. The thesis that firmly believed propositions are treated as if they were necessary truths – for a time, and in a specific domain – is a central and recurring theme in Sellars' thought.[11] In Sec. 3.7 I introduced the term "guiding assumption" (GA) for such propositions; I will use this term henceforth.

Since a rule is associated with every universal proposition, accepting a proposition as a GA is equivalent to accepting the legitimacy of the inference licensed by that rule. Moreover, for practical purposes it is the rule that is important since it allows us to infer directly from one concept to another. Thus Sellars holds that once we accept a GA we should undertake to modify our behavior so that the associated inference becomes spontaneous when we are *using* the object language:

> suppose that "$\Phi$" and "$\Psi$" are empirical constructs and that their conceptual meaning is constituted, as we have argued, by their role in a network of material (and formal) moves. Suppose that these moves do not include the move from "x is $\Phi$" to "x is $\Psi$". Now suppose that we begin to discover (using this frame) that many $\Phi$'s are $\Psi$ and that we discover no exceptions. At this stage the sentence "All $\Phi$'s are $\Psi$" looms as an "hypothesis", *by which is meant that it has a problematical status with respect to the categories of explanation*. In terms of these categories we look to a resolution of this problematical situation along one of the following lines.
>
> (a) We discover that we can derive "All $\Phi$'s are $\Psi$" from already accepted nomologicals. (Compare the development of early geometry.)

> (b) We discover that we can derive, "If C, then all $\Phi$'s are $\Psi$" from already accepted nomologicals, where C is a circumstance we know to obtain.
> (c) We decide to adopt – and teach ourselves – the material move from "x is $\Phi$" to "x is $\Psi$". In other words, we accept "All $\Phi$'s are $\Psi$" as an unconditionally assertable sentence of L, and reflect this decision by using the modal sentence "$\Phi$'s are *necessarily* $\Psi$". This constitutes, of course, an enrichment of the conceptual meanings of '$\Phi$' and '$\Psi$'.
>
> (SRLG 357)

Note especially the final sentence: When we adopt a new material rule we alter the concepts involved in that rule. As a result, terms acquire new meanings, and this part of their meaning is expressed (in the metalanguage) by a synthetic proposition S. Thus even though S is not analytic, when we reflect on the meanings of its terms we find that S is true *ex vi terminorum*. But this "truth in virtue of meaning" is the result of prior decisions about what to include in these meanings (cf. CDCM 287–88).

As Sellars points out (IM 337–38; PT 293–94), there is substantial similarity between his views and those of Lewis; it is worthwhile pinning down just where they agree and disagree. Both hold that associating a new inference with a term changes the meaning of that term, and they agree that the motivation for the conceptual change derives from the evidence that has convinced us that all $\Phi$ are $\Psi$. They also agree that there is no epistemically significant sense in which the evidence *forces* this change. Rather, conceptual change requires a decision (cf. SLRG 358 where Sellars reiterates that the change is the result of a decision, and CDCM 287–88, 297). Moreover, Lewis and Sellars agree that on the level of the object language – that is, from the perspective of someone using the language to deal with items in the relevant domain – we have a license to immediately infer "x is $\Psi$" from "x is $\Phi$." But when we consider the metalinguistic grounds for this inference, Lewis insists that it derives from an analytic proposition while Sellars holds that it derives from a synthetic proposition. In other words, Lewis contends that after the conceptual change "All $\Phi$ are $\Psi$" is analytic, while Sellars contends that this sentence remains synthetic, but that we now treat it as a necessary truth:

> we have come out with C. I. Lewis at a "pragmatic conception of the *a priori*". Indeed, my only major complaint concerning his brilliant analysis in *Mind and the World Order*, is that he speaks of the *a priori* as *analytic*, and tends to limit it to propositions involving only the more generic elements of a conceptual structure (his "categories"). As far as I can gather, Lewis uses the term "analytic" as equivalent to "depending only on the meaning of the terms involved". In this sense, of course, our *a priori* also is analytic. But this terminology is most unfortunate, since

> in a perfectly familiar sense of "synthetic", some *a priori* propositions (including many that Lewis recognizes) are synthetic and hence *not* analytic (in the corresponding sense of "analytic"). That Lewis does not recognize this is in part attributable to his ill-chosen terminology. It is also undoubtedly due to the fact that in empirically-minded circles it is axiomatic that there is no synthetic *a priori*, while the expression itself has a strong negative emotive meaning. Whether or not it is possible to rescue this expression from its unfortunate associations I do not know. I am convinced, however, that much of the current nibbling at the distinction between analytic and synthetic propositions is motivated by what I can only interpret as a desire to recognize the existence of synthetic *a priori* propositions while avoiding the contumely which the language traditionally appropriate to such a position would provoke.
>
> (IM 327–28)

I will argue that Sellars' view is preferable for two reasons. The first reason may be described as "pragmatic"; the second is a matter of epistemic principle.

In order to understand the pragmatic difference we must keep in mind that a theory of concepts should provide a basis for studying systematic changes in concepts as knowledge develops. Sellars' approach suggests that we think of such changes as typically involving relatively small alterations of existing concepts; it is then a short step to making detailed comparisons of concepts at various stages in the development of a subject. There is nothing in Lewis' view that prevents this kind of comparison. Given the concepts associated with a word at two different stages, we can compare the ways in which those concepts are similar and different. But the tenor of Lewis' approach does nothing to encourage this kind of comparison. Rather, on Lewis' picture, one set of analytic propositions in our active repertoire has been replaced by another, end of story. On Sellars' approach such comparisons are a natural step. Moreover, we will see that on the full Sellarsian theory of concepts examination of licensed inferences is just one of the dimensions on which concepts can be compared.

The difference of epistemic principle involves a point at which Sellars is closer to Kant than is Lewis. We saw in Sec. 3.7 that GAs guide behavior in particular domains. Once we have accepted "All $\Phi$ are $\Psi$" as a GA, if we encounter a $\Phi$ that is not $\Psi$ we are directed to look for some other factor that is responsible for this deviation (such as failure to include the weight of the smoke in Kant's example). This is a common form of everyday and scientific behavior that makes sense if the proposition in question is a GA since we can encounter items in experience that appear to contradict GAs. It is not clear why we should engage in this behavior on Lewis' view given that *no experience can contradict an analytic proposition.* For Lewis, when an identification fails we have only two options – withdraw the identification or change concepts. There is no room in Lewis' account for the crucial option of

retaining both our identification and our concepts while seeking another reason for the anomaly. The concept of a GA, then, can play a role in explaining features of human epistemic behavior that analytic propositions cannot, *as a matter of principle*, play. Treating these propositions as analytic explains why we consider them immune from empirical refutation, but cannot explain how they guide research.

GAs express our current understanding of the laws of nature – where the expression "law of nature" applies to any synthetic generalization that we take to be established, and are ready to use as a basis for action and research. This status accrues to a proposition as a result of our decision to hold it immune from empirical refutation. Our decision can be reversed as we continue the process of seeking the correct laws of nature – a process that goes hand-in-hand with finding the correct concepts for describing nature:

> I see nothing horrendous in the notion that a language or conceptual framework brings with it a commitment to certain logically synthetic propositions, provided it is recognized that there is more than one pebble on the beach, i.e., that there are many alternative frameworks, one of which the world *persuades* us to adopt (or, better, adumbrate), only to persuade us later to abandon it for another.
>
> (PT 293)

This results in shifting meanings of our terms, but such shifts are part of the ongoing process of figuring out what the world is like. GAs and material rules are two sides of the same coin. We focus on material rules when we are discussing concepts in use; we focus on GAs when we are considering the epistemic basis of our practice.[12]

### *4.2.2 Implicit Definitions*

Another Sellarsian theme overlaps the above discussion and will clarify how he integrates his account of material rules into an holistic account of conceptual status. Sometimes Sellars treats languages as axiom systems in which terms are *implicitly defined* by the axioms in which they occur.[13] The GAs we accept for a domain provide these axioms, which include terms that are specific to that domain and establish relations between these terms. We encountered this idea in Sec. 3.5 when we discussed the view of a scientific theory as a formal calculus plus a set of correspondence rules. Sellars' account of *conceptual status* is a generalization of the "formal calculus" part of this approach, which Sellars applies to all expressions that have conceptual status, not just to theoretical terms of science.[14] By specifying relations between the characteristic terms of a language, axioms provide at least part of the meanings of these terms, and thus provide part of their definitions. They also constrain any further interpretations we may impose on these

terms. These definitions are "implicit" because they are not of the form "'C' if and only if ' . . . '", which is the form of explicit definitions. The analytic sentences that many empiricists invoke to reduce sentences of the secondary language to sentences in the observation language are explicit definitions and presumably analytic since they provide rules for eliminating secondary terms from expressions in which they occur (cf. SAP 302). Implicit definitions do not provide rules for eliminating terms, they provide only relations between terms. In other words, a major component of a conceptual system for a domain is a set of axioms that serve as GAs. The relations between the concepts in these axioms are reflected in the material rules that guide our behavior and thought in that domain.

I want to underline a central dialectic that Sellars identifies in the development of human thought. Adult thought about a subject matter always begins with a conceptual framework that includes propositions we consider axiomatic for that domain. Sometimes we entertain a new claim on empirical grounds. When we become convinced that this claim is true we accept it as an axiom, and adjust our axiom-set accordingly. In doing so we build the new claim into our language, which typically involves some alteration in the original conceptual system. The detailed alterations may range from minor (when we just add a new feature to our account of some item) to wholesale reconstruction. Once we have made this change, we (and our successors, assuming that we are dealing with a fairly stable case) can recapture the claim by reflection on the meanings of our terms – which is why we think of propositions that express meanings as a priori. But this a priori status is the result of human decisions and thus subject to revision, which is why "yesterday's necessities, are today's contingencies and vice versa . . . " (LRB 311). This process will continue until we achieve the final account of the items in the domain.

### *4.2.3 Entry Transitions*

In his early papers Sellars was strongly attracted to a coherence theory of meaning, which would make the content of any concept solely a matter of relations to other concepts. He sometimes described his theory as coherentist (RNWW 617) and maintained that the non-logical content of descriptive terms is completely captured in material rules. For example:

> The meaning of a linguistic symbol *as a linguistic symbol* is entirely constituted by the rules which regulate its use. The hook-up of a system of rule-regulated symbols with the world is not itself a rule-governed fact, but . . . a matter of certain kinds of organic event. . . . But if the linguistic as such involves no hook-up with the world, if it is – to use a suggestive analogy – a game played with symbols according to rules, then what constitutes the linguistic meaning of the factual, non-logical expressions of a language? The answer, in brief, is that the undefined

> factual terms of the language are *implicitly* defined by the conformation [i.e., material] rules of the language. These specify the proper use of the basic factual expressions of the language in terms of what might be called an axiomatics. Thus, for each basic factual word in the language there are one or more logically synthetic universal sentences which, as *exhibiting* the rules for the use of these words, have the status of "necessary truths" of the language.
>
> (LRB 310)

One attraction of a coherence approach is its ability to provide a unified account of all concepts. Nevertheless, Sellars exhibits some discomfort with a pure coherence approach at a fairly early stage. For example, he notes that for a language to be *applied*, some of its descriptive predicates must be learned responses to extra-linguistic objects (IM 334), and that a language that is not applied is, in some sense, empty (RNWW 611). Still, in these early papers, Sellars insists that questions of application and meaning are distinct: "the difference between an applied and a non-applied language has nothing to do with the *meanings* of its expressions" (RNWW 611, cf. IM 335). In later work Sellars adopts a coherence account of *conceptual status*, but concludes that conceptual status does not exhaust meaning for descriptive predicates. Thus, comparing English and German speakers, Sellars writes: "if they did not (tend to) respond to red things in standard conditions with '*rot*' – when 'looking to see what colour it has' – it could not be true that the German word '*rot*' means *red*" (SRLG 335, cf. SAP 316). Even more strongly, in the context of a rejection of "the abstractive theory of concept formation in all its disguises," Sellars insists that "one does not have the concept red until one has directly perceived something *as* red" (P 90). Eventually Sellars concludes that: "a non-logical predicate constant which isn't connected with extra-linguistic objects is not, in the full sense, meaningful" (SRII 176). Indeed, Sellars gives the key argument against a pure coherence account of descriptive concepts (SAP 304–5): A set of predicates that is implicitly defined by a system of propositions is an abstract structure which may have many "real" (i.e., extra-linguistic) meanings. More propositions increase the constraints on the set, but never generate unique concepts. Something in addition to a purely formal structure is required for descriptive meaning. However, Sellars stresses that, contrary to concept empiricism, *no concept gets its content solely as a result of association with some experienced item*. I will take the view that descriptive concepts get part of their content from some connection to extra-conceptual items to be Sellars' mature view and will focus on that part of his account in this section.

Sellars calls the connections we are concerned with *entry transitions* (ETs), and maintains that these are *non-inferential* transitions from noticing some item to subsuming it under a concept. These transitions are embodied in stimulus-response (S-R) habits: "the *observational application* of a concept cannot be the obeying of a rule at all. It is *essentially* the actualization of a

thing-word S-R connection" (SRLG 334). For example, when I look at a typical stoplight I spontaneously tend to think of it as red, and when I rub my hand over a piece of sandpaper I tend to think of it as rough, without mediation by any reflective process. Sellars holds that mastery of a descriptive concept requires the development of such habits (e.g., MFC, SAP, SLRG). But while I have not mastered RED and ROUGH until I have developed the appropriate habits, the tendency to utter or think "red" in the presence of red objects will not distinguish someone who has the concept RED from someone who is just applying a label. A digital thermometer attached to a voice synthesizer can correctly announce temperatures without acquiring temperature concepts. Mastery of a descriptive concept requires learning the appropriate implications as well as the ETs; I do not see anything *as red* until I have learned those implications.

A somewhat more complex example will underline the importance of not identifying mastery of a descriptive concept with learning ETs. In his early work on the spectrum Newton acknowledged only five spectral colors; orange and indigo were not included. He saw these colors and he named them, but considered them to be boundaries between colors, rather than distinct members of the spectrum (Topper 1990). At this stage of his research Newton's concepts of orange and indigo entered into the same ETs as his later versions of these concepts. At both stages Newton would have said "orange" and "indigo" in the same circumstances, but these terms enter into different networks of implications in his earlier and later accounts. For Sellars, Newton's move from a five-color spectrum to a seven-color spectrum involves a degree of conceptual change, even though there may have been no changes in the relevant ETs.

As the last example suggests, a single item can provide the basis for different ETs – and thus be conceptualized differently – under different circumstances. Which transition is made will depend on available concepts plus such factors as local observational conditions and the individual's beliefs and aims. Adapting an example from EPM (142–44), a person working in a tie shop who is familiar with the effects of the shop's lighting may learn to spontaneously classify a tie that looks green in the shop as blue. In daylight the same apparent color might lead this individual to classify this tie as green. But more drastic differences can also occur. The tie-shop example involves transitions to different concepts within a system of color concepts, but a single item may be involved in transitions into different conceptual systems. Thus Sellars notes that one may respond to the noise *red* as an English word, or by a singing instructor as a flat note (SRLG 329–40). Consider some other examples. I may identify an item on the table before me as a fork – an eating utensil. But given a different state of mind, I might identify the item as an instance of tarnished silver without any concern for its customary use. In the same way, a biologist glancing at a tree in different circumstances might conceptualize it in terms of its species, or as a hard wood, or as an obstacle. All of these classifications involve background

knowledge, but Sellars holds that this does not make them inferential. In general, he rejects the view that "knowledge (not belief or conviction, but knowledge) which logically presupposes knowledge of other facts *must* be inferential" (EPM 164). Indeed, given Sellars' view of the role of material rules in determining conceptual content, all classification involves some background knowledge: "knowledge 'at the perceptual level' essentially involves *both* knowledge of singular matters of fact and knowledge of general truths" (SK 297).

In addition, different items can lead to the same concept. We have already seen this in the tie-shop example, where different color experiences lead to BLUE in the shop and in sunlight. This example throws light on Molyneux's problem (Sec. 3.1) and, in general, on situations in which we recognize an item by more than one sense. In Molyneux's case different sensations serve as the experiential side of an ET to a shape concept. The blind patient who first learns to identify shapes by touch and then acquires sight must learn a visual ET for each shape. A parallel with Carnap's reduction sentences is also worth noting. Both Carnap and Sellars hold that some concepts are open ended (recall Sellars' remark about the prevalence of such concepts quoted in Sec. 4.1). For Carnap, we can extend a theoretical concept by adding new reduction sentences that introduce new ways of detecting instances of that concept. For Sellars, acquiring new ways of identifying *any item* may involve learning new ETs, and adding ETs to a concept involves a degree of conceptual change. Sellars suggests that this change is better described as enrichment of the concept than as replacement (CDCM 287).[15] I will return to this situation as we proceed; for now I want to emphasize that, for Sellars, qualitatively different sensory experiences do not entail distinct concepts.

Although Sellars usually illustrates ETs by simple perceptual examples, he recognizes other kinds of cases. For example, he tells us that once we have reached the stage at which we can talk about language, "Language entry transitions now include 'This is a "2 + 2 = 4"' as well as 'This is a table'" (MFC 425). I think it is also appropriate to allow for ETs that involve more than one sense. For example, I might non-inferentially recognize a problem with a car's exhaust system as a result of what I simultaneously hear and smell.

Consider another issue. On standard empiricist theories a primary predicate can acquire meaning only by being associated with an instance; as a result, there cannot be any primary predicates that lack instances. On Sellars' approach there are no primary predicates, and requiring ETs as part of the content of all descriptive concepts does not eliminate fully meaningful predicates without instances. In Aristotle's system of physical concepts, for example, falling stones and rising flames provide the basis for ETs to the concept of an object moving to its natural place. In later physics the concept of natural place is eliminated, and these phenomena are no longer subsumed under a single concept. But we can still describe the Aristotelian concept and the ETs that are appropriate to that concept. The

same point holds for phlogiston, prepotency, metabolons, and other abandoned concepts. Let me emphasize again the importance of distinguishing between giving an account of the content of a descriptive concept and asking whether that concept has instances.

I now want to consider in some detail Sellars' thesis that ETs are S-R connections, not inferences. Sellars holds that all inferences are rule-governed. Even in the case of habitual inferences, on reflection we can recover the formal or material rules that justify them. Thus when Sellars maintains that ETs are non-inferential, his claim is that *even on reflection*, we cannot find rules that justify these moves. This is because inferences are possible only after we are in a conceptual system, while ETs are moves from the world into a conceptual system. Sellars justification for this claim rests on an argument that he repeats in many places (e.g., CC 82–89; EPM 163–64; MFC 430–32; SAP 314–16). The argument is mainly aimed at Carnap who, in his account of formal languages, distinguishes syntactical rules that connect items within a language from *semantical rules* that relate the system to an extra-linguistic subject matter. Semantical rules give the meanings of basic predicates by relating words in the basic vocabulary to the extra-linguistic items that determine their meaning; "'red' means red" is a paradigm example. Sellars attempts to show that there is something fundamentally wrong with the notion of such a rule, and that the root of the problem lies in the mistaken belief that "means" expresses a relation.[16]

We can begin our discussion of Sellars' argument by noting that the general form of a semantical rule is:

" . . . ." means _. (M0)

To explore this form Sellars introduces examples such as:

"red" means red, (M1)

"red" means *rot*, (M2)

*"und"* means and. (M3)

Consider M1 first and note that there are no quotation marks around the second occurrence of "red." For Carnap, quotation marks would be inappropriate in this location since "red" is being *used* to name a perceptible quality. In its first occurrence "red" is in quotation marks because it is *mentioned* – it is the word we are talking about. M3 looks like another instance of M0, but something seems wrong because "and" does not mention an item in the world.[17] Sellars holds that M3 is appropriate when we explain the meaning of the German "*und*" to someone who already speaks English, and our goal is to inform the English speaker that "*und*" plays the same role in German that "and" plays in English. It might seem that "and" should be in quotation

marks because we are mentioning two words and asserting that they have the same meaning in their respective languages, but Sellars disagrees. He holds that M3 is correct as written because "and" is in fact used, not mentioned, in this sentence, although it is used in a special way: "and" is being *displayed* to the language learner, who already understands this term. The English speaker is being invited to reflect on the role that "and" plays in English and recognize that "*und"* plays that role in German. A different example may clarify Sellars' point. Suppose I am playing chess with an unusually shaped set of pieces. Since I am playing in a public place with many kibitzers, I keep a box containing standard chess pieces nearby. When someone points to a piece on the board and asks "What's that?" I respond by reaching into this box and displaying a familiar piece – say, a knight. I thereby inform the kibitzer that the unfamiliar piece plays the role of a knight in the chess set I am using.

Let us pursue the chess example a bit further. Being a knight is completely determined by the role that knights play in the game; the size, shape, or color of the piece is irrelevant. Sellars underlines this point by introducing a version of chess played in Texas in which the squares are counties, the king is a Cadillac, pawns are Fords, and so forth (SLRG 344). It is not even necessary that the counties be laid out in a rectangular array. All that is required for this game to be chess is that there be an appropriate mapping between the moves and pieces in this game and those in the more familiar version. In other words, to be a knight or a specific square on a chessboard is to play a particular role – which amounts to functioning in accordance with a set of rules. Two apparently different structures are, in this respect, identical when there is an isomorphism between the pieces and rules in one structure and those in the other – although the specific mapping may be quite complex.

The meaning of a word, Sellars holds, is determined by its role in a language; words in different languages have the same meaning when they play the same role. I can teach the meaning of a word in language L to someone who is not familiar with L by displaying a word in a familiar language that plays the same role. This, Sellars contends, is the proper interpretation of expressions of type M0:

> Meaning statements, by their very nature, focus attention on the functional equivalence of expressions. They do not tell us *how* an expression functions, except *indirectly*, by presenting us with another expression, with the functioning of which we are presumably familiar and giving us the task of 'getting with' this functioning by a rehearsal in imagination of the patterns of inferential and non-inferential transitions characteristic of the latter expression.
>
> (NO 113)

Thus M3 states that "*und*" plays the same role in German that "and" plays in English by displaying an "and." The same account applies to M2 except

that in this case the meaning of the English "red" is being explained to a German speaker.

Sellars introduces *dot quotes* to capture this idea. The expression "• and •" indicates the role that the word "and" plays in language. Thus M3 can be rewritten:

"*und*" is an • and •. (M3′)

In this notation • *und* •s are identical to • and •s so that *M3′* is equivalent to:

"and" is an • *und* •.

Which expression we use depends on which term we are explaining. In the same way, M2 becomes:

"red" is a • *rot* •,

and so forth.[18]

In the stronger version of this account Sellars holds that explaining the meaning of an unfamiliar term by displaying a familiar term is the only proper use of format M0. "Means," Sellars tells us, is not a relation term except in "a purely grammatical sense" (SAP 315). Rather, "'means' is a special form of the copula" (MFC 432, cf. SM 81). If we interpret M1 in this way, we should reject the view that M1 is a semantical rule that specifies the relation between a word and a non-linguistic item. However, to establish the conclusion we need a reason for accepting this interpretation of M1. So far, we have only a possible interpretation.

Sellars' main reason for rejecting the usual interpretation of M1 will take us, briefly, into the realm of language learning. Price's strictures notwithstanding, the usual empiricist account of meaning is largely built on an account of language learning and has been regularly defended on this basis. But, Sellars argues, expressions such as M1 cannot play any role in learning the meaning of "red" because understanding the rule requires that the learner already knows this meaning (IM 335–36; SAP 312; SRLG 333–34).[19] Thus, Sellars concludes, the only rules involved in learning the meaning of a word are *syntactical* rules: "the conceptual status of descriptive as well as logical – not to mention prescriptive – predicates is constituted, *completely* constituted, by syntactical rules" (SAP 316). Several comments on this argument are in order.

First, it should be clear that the scope of "syntactical rule" in Sellars' usage goes beyond its customary use in logic and logical theory since Sellars includes material rules among the syntactical rules. This usage is, however, in conformity with Carnap's usage which includes both L-rules and P-rules among the syntactical rules (1959: 315–16).

Second, while Sellars' point about the irrelevance of these semantical rules for language learning can be accepted without qualms, attribution of the view that we learn our primary language via such rules to empiricists is unfair.

Their account of primary-language learning is built on the notion of ostensive definition. The semantical rules that Sellars is criticizing occur in sophisticated accounts of formal systems. Such accounts are developed by and for people who already know the language in question; the function of these rules is to express a thesis about meaning, not to teach meanings of words. The thesis may be false, but it is not unintelligible.

Third, Sellars holds that statements such as M1, on their usual interpretation, lead to bad metaphysics. If we take M1 as giving the meaning of a word by relating it to an entity, then we are tempted to treat sentences such as "'Triangular' means (stands for, designates) triangularity" as also expressing a relation between a word and an entity, and we are on the path to Platonism (e.g., EAE, NO). However, we can avoid this temptation while also denying that all sentences of type M0 have the same sense.

Fourth, Sellars' account provides a uniform interpretation for the variety of different contexts in which "means" occurs, and he considers this to be a desirable result. On an empiricist view of our basic descriptive vocabulary we need a different account for the meaning of "many of our most familiar concepts, among others those of logic and mathematics." This results in a "radical dualism" since we now require a "second mode of concept formation, namely the learning to use symbols in accordance with rules of logical syntax" (SAP 312). Sellars, on the other hand, seeks a unified theory of meaning that applies even to the three classes of expressions that have been problematic for empiricists exactly because they do not get their meaning from associations with extra-linguistic items: logical constants, theoretical terms of science, and prescriptive terms.

We encountered the thesis that empiricists require different accounts of concept formation for empirical terms and logical constants in Sec. 3.4; I submit that they can accept this result with equanimity. There is nothing wrong, they could reply, with there being different modes of concept formation for different kinds of concepts. Moreover, while Sellars provides a uniform account of *conceptual status*, he recognizes that conceptual status is not the entire story with respect to conceptual content. He admits different kinds of concepts and gives partially different accounts of the content of these different kinds. Recall that we are currently in the midst of a discussion of ETs, an additional element besides conceptual status that, Sellars holds, is required for descriptive concepts. We will shortly find that Sellars' account of logical constants has much in common with empiricist accounts. Sellars' key thesis with regard to descriptive concepts is that none get their content *solely* from correlations with extra-linguistic items. But the defense of this thesis does not require denying that M1 expresses a relation. The upshot of this discussion, then, is that Sellars has not made a case for a single uniform interpretation of all versions of M0, or shown that "means" never expresses a relation, or demonstrated a defect in Carnap's use of semantical rules.

At this point two Sellarsian theses seem clear and defensible: part of the content of any descriptive concept is specified by a connection to its extra-conceptual subject; and these connections alone do not determine the content of any descriptive concepts – some relations to other concepts are also required. Having criticized Sellars' reasons for rejecting the view that the required relation to extra-conceptual items is established by semantical rules, I will now argue that Sellars' doctrine of ETs does not provide the account we need.

Focus first on the role of habits in establishing this connection. While the habitual application of concepts may have pragmatic virtues when we use them to find our way around some domain, habits become irrelevant when we shift to reflection on our concepts. Recall Sellars' description (quoted above) of how we can first decide to adopt a material move and then undertake to make the move habitual. A parallel story applies with regard to ETs: I might encounter an unfamiliar item, decide that it is an instance of a familiar concept, and adopt a new ET. Molyneux's patient is a striking example, and the history of science is full of cases in which new ways of identifying acids, or specific elements, or radioactivity were discovered. In each case we may decide to adopt a direct link from the item to the concept, and we may *then* undertake to make the link habitual. We have a reflective understanding of the concept *before* we establish the habit – and this reflective understanding is required if we are to have reasons for undertaking to establish the habit. Whenever we consider a concept reflectively – perhaps we are contemplating a conceptual change that will involve adding or deleting ETs or implications, proposing a new concept, analyzing a familiar concept, or studying concepts of another society or historical period – mastery of conceptual content does not require that we achieve *habitual* use of any aspect of that concept. Indeed, reflection on the content of a concept may convince us that we do not want to form such habits. An immediate consequence is that a theory of conceptual content should not include any reference to habits. Sellars appears to have run together two different issues that we should keep distinct: the content of descriptive concepts and the conditions for efficient use of established concepts.

A second problem with Sellars' doctrine of ETs arises when we turn to theoretical concepts. These provide a crucial test for any theory of concepts and are especially important for Sellars who is both a scientific realist and a fallibilist: He holds that discovery of the correct concepts for describing various domains is a long-term goal of science, and that pursuit of this goal often requires revision of earlier views. We have learned from the history of science that a substantial part of this pursuit involves discovery of items that are not available to our unaided senses. Sellars holds that this is just what we should expect: "it would be odd if the only qualitative dimensions of the world were those which are, in the last analysis, tied to the sensory centers of the human brain" (TE 70, cf. 77–78). The discovery of such items results in the introduction of new descriptive concepts – the *theoretical concepts* that have troubled empiricists.

Sellars provides two different accounts of theoretical concepts, one that applies to developing science, and one that applies to the final science we may achieve some day. One might think we should restrict discussion to science as we find it – that is, to developing science – but Sellars holds that philosophers should be more adventurous:

> the perspective of the philosopher cannot be limited to that which is methodologically wise for developing science. He must also attempt to envisage the world as pictured from that point of view – one hesitates to call it Completed Science – which is the regulative ideal of the scientific enterprise.
>
> (TE 77)

The latter-day-logical-empiricist account of theoretical concepts in terms of axiom systems plus correspondence rules plays a key role in both accounts of theoretical concepts. In addition, models and analogies play a central role in Sellars' account of developing science, but I will postpone discussion of that aspect until Sec. 4.5.

The key difference between developing science and final science concerns the role of the observation/theory dichotomy. Final science will be expressed in terms of concepts that describe the items that actually exist in each domain, and (Sellars holds) at that point we should adopt these concepts for our descriptions. Even our everyday observation concepts should be replaced so that we will then respond to experience directly in terms of these new concepts. In other words, ETs will take us directly from experience to these fundamental concepts, and the content of each concept will be jointly determined by its ETs plus its implicit definition in terms of other theoretical concepts. When this occurs, the distinction between observation concepts and theoretical concepts will no longer have any semantic import. But, Sellars maintains, we should not attempt to make these replacements until that final stage is reached (e.g., SM 146).[20] In developing science the observation/theory dichotomy is methodologically central and we have a *three-layer structure*: experience, ETs that take us from experience to observation concepts, and correspondence rules that connect the axiom system to observation concepts.

> Those descriptive predicates which are conditioned responses to situations of the kind they are correctly said to mean are called *observation predicates*. If a language did not contain observation predicates it would not be *applied*. Descriptive predicates other than observation predicates gain application through rules tying them to observation predicates. . . . One can, indeed, say that all the other descriptive predicates of a language must be "defined" in terms of observation predicates; but it would be a mistake to suppose that in every case these definitions will be *explicit* definitions.
>
> (SAP 316)

Moreover, "There is a core of truth in the concept of '*the* observation framework' and, indeed, of the abstractionist approach to basic empirical concepts which survives the exorcizing of givenness" (SRII 187). This core of truth is that in developing science theoretical meaning depends on observables: a theoretical concept "must belong to a framework which is logically connected with the language of observable fact . . . " (EPM 193). In our current framework observation concepts constitute a rock-bottom epistemic stratum, although "it is *still* in principle replaceable by another conceptual framework in which these predicates do not, *strictly speaking*, occur. It is in this sense, and in this sense *only* that I have rejected the dogma of givenness with respect to *observation* predicates" (SRII 187).

Note that Sellars provides a clear characterization of observation predicates: "Those descriptive predicates which are conditioned responses to situations of the kind they are correctly said to mean are called *observation predicates*." Predicates such as red, rough, table, and tree meet this requirements; an ET takes us directly from noticing such an item to the corresponding concept. On the other hand, while I may use a magnetic compass to detect the local direction of the earth's magnetic field, or a Geiger counter to detect the presence of radioactivity in my immediate environment, MAGNETIC FIELD and RADIOACTIVITY are not observation concepts. We do not see magnetic fields or hear radioactive decays; rather, we see a needle on a compass or hear a series of clicks. These are the observables that we connect to MAGNETIC FIELD or RADIOACTIVITY via correspondence rules. Sellars is quite clear that unaided observation is basic. Thus he acknowledges that it is:

> not absurd to speak of observing viruses and protein molecules through an appropriately constructed electron microscope. But, as is evident, this extended use of the term is built on the physical theory of the instrument and how it relates to the physical systems which can be observed by its use. Again, to identify the objects observed by its use as 'protein molecules' or 'viruses' presupposes biochemical theory and pathology. Furthermore, it is particularly clear that the observations made by the use of an instrument cannot be the grounds on which we accept the theory of the instrument. For until we have the theory of the instrument, we *logically* can't make observations with it. . . . Thus, although observation in the extended sense provides data for the elaboration of theories pertaining to objects which are not observable in the absence of theory-laden instrumentation, the concept of such observation presupposes the concept of unaided perception. . . .
>
> (TE 61–62)

Even if we train ourselves to respond to the compass using MAGNETIC FIELD, MAGNETIC FIELD is not an observation concept since we cannot detect these fields without the instrument. Our justification for considering the direction

of the needle to be the direction of the earth's magnetic field depends on the theory of the instrument which, Sellars holds, ultimately depends on what we can perceive without instruments. Note how this case differs from the tie-shop example where the trained employee moves directly from the green appearance of the tie to describing it as blue. Blue is an observation predicate, which is why we can check the color by taking the tie into sunlight. We cannot check the compass by a comparable shift because we are biologically barred from perceiving magnetic fields.[21]

Now consider the perspective of final science. At this stage we adopt an ET that takes us directly from what we see to DIRECTION OF THE LOCAL MAGNETIC FIELD; an analysis of this concept would not include any mention of the intervening observable. The Geiger-counter case is similar. At the present stage of physics we may habitually move from hearing a particular sound to RADIOACTIVITY, but an analysis of this concept would include an ET from what we hear to the concept CLICK, and then (via correspondence rules and implicit definitions) to RADIOACTIVITY. Including reference to the observable output of a specific detector as part of the analysis of a theoretical concept may seem odd, but this is a consequence of the requirement that all concepts be tied to observables. For Sellars this oddity holds only for developing science. In the final stage the intermediate ET to the concept describing something we can sense will vanish. An account of RADIOACTIVITY will include just a description of its relations to other concepts plus a set of ETs that takes us directly from extra-conceptual items to RADIOACTIVITY.

Unfortunately, Sellars' account of the role of ETs in determining conceptual content clashes with important cases in contemporary physics whether we consider these cases from the perspective of developing or final science. The problem arises in a particularly sharp form because of the statistical nature of much contemporary physics; it is prominent in high-energy physics where cases occur in which specific observables provide characteristic evidence for the presence of two or more different particles or processes. For example, the experiments that established the existence of top quarks (Abachi *et al.*, 1995; Abe *et al.* 1995) provided a number of instances of a data-pattern that could have resulted from the occurrence of a top quark – among other possibilities.[22] Statistical analysis shows it to be highly improbable that none of these patterns resulted from the passage of top quarks through the detector; so one can conclude that top quarks exist and behave as predicted. But *no specific case involving this particle is ever identified*.[23] Whether we are considering developing or final science, anyone who spontaneously identified a pattern as the signature of a top quark *would be making a mistake*.[24] Note especially that physicists had a fully developed concept of a top quark before they did any of the experiments or statistical analyses that confirmed the existence of this particle. Indeed, the concept was required to design the experiments and determine what data-patterns are relevant.

Detection of elusive particles and processes are not the only cases in which ETs are inappropriate; other examples include SPACETIME INTERVAL,

ENERGY, and ENTROPY. Given our sensory biology, we will not be able to apply these concepts via S-R connections even if we achieve final science. Nor are there any reasons for believing that the future development of physics will move us in the direction of concepts that are applicable by ETs. Rather, the development of physics has generally moved in the direction of more abstract concepts whose instances cannot be picked out at a glance, even if we are well equipped with the appropriate conceptual framework (see Ch. 10). If Sellars' general approach to descriptive concepts is correct, the tie between a formal structure and its domain will have to be introduced by some other means. In Ch. 5 I will propose a modified Sellarsian account of descriptive concepts without ETs. The view that descriptive concepts derive part of their content from a relation to appropriate extra-systemic items will be central to that account, but the account will not require any encounter with actual cases in which the concept is instantiated. Nor will it include any habits or non-inferential moves as part of the content of descriptive concepts, or as a requirement for the mastery of such concepts.

### *4.2.4 Individual Concepts*

Our discussion of descriptive concepts has focused on predicates, but Sellars' view also includes concepts we use for thinking about specific individuals: "we must recognize individual concepts as well as universal concepts (and, indeed, other kinds of concepts as well) . . . " (LT 112). Historically, there has been a debate between the view that individuals are labeled but there is no conceptual content to these labels, and views which hold that there are full-blown individual concepts. The Frege-Russell view, which treats referring terms as descriptions, is an important instance of the latter view. At present, under the influence of Kripke, the label view is widely held.

Sellars' remarks on the topic are generally critical of the Frege-Russell view, which was dominant when Sellars was writing on these issues. His own view characteristically seeks to synthesize elements from both approaches. "Names of objects have a function which, like that of the origin of a coordinate system, is to be a *fixed center of reference*, a peg, so to speak, on which to hang definite descriptions" (NO 123). Note how the distinction between a label and a concept applies in this case. On encountering a new individual, I may initially give it a label. As I learn about this individual I develop a concept that contains my beliefs about it. This is particularly clear in cases of persons I know, well-known historical figures, and contemporary figures in the public eye. I have beliefs about Mozart, Kant, and Einstein that I use for assessing whether claims about some individual actually refer to these figures. Moreover, these beliefs are subject to revision, with the usual impact on my concepts. Many of these beliefs will be expressed as definite descriptions, but they need not all be so expressed. I can believe that Mozart was a great composer without having to believe that he was *the greatest*, and my belief can still serve to fix reference: If I overhear a conversation about

someone named Mozart, but realize that the person being discussed is a stock broker, I will not make the inferences I normally make about the composer. The same account applies to non-human individuals such as the planet Pluto or Mount Ranier. For Sellars, the special feature of individual concepts is reflected in the ETs. Many different items may serve as the basis for such an ET, but since the concept applies to just one individual, rather than to any member of a class, any ET must take off from something that is appropriately connected to that individual. Presumably this will require a causal chain that involves that individual.

## 4.3 Formal Concepts

The content of formal concepts is completely determined by implicational relations. Sellars considers these full-blown concepts, although concepts of a different kind than descriptive concepts; mathematical and logical concepts are the prime examples. Consider the mathematical concept GROUP, defined as a set of elements plus a binary operation that has just four properties.

1. The set is closed under the operation: using $e$, and $f$, to represent two elements, and $\circ$ for the operation, $e \circ f$ is a member of the set.
2. The operation is associative: for any three elements, $(e \circ f) \circ g = e \circ (f \circ g)$.
3. There is an identity element, $I$: for any element, $I \circ e = e \circ I = e$.
4. Every element has an inverse: elements, $e$ and $f$ are inverses of each other just in case $f \circ e = e \circ f = I$.

These are formal properties in that we need not have any idea what $e$, $f$, $g$, and $\circ$ stand for; the relevant properties are completely determined by the axioms. A mathematician can study the properties of groups without considering any specific items or operations that fit this definition. One advantage of such formal exploration is that its results apply to any subject matter that has the group structure. The set of all integers – positive, negative, and zero – is a group if the operation is addition and zero is the identity element. It is not a group if the operation is taken to be multiplication, since most elements lack inverses. The set of positive and negative rational numbers – omitting zero – is a group with multiplication as the operation and one as the identity element. I consider some applications of groups to physical science in Ch. 10.

Logic provides a second standard example of formal concepts. However, we must be careful exactly because modern logic is mathematical logic. Standard truth-functional propositional logic consists of a specific formal system – Boolean algebra – plus an interpretation. Letters of the alphabet ($p$, $q$, etc.) are variables that stand for truth-values of propositions (and, to this extent, for the propositions themselves); the binary connectives "&" and "$\vee$" are functions that map pairs of truth-values to a single truth-value; the monadic operator "~" is a function that maps a single truth-value to a truth-value. The functions are chosen so that they mimic familiar linguistic

operations commonly expressed by "and," "or," and "not." The result is an interpreted system that is still "formal" in the sense that the sentence letters do not stand for specific truth-values; they are variables that range over the permitted truth-values. The point that propositional logic is an interpreted system can be seen most clearly by noting that the underlying formal system can be give quite different interpretations. For example, we could let the variables range over real numbers in the closed interval from zero to one, and the three functions represent the smaller of a pair of numbers (&), the larger of a pair ($\vee$), and one minus a number (~). (When the inputs to the binary functions are equal, the output is just that number.) All the theorems of propositional logic hold under this interpretation – and other interpretations as well. We generate the familiar logical formalism by adopting a set of ETs that take us from a specific natural language word, such as "and," to the formal concept CONJUNCTION, and so forth. Ambiguities of natural languages sometimes prevent these transitions from being automatic (another case in which habits should be avoided or restrained). For example, the United States Constitution forbids "cruel and unusual punishment," but (without further context) this could express either a ban on cruel punishments and unusual punishments or only a ban on punishments that are both cruel and unusual.

This way of looking at mathematical logic provides some insight into what is involved when people advocate alternative logics, such as three-valued, intuitionist, or relevance logic. The underlying disagreement is over which arguments are valid. Intuitionist logic does not include a connective that has exactly the same properties as "~" in classical systems because the claim that ~~$p$ implies $p$ is rejected. There is, instead, a connective that has many, but not all, of the properties of standard negation, which is why we may want to call it an alternative form of negation, or a variant negation concept (CC). In a three-valued logic the propositional variables range over three permissible values, and the functions associated with the operators must be adjusted. Some alternative propositional logics, such as relevance logic, require attention to properties other than truth-values. Given the existence of alternative logics, we can conceive of situations in which we develop a subject using different logics, keeping the central claims of the subject unchanged, but accepting different consequences. In an analogous way, we can consider the Galilean and Lorentz transformations as different formalisms for transforming expressions between frames of reference that are in uniform motion with respect to each other. Applying the different transformation rules to Newton's laws or Maxwell's equations yields different results as to which expressions retain their form under a transformation. Special relativity also replaces the formalism used in classical physics for combining velocities. Whatever reasons we have for believing that relativity is correct are also reasons for accepting this formalism. Whether we are working within a formalism, or applying that formalism to some descriptive system, formalisms provide powerful means of arriving at new results. Such results can lead us to change our views about the adequacy of a system of descriptive concepts, or the formalism we are using, or both. In subjects that

can be treated mathematically, this interplay is one generator of conceptual change. These cases also raise the question of what constitutes a single conceptual system; I defer this question until Sec. 5.8.

The content of a formal concept is determined by the full set of implications in which it plays a substantive role. The notion of a *substantive role* can be illustrated by comparing the implication in standard propositional logic from $p$ to $p \vee q$ with the implication from $p$ to $p \vee (q\&r)$. The first implication is licensed by the concept DISJUNCTION and expresses part of the content of that concept. The second implication is a special case of the first that does not depend in any way on the properties of CONJUNCTION. Since no properties of conjunction are relevant to this implication, conjunction does not play a substantive role and the implication tells us nothing about conjunction.

Considering some other examples can bring out the significance of attending to the full set of implications. Substantive implications involving conjunction include those from $p$ and $q$ to their conjunction $p$ & $q$, as well as those from $p$ & $q$ to $p$ and to $q$. These are sometimes referred to as "conjunction introduction" and "conjunction elimination" but it would be a mistake to hold that these implications *fully* determine the concept of conjunction. Rather, part of this concept is expressed in De Morgan's laws (ME 330) that relate conjunction, disjunction and negation. Other principles of logic, such as the distribution rules, express other aspects of the logical concepts that occur in them in a substantive way. It is possible to build an alternative logic (such as certain quantum logics) that include the introduction and elimination rules for conjunction, but not the law that allows distribution of conjunction over disjunction. Thus the claim that the full set of substantive implications is included in the content of each connective allows for cases in which multiple axioms (or rules in a natural deduction system) enter into the implicit definition of an operator.

I have been treating logical concepts as a species of descriptive concepts; this may be surprising since logic is often considered a prescriptive enterprise. I will return to this topic in the next section were we will see that some systems of concepts have both descriptive and prescriptive roles.

For Sellars, every conceptual system has a formal system at its core. This formal system embodies the implications that constitute conceptual status. A system of descriptive concepts is a more-or-less complex system of formal concepts plus a set of ETs that ties the formal system to a specific subject matter. In addition, material rules provide the justification for some of the implications embodied in this formal system.

## 4.4 Prescriptive Concepts I

Prescriptive concepts are also, for Sellars, genuine concepts, and thus partly constituted by implications in which they play a substantive role. I will focus initially on OUGHT, which is:

> the central term in the 'language of norms', a mode of discourse which presupposes, but is irreducible to, the 'language of fact'. The term 'ought' has a characteristic syntax by which it is related to other normative expressions, as well as to logical and descriptive categories.
>
> (OM 516)

Elsewhere Sellars writes:

> 'ought' has as distinguished a role in discourse as descriptive and logical terms, in particular . . . we *reason* rather than 'reason' concerning *ought*, and once the tautology 'The world is described by descriptive concepts' is freed from the idea that the business of all non-logical concepts is to describe, the way is clear to an *ungrudging* recognition that many expressions which empiricists have relegated to second-class citizenship in discourse, are not *inferior*, just *different*.
>
> (CDCM 282)

Sellars underlines this point in SE:

> Concept empiricists were dominated by the ostensive training aspect of learning how to use words: the formation of habits of responding to *things* with *words*. But it is obvious that we learn the use of many words where such a correlation does not even make sense. This is surely the case with *logical* words and reflection shows it to be equally true of such words as "was," "will be," "this," and, to move closer to practical discourse, such words as "shall," as in "I shall do A."[25]
>
> (SE 407–8)

To sustain the claim that OUGHT is a concept Sellars must show that it has conceptual status: "The criterion I propose is that a word stands for a concept when there are good arguments in which it is essentially involved" (SE 408). This can be illustrated by adapting an example from Solomon (1977: 156). Consider the argument:

> Jones ought to do A and B.
>
> Thus, Jones ought to do A. (O1)

This is a valid argument, but its validity does not follow from the logical properties of "and" alone. To see why, compare the following argument that is clearly invalid:

> The weight of the building is supported by columns A and B.
>
> Thus, the weight of the building is supported by column A. (O2)

O2 reminds us that in non-extensional contexts a proposition of the form *C(A and B)* does not entail *C(A)*. Since the premise of O1 is non-extensional, the fact that the argument is valid depends, in part, on the logical features of OUGHT.

Given that OUGHT has conceptual status, we must pin down the additional feature that distinguishes prescriptive from formal concepts. Sellars' complete ethical theory and theory of practical reason are extremely intricate, but we need not explore their details for present purposes. We can draw out the features that concern us by noting that (as in other cases) Sellars endeavors to synthesize elements from extant theories, no one of which is adequate by itself. In the present case Sellars draws on features of deontological, emotivist, and teleological approaches, although only the first two need be considered here.[26] Emotivism, Sellars tells us, has the virtue of insisting on a necessary tie between "thinking that one ought to do A" and "being motivated to do A." Emotivists erred in holding that ought-statements lack content and are thus pseudo-concepts. This error was avoided by intuitionists "of the deontological variety" who recognized that prescriptive discourse is genuine discourse, although of a unique kind. Unfortunately, many intuitionists lost sight of the motivational side of prescriptive discourse. The correct ethical theory, Sellars maintains, will include both the inferential role of prescriptive concepts and the need for a connection between moral thinking and doing (IILO 160–62). This connection is the additional feature we are seeking; it is captured in Sellars' notion of a *departure transition* (DT) plus the claim that these transitions are constitutive of prescriptive concepts.

DTs are moves from a conceptual system to an extra-conceptual domain (e.g., SLRG 329; TA 108–9). Suppose that having decided to sit on a chair, I actually sit; sitting is one example of a DT. DTs occur in many contexts where they are optional, but Sellars holds that a habitual tendency to make a DT is *constitutive of prescriptive concepts*, and this parallels the sense in which habitual ETs are constitutive of descriptive concepts. Both descriptive and prescriptive concepts require a connection to the world, but the connections are in opposite directions for the two cases. If a sequence of intra-systemic inferences leads to the conclusion that I ought to do A, I should feel an inclination to do A. If I do not feel this inclination, I do not fully grasp the concept OUGHT – just as I do not fully grasp the concept RED if I have no tendency to conceptualize red items as RED. Paralleling the case of descriptive concepts, neither the ability to infer that I ought to do something nor a tendency to carry out an action is *by itself* sufficient for a full grasp of OUGHT. Both are required (e.g., SLRG 350–52; TA 108–9).[27] However, Sellars does not require an actual DT as a condition for mastery of OUGHT, only a tendency to act; I may sometimes conclude that I ought to do A but override my inclination to act. Cases in which I do not do what I ought to do may involve a moral lapse, but it does not follow that I fail to *understand* what is involved in the prescription. Recall that descriptive

concepts also require only a tendency to make an ET: for "*rot*" to have the same meaning as "red," German speakers must "(tend to) respond to red things in standard conditions with '*rot*' – when 'looking to see what colour it has'"(SRLG 335).

Sellars also recognizes different kinds of oughts. In the first place, Sellars distinguishes rules which tells us what one "ought to do" from those which assert what one "ought to be" (e.g., SM 75–77; LTC 506–9). The former require some action; the latter are "rules of criticism." When we say that Jones ought to be grateful for a benefit received, we are describing a required state, not an act that Jones can carry out now. Still, to accept an ought-to-be implies that one ought to do whatever is necessary to assure being in the appropriate state when a relevant situation arises (e.g., SM 76). Thus ought-to-do rules are fundamental and it is these that are directly tied to DTs. Another important case concerns prohibitions – prescriptions that tell us not to carry out certain acts. We can assimilate such cases to Sellars' model by treating *ought-not* as involving a tendency to block specific DTs should an urge to carry them out occur.

A more complicated case is illustrated by principles of logic since these are rules of *permission*: they formulate inferences that we may make in a deductive context, although we are never *required* to make a specific inference from this set (e.g., IM 328–31).[28] The normative bite of logic comes from two features. First, we are prohibited from making inferences not licensed by the accepted deductive system. Our conformity to logic is thus exhibited in "negative uniformities" (MFC 422), for example, tendencies not to utter statements such as "it is raining and it is not raining." These prohibitions carry the main normative force of logic. Second, a system of logic provides a menu of specific options among which we may choose. In this respect logical constraints are analogous to the constraints provided by the rules of a game such as chess, which allow a limited array of choices at a given juncture. Still, the game analogy should not be pushed too hard since logic is presumably mandatory while playing chess is not mandatory. There are also cases that differ from logic and games in involving a prohibition that is not accompanied by any suggestions as to what acts are appropriate. In all these cases, ought-to-do remains the fundamental notion: "consciousness of ought to do is the basic consciousness involved in recognizing a set of rules, . . . consciousness of *may do* is to be defined in terms of it" (IM 332). The claim that I may do $A$ can be introduced as the negation of the claim that I ought not do $A$. For the remainder of this discussion I will use "ought" as a synonym for "ought to do" unless a different reading is explicitly noted. The upshot of the discussion thus far is that the basic prescriptive claim is of the form "Ought $A$," which requires a tendency to carry out the DT indicated by $A$.

ETs and DTs are not mutually exclusive, so a single concept could involve both. Such a concept would have both descriptive and prescriptive aspects. Sellars does not discuss such cases in detail, but he is aware of them (e.g.,

NO 131). In fact, such concepts are common. STOPLIGHT, for example, both describes a kind of object and prescribes a certain behavior. On a Sellarsian account the descriptive side of the concept includes a tendency to conceptualize certain items as stoplights, while the prescriptive side requires a tendency to stop in the presence of these items. DEDUCTIVE VALIDITY is also both normative – as discussed above – and descriptive since it describes a formal pattern as truth preserving. Consider two more examples: PROMISE describes a kind of act and expresses a commitment; TRUTH describes a property of propositions, but is also prescriptive since the claim that a proposition is true includes the requirement that it ought to be believed. Someone who agrees that a proposition is true but has no inclination to believe it does not have a full grasp of TRUTH. An adequate analysis of truth would thus require an account of its implicational relations to other concepts, as well as its entry and departure transitions. I will undertake such an account Sec. 8.4.

Unfortunately, Sellars' DTs are habits and including them in the content of concepts generates problems parallel to those we encountered in the case of ETs. We can study a culture and learn its prescriptions without integrating those prescriptions into our own behavior. The requirement to act in a particular way may be built into a concept, but *understanding* the concept does not require developing a tendency to meet this requirement. Sometimes we *first* learn the prescriptions of another culture and *then* decide that we do not want to integrate the prescribed action into our own behavior (e.g., propitiating various deities, or finding and executing witches, or continuing a vendetta). This point also holds when we consider introducing a new prescriptive concept: We want to be clear on the prescribed behavior *before* we decide whether to include it in our active lives.

An additional problem arises because of Sellars' attempt to reduce all prescriptions to ought-to-dos. The proposed reductions are not all compelling. To be sure, we can define "ought not" and "may" in terms of "ought" and "negation", but such definitions provide no insight into cases such as deductive logic where we have a set of permissible options, but no clear guidance as to which option we ought to take at a particular juncture. More generally, permissions and prohibitions have considerably more structure than is encompassed by these definitions. I suggest that at least part of the motivation for the proposed reductions lies in the attempt to capture the relation between a prescriptive concept and its domain in a single psychological phenomenon. I will return to this topic in Sec. 5.5.

Many prescriptive concepts share a feature with formal concepts: When the concept is applied to a proposition, the content of that proposition is not changed, while the proposition determines the required behavior. "Ought A," for example, requires a DT, but the specific DT depends on A. Such prescriptive concepts may be described as *quasi-formal*. However, not all prescriptive concepts have this feature (e.g., STOPLIGHT).

## 4.5 Models, Analogies, and Conceptual Change I

It is a striking feature of human cognitive history that we make new discoveries and come to think new thoughts, which require new concepts. Yet our ability to think about a topic is based on available concepts, and our ability to communicate with others requires shared concepts. As a result, existing concepts must provide the material for introducing new concepts. Sellars holds that the key to this process lies in the construction of analogies. Analogy is central to Sellars' philosophical project:

> If the notion of one family of characteristics being *analogous* to another family of characteristics is obscure and difficult it is nevertheless as essential to the philosophy of science as it has been to theology and, it would seem, somewhat more fruitful. That it is a powerful tool for resolving perennial problems in epistemology and metaphysics is a central theme of this book.
>
> (SM 18)

He later adds: "the use of analogy in theoretical science, unlike that in theology, generates new determinate concepts" (SM 49).

Sellars' most detailed discussions of analogy focus on the introduction of the new descriptive concepts that are required when we postulate the existence of entities not previously considered. I will focus first on this case – in particular on the so-called "theoretical entities" introduced in the development of science. Note especially that the phrase "introducing a theoretical entity" is a euphemism for *introducing a new descriptive concept*. The aim of such concepts is to describe items that exist – if they exist at all – independently of the theories we invent to describe them. When scientists "introduce molecules" they are introducing the concept MOLECULE into a physical theory, and (from a realist perspective) in accepting that theory, we accept the hypothesis that this concept is instantiated in the domain under study.

### *4.5.1 Theoretical Entities*

I will begin with Hesse's (1966, originally published in 1963) account of the role of analogy in introducing new theoretical entities, and Sellars' critique of that account. Hesse holds that new theoretical entities are always introduced into science on the basis of an analogy with familiar objects that provide a *model* for the new entities. The properties that characterize the model fall into three classes: the *positive analogy* consists of features that also characterize the new entity; the *negative analogy* encompasses features that the new entity does not share; the *neutral analogy* consists of features whose possession by the new entity is an open question (1966: 8). The neutral analogy is especially important since, Hesse argues, attempts to move properties in this class to either the positive or negative analogy provide the

driving force for new discoveries. For example, molecules were once conceived of as small, hard particles like tiny billiard balls. Mass was part of the positive analogy, color part of the negative analogy, and being composed of smaller particles was part of the neutral analogy. On this account we can, if we are clever enough, find analogies between any items whatsoever, and this is as it should be. Analogy is a pragmatic notion, and the analogies we draw are a function of what we are attempting to accomplish in a specific case.

Thus far Sellars and Hesse agree; disagreements arise because Sellars pursues a more ambitious use of analogies than does Hesse. Hesse argues that while a new theoretical entity is conceived of as similar to, but not identical with, the model, to avoid an infinite regress similarities must be analyzed in terms of identities and differences. After discussing some examples Hesse writes:

> These examples suggest that when similarities are recognized they are described in some such way as, "Both analogues have property *B*, but whereas the first has property *A*, the second has instead property *C*. It may be that when the nature of the similarity is pressed, it will be admitted that the analogues do not have the *identical* property *B*, but two *similar* properties, say *B* and *B′*, in which case the analysis of the similarity of *B* and *B′* repeats the same pattern. But if we suppose that at some point this analysis stops, with the open or tacit assumption that further consideration of difference between otherwise identical properties can be ignored, we have an analysis of similarity into relations of identity and difference.
>
> (1966: 70–71)

Sellars' problem with this account is its requirement that analogies ultimately be traced back to identities and differences among properties that are not themselves introduced by analogy. These properties provide the ultimate content of all concepts, and the resulting view looks very much like the standard empiricist approach that Hesse later criticized (1970b). As a result, Hesse's account prevents us from appreciating "how the use of models in theoretical explanation can generate *genuinely new* conceptual frameworks and justify the claim to have escaped from the myth of the given" (SRII 183–84).

To allow for the introduction of new frameworks and new content, Sellars maintains, analogies can be based on similarities, not on identities, at the level of *first-order* properties. Identities are required, but these can be identities of *higher-order* properties. In SRII Sellars does not actually give an example of a new concept introduced in this manner. Instead, he illustrates the role of second-order properties by noting that we can draw an illuminating analogy between ordered points on a line and successive moments of time without maintaining that spatial points and temporal moments share

any first-order properties. It is sufficient that the ordering relation share such second-order properties as transitivity and asymmetry (SRII 180).[29] To fully appreciate the significance of higher-order properties in the construction of analogies, we must consider another Sellarsian theme. Sellars stresses that analogies are always accompanied by a *commentary* "which *qualifies* or *limits* – but not precisely nor in all respects – the analogy between the familiar objects and the entities which are being introduced by the theory" (EPM 182; cf. SRII 182.) A similar idea is implicit in Hesse's discussion, but a Sellarsian commentary can be considerably richer than just noting the positive, negative and neutral analogies. In particular, we are not limited to explicitly formulable identities of the kind used in the analogy between space and time. Rather, we are free to use the full resources of available language, which "*does* contain adequate resources for referring to second order attributes by more complex locutions" (SRII 181). These include;

> all the techniques of indefinite reference and definite description. And when we take into account the open texture and vagueness with which reference can be made, we begin to see how models can be the *fundamenta* of open-textured reference to second-order attributes.
>
> (SRII 182)

In a reply, Hesse (1970a: 177–78) argues that Sellars' objection to the reduction of analogy to identity of first-order properties would seem to hold for higher-order properties as well. It is not clear how, on the view that Sellars presents in SRII, we could ever introduce new second-order properties. I will argue below that the resources for replying to this objection are available in Sellars' theory of concepts. I want to work my way towards that argument by examining a further point of disagreement between Sellars and Hesse.

For Hesse the *meaning of a theoretical term* is determined by the model; but for Sellars "one knows the meaning of a theoretical term when one knows (*a*) how it is related to other theoretical terms, and (*b*) how the theoretical system as a whole is tied to the observation language" (EPM 192). It is possible, in principle, for the meanings of the terms of a scientific theory to be "fully captured by the working of a logistically contrived deductive system" (SRII 179).[30] Moreover, Sellars *contrasts* this account with the account in terms of a model (EPM 192). This suggests that while we may use a model as a vehicle for introducing a new concept, models are heuristic devices that do not play a permanent role in determining the meanings of theoretical predicates. This temporary role for models is in accord with Sellars' view that there is no special set of concepts that forms a permanent part of our cognitive endowment. Any system of concepts we use in any domain can be replaced (even though, as we have seen, Sellars holds that it would be a methodological error to replace our observation concepts now). In spite of these clear statements, Sellars seems ambivalent on the point. While he rejects Hesse's account of the meaning of theoretical terms, he is

equally concerned to reject Nagel's view that models have an important heuristic function, but make no contribution to the content of theories. Sellars attributes to *Nagel* the view that:

> the scientific *content* of the theory derives not from models or analogies but from the implicit definition of theoretical predicates by the postulates in which they occur, together with the correspondence rules which connect theoretical terms with expressions in the empirical hierarchy.
>
> (SRII 178)

But, Sellars adds, "This is at best a half-truth." Half-truth is 50 percent more truth than Hesse would concede to Nagel's view. Sellars explains:

> It is a half-truth because theoretical postulates are often specified in a way which *logically* involves the use of the model. And even when a set of postulates is explicitly given in the form prescribed by contemporary logical theory, it turns out, in actual practice, (although *ideally* it *need* not) that the conceptual texture of theoretical terms in scientific use is far richer and more finely grained than the texture generated by the explicitly listed postulates.
>
> (SRII 178–79)

But Sellars' use of "logically" in this passage is confusing. He seems to be aligning what is logically possible with what occurs in practice, and contrasting this with what may occur ideally.

This apparent ambivalence on the role of models in determining the meaning of theoretical terms occurs in other places as well. In one discussion Sellars emphasizes that models have several purposes:

> The most obvious is to make the theory intuitive, and aid the imagination in working with it. But more than this it fills an important need in that whereas the basic magnitudes of the empirical framework are operationally defined and are therefore rooted in a background of qualitative content, the basic magnitudes of the theoretical framework, in the absence of a model, would in no way point to a foundation in nonmetrical, qualitative distinctions which might stand to the them as the qualitative dimensions of observable things stand to empirical properties which are operationally defined with respect to them. . . . Now by virtue of their vizualizable character, models provide a surrogate for the "qualitative" predicates which must, in the last analysis, be the underpinning of theoretical magnitudes if they are to be the sort of thing that could "really exist," if this phrase can be given a stronger interpretation than that of the irenic instrumentalist.
>
> (TE 70)

In spite of these strong words, one paragraph later Sellars concludes:

> The reference of the theory, if it can be said to have reference, and the meanings of the predicates of the theory, insofar as these are more than an adumbration of things to come, are to be understood in terms of the deductive system and the coordination of the theory with the empirical generalizations it is designed to explain.
>
> (TE 71)

In another discussion (EPM 182) Sellars insists that while the standard account of theories "does throw light on the logical status of theories, it emphasizes certain features at the expense of others." In particular, the standard view "gives a highly artificial and unrealistic picture of what *scientists have actually done in the process of constructing theories* [emphasis added]. I do not wish to deny that logically sophisticated scientists today *might* and perhaps, on occasion, *do* proceed in true logistical style." Nevertheless, "the fundamental assumptions of a theory are usually developed . . . by attempting to find a *model*, i.e., to describe a domain of familiar objects behaving in familiar ways such that we can see how the phenomena to be explained would arise if they consisted of this sort of thing."

One way of reading these remarks is in terms of the distinction between what occurs in the context of discovery, when scientists are building theories, and what occurs when we are analyzing completed theories. From this perspective models would play their central role only in the former context, but would vanish in the latter case. Yet if this is Sellars' point, he is unfair to Nagel who would invoke exactly this distinction in his own account of the place of models in science.

I submit that Sellars is indeed drawing on the distinction between the two contexts, but not *using* it in the way that logical empiricists used it. In the first place, Sellars considers the analysis of what occurs in the context of discovery to be a proper philosophical endeavor, and central to epistemology: "human discourse is discourse for *finding things out* as well as for expressing, in textbook style, what we already know" (CDCM 250).[31] Second, for Sellars, we are in the process of "finding things out" for the long-term – until we reach the final science. In our present situation models, along with observation concepts, play a special methodological role: Theoretical concepts are introduced by means of models, and our current understanding of these concepts depends on those models. For working scientists it is models – not correspondence rules – that connect observation terms to the theoretical system. As a result, models are deeply implicated in our ongoing use of theoretical concepts. When the final science is achieved, models will vanish along with the distinction between observational and theoretical concepts.

One consequence of this reading is that it seriously reduces the significance of Sellars' official account of the meanings of descriptive terms. Meanings determined in accordance with the official account now seem to

play just two roles. First, the account applies to observation terms. In this case Sellars' key thesis is that there is a systemic aspect even to the meanings of these terms. Second, the account will apply to theoretical terms in the distant future. But until we reach that point, models that are built on observables will continue to play a fundamental role in scientific thinking. A further consequence of this reading is that Sellars' attempt to find a way between Hesse's and Nagel's accounts of theoretical terms amounts to holding that Hesse's account of the role of models in scientific thinking is right for ongoing science, while Nagel's view is right for completed science.

We can further clarify Sellars' account of the role of models in ongoing science by distinguishing two views:

> M1. Introduction of new theoretical concepts requires a model that is based on familiar concepts – theoretical or non-theoretical; once we have mastered a new concept we can dispense with the model.
>
> M2. Introduction of new theoretical concepts requires a model that must be traced back to some non-theoretical concepts; the model used to introduce a theoretical term is permanently implicated in the meaning of that term.

The difference between M1 and M2 can be sharpened by a schematic example. Suppose we use A as a model for introducing B, then B as a model for introducing C. What role does A play in the introduction of C and in our subsequent use of C in thinking about Cs? On M1, A need not play any role in the introduction or use of C. Whatever role A played in introducing B, B can become an autonomous concept (more precisely, a member of an autonomous conceptual system); once this occurs, B is all we need for the next cognitive step. On M2, A plays a continuing role in our understanding of B. As a result, A is also implicated in our understanding of C, and in our grasp of any further concepts that are introduced taking C as our model.

It seems clear that M2 is Sellars' view of theoretical concepts for ongoing science. To be sure, Sellars insists that the world may be very different than it appears to us in our sensory experience, and that we are capable of learning this. He is quite clear that questions about the meaning of theoretical terms, and about the ontological status of the items these terms are about, are distinct questions. He is also clear that our ability to detect items with our unaided senses is irrelevant to their ontological status. But on Sellars' view of scientific *methodology*, observation concepts should play a continuing role in determining the meanings of theoretical terms until some hypothetical distant date when the scientific quest is complete. In ongoing science, theoretical terms get their working meanings from models, not from implicit definitions and entry transitions. Since the process of constructing models begins from observation terms, these terms enter into the meanings of theoretical terms through these models. Still, Sellars holds, it does not

follow that theoretical concepts introduce nothing conceptually new exactly because we do not explicitly define theoretical terms by means of observation terms: "analogical concepts in science are methodologically dependent on a conceptual base to which they are not reducible" (SM 21).

In Ch. 5 I will defend M1 by extending Sellars' views on concepts and analogy. I want to pave the way for that discussion by considering Sellars' two most developed examples of how analogies could be used to introduce new concepts: Introduction of the concepts SENSORY IMPRESSION and THOUGHT into a community that does not already have these concepts. For reasons that I will consider towards the end of the discussion, Sellars holds that these are not genuine theoretical concepts. Nevertheless, he holds that it is illuminating to view them as if they were. Note especially that Sellars is not defending an historical thesis about how these concepts were introduced. Rather, he is attempting to show how these concepts *could have been introduced* given a restricted language in which they do not occur.[32] Sellars begins with an imaginary "Ryleian" community whose language allows only for the description of overt behavior, and he attempts to show how we could use analogies to transcend these limits. I am not concerned to evaluate these accounts; they deal with issues in the philosophy of mind that are beyond the scope of this book. I am using them only as means of further explicating Sellars' doctrine of analogy.

Consider SENSE IMPRESSION first.[33] We begin by imagining a stage in the development of knowledge at which this concept does not exist. People talk about seeing items in the physical world without any notion that internal processes in the perceiving organism are implicated in perception. However, they find two features of perception puzzling. First, objects sometimes looks as if they have properties that, on further examination, they turn out not to have. Second, they sometimes have the experience of seeing an object when no such object exists. To account for these situations they introduce a new concept and a causal hypothesis. I will consider the hypothesis first, taking the concept for granted; then I will examine Sellars' account of how the concept is formed.[34]

The causal hypothesis is that normal vision takes place when light impinging on a perceiver's eyes produces an internal state of the perceiver which is called a *visual impression*. The presence of a visual impression is causally sufficient for the experience of seeing an external object to occur, and the specific features of the impression determine the visual properties that we perceive. Misperception of an actual object occurs when features of an impression differ from what those features would be in the normal perception of that object; hallucinations occur when an impression is produced without being caused by any physical object. Detailed explanations of how these aberrant conditions come about are subjects for empirical investigation. Here is Sellars' outline of the account:

> the framework of sense impressions involves a causal hypothesis, the general character of which can be indicated by saying that the fact that

> blue objects appear in certain circumstances to be green, and that in certain circumstances there appear to be red and triangular objects in front of people when there is no object there at all, are explained by postulating that in these circumstances impressions are brought about of the kinds that are normally brought about by blue objects (in the first case) and by red and triangular objects (in the second).
>
> (P 94)

The concept of an impression required by this account can, to a first approximation, be viewed as a theoretical construct (P 91–92, cf. EPM 150–51, 190–95; SM 9) introduced by analogy with properties of the physical objects that normally cause those impressions: "Interpretation of the framework of sense impressions as a theoretical framework suggests that the analogy between the attributes of impressions and the perceptible attributes of physical objects is but another case of the role of analogy in concept formation" (SM 21). As a first step, the impression of a red triangle is introduced as the internal state caused by objects with red triangular surfaces under normal conditions. But this is not the whole story since we can say more about this new entity than just describe its usual cause: "The fact that *impressions* are theoretical entities enables us to understand how they can be *intrinsically characterized* . . . " (EPM 192). Analogies come into play in developing this intrinsic account: "visual impressions of red triangles are conceived as items which are analogous *in certain respects* to physical objects which are red and triangular on the facing side" (P 93, cf. EPM 192–93; SM 18–23). The qualification "in certain respects" is vital. The impression of a red triangle is neither red nor triangular.

> The *essential* feature of the analogy is that visual impressions stand to one another in a system of ways of resembling and differing which is structurally similar to the ways in which the colours and shapes of visible objects resemble and differ.
>
> (EPM 193)

An impression need not have all the properties of its model, and it can have properties that are incompatible with properties of the model as long as we leave out those features of the model that would generate a contradiction. Note especially that in the present case the analogy is between "sense impressions and physical objects, not between sense impressions and *perceptions* of physical objects." Thus "the analogy is a trans-category analogy, for it is an analogy between a state [the sense impression] and a physical thing [the perceived object]" (P 93, cf. EPM 191).

With these points in mind Sellars states "the positive analogy" in two parts:

> (*a*) Impressions of red, blue, yellow, etc. triangles are implied to resemble-and-differ in a way which is formally analogous to that in

> which physical objects which are triangular and (red or blue or yellow, etc.) on the facing side resemble-and-differ; and similarly *mutatis mutandis* in the case of other shapes.
>
> (P 94)

Part (*b*) of this account is the same as part (*a*), but with the roles of colors and shapes interchanged. Here is the upshot:

> In effect, these analogies have the force of postulates, implicitly defining two families of predicates '$\Phi_1$' . . . '$\Phi_n$' and '$\Psi_1$' . . . '$\Psi_n$', applicable to sense impressions, one of which has a logical space analogous to that of colours, the other a logical space analogous to that of the spatial properties of physical things.
>
> (P 94)

In addition, relations between visual impressions of colors and of shapes are analogous to the relations between physical colors and shapes (SM 24–26). "Succinctly put, impressions have attributes and stand in relations which are counterparts of the attributes and relations of physical objects and events" (SM 26).

Several features of impressions follow from the way they are introduced. First, consider the claim that the impressions associated with seeing red are just those that we experience when we see red surfaces under normal conditions. This claim is neither a *definition* of "impression of red," nor a part of such a definition:

> 'impression of a red triangle' does not simply mean 'impression such as is caused by red triangular objects in standard conditions', though it is true – *logically* true – of impressions of red triangles that they are of the sort which *is* caused by red and triangular objects in standard conditions.
>
> (EPM 192)

Sellars takes this to be a logical truth because, given the role that red surfaces seen in standard conditions play in introducing IMPRESSION OF RED, it would be a contradiction to claim that any impression other than an impression of red occurs when we perceive a red physical object in standard conditions.[35] Still, if impressions were defined by their causes, SENSE IMPRESSION would not introduce new conceptual content. It is the analogical construction that is responsible for this new content.

Second, once we have the concept of a sense impression, a visual encounter with a red surface might result in an ET either to the concept RED OBJECT, or to IMPRESSION OF RED, depending on our current interests. The difference between the two conceptualizations lies in the further inferences that are licensed. If I move into the conceptual system of physical objects, it follows that there actually is a red physical object in front of me, an object that is avail-

able for further sensory exploration by myself and by others. If I move into the conceptual system of impressions, this implication does not obtain (P 92).

Third, since impressions are inner episodes they are, in a sense, private: each of us has a form of privileged access to our own impressions. However, while I have a special way of detecting my own impressions that is not available to others, it does not follow that others lack access to my impressions. Given the way impressions have been introduced, there clearly are situations in which other people have reasonable grounds for attributing specific impressions to me. If you see me looking in the direction of a red object in good light with my eyes open, it is reasonable for you to conclude that I am having an impression of red. Thus impressions "combine *privacy*, in that each of us has privileged access to his own, with *intersubjectivity*, in that each of us can, in principle, know about the other's" (EPM 176, cf. 195). Moreover, my access to my own impressions is not guaranteed to be more reliable than other people's access. Identification of an impression involves classification, so mistakes are possible. Someone who is aware of the circumstances in which an impression has been produced may be able to tell me that I have misclassified my present sensory state.

Fourth, the fallibility of my descriptions of my own impressions stands as one feature that distinguishes impressions from sense data. Three other features will underline this distinction. (1) Unlike sense data, the impression of red is not literally red, although it is analogous to red in certain respects. (2) "Sense-impressions are non-conceptual states of consciousness" (SM 10), not objects of consciousness.[36] (3) While one primary function of sense data is to provide the epistemic basis of empirical knowledge, impressions have no such epistemic function:

> the direct perception of physical objects is meditated by the occurrence of sense impressions which latter are, in themselves, thoroughly non-cognitive . . . this mediation is causal rather than epistemic. Sense impressions do not mediate by virtue of being known.
>
> (P 90–91)

Impressions share this feature with many other states. But, again, a non-epistemic state can still be known, and known by its subject in ways that are not available to others (see SS for further discussion).

Now consider a second class of "inner episodes": *thoughts*. Again we are assumed to be in a community that uses language only to talk about public objects and behaviors; Sellars' project is to show how a substantive concept describing inner episodes can be introduced by analogy. There is, once more, a key phenomenon to be explained: "a person's verbal propensities and dispositions change during periods of silence as they would have changed if he had been engaged in specific sequences of various types of candid linguistic behaviour called 'thinkings-out-loud' by our Ryleians . . . " (SM 151, cf. 87–88, 159). The explanation postulates

thoughts: a class of inner episodes that can cause overt speech. Thoughts occur independently of speech, and do not always result in speech, but overt speech provides the model for introducing THOUGHT. Sellars assumes that before this concept is introduced the community's language has already been enriched to include semantical discourse: "the resources necessary for making such characteristically semantical statements as '"*Rot*" means red', and '"*Der Mond ist rund*" is true if and only if the moon is round' (EPM 179). Although this is a significant enrichment, it still deals with purely public items. Then, "*using the language of the model, the theory is to the effect that overt verbal behaviour is the culmination of a process which begins with 'inner speech'* . . . " (EPM 186). However, "It is essential to bear in mind that . . . 'inner speech' is not to be confused with *verbal imagery*" (EPM 186). Since THOUGHT is modeled on overt speech, we can apply semantical categories to these inner episodes and describe them "as *meaning* this or that, or being *about* this or that" (EPM 187). But we must not forget that this is an analogous use of semantical terms, whose primary use is to describe overt linguistic behavior (EPM 188).

Thoughts, then, are private mental episodes that are introduced by analogy with overt speech. Moreover (as in the case of sense impressions) once we have learned the theory, "it is but a short step to the use of this language in self-description" (EPM 189). People first learn to attribute thoughts to others, taking speech as evidence for the occurrence of these thoughts. They then learn to attribute thoughts to themselves, giving "reasonably reliable self-descriptions, using the language of the theory" (EPM 189) without having to first describe their own behavior: "*What began as language with a purely theoretical use has gained a reporting role*" (EPM 189). Again we have a kind of privileged access in that people have ways of detecting their own thoughts that are not available to others. But we also have ways of detecting other people's thoughts, and our accounts of our own thoughts are fallible and subject to correction by others.

I noted above that Sellars describes impressions as theoretical concepts only to a first approximation. Similarly, in the case of thoughts he holds that "their status might be illuminated by means of the contrast between theoretical and non-theoretical discourse" (EPM 188), although they are not genuine theoretical concepts. The reason for this limitation is our ability to directly detect our own impressions and thoughts. Once we learn to report our own impressions and thoughts, they fit his characterization of observation predicates as "Those descriptive predicates which are conditioned responses to situations of the kind they are correctly said to mean . . . " (SAP 316). This characterization applies to impressions and thoughts *now*, even if they might have been introduced by a process that parallels the process by which theoretical entities are introduced. Recall also that Sellars does not claim that his account of the introduction of these concepts is historically correct. In any case, my aim in discussing these examples is solely to illustrate Sellars' account of the use of analogy to introduce new concepts.

### *4.5.2 Modifying Formal Concepts*

I have been examining the use of analogy to introduce concepts that describe entities, but we introduce new concepts whenever we introduce new ways of thinking into a subject. Sellars also applies his account of analogy to formal concepts, although his discussions are rather sketchy; I will go beyond his texts. Consider Euclidean geometry (e.g., CC; MFC 344–46; SM 128–30). Before the introduction of non-Euclidean geometries in the nineteenth century there was no reason to attach a modifier to "geometry"; once the new geometries were developed, "geometry" became a generic term with many species.[37] Why should we consider these new constructs to be *geometries*? The Sellarsian reply is that there are systematic similarities between these constructs that justify treating them as specific instances of a more general class. Moreover, doing so provides considerable insight into the properties of these constructs. For example, the first non-Euclidean geometries were constructed by making specific alterations in Euclid's parallels postulate. The changes yield new systems that can be systematically compared with Euclid's original version. Euclid does not use his parallels postulate until his proof of Proposition 29. Theorems proved without using this postulate are known as "absolute geometry" and are common to several geometries. Systematic differences among geometries appear when we explore further theorems. We can, for example, specify a common definition of a triangle that holds for various geometries, and then show that the relation between the area of a triangle and the sum of its internal angles differs for different variations on the parallels postulate. Comparisons of this sort are common in geometry textbooks. The generalized Riemannian geometry mentioned in the most recent note carries this approach further, allowing us to completely characterize different geometries in terms of numbers in a matrix (the metric tensor). The Sellarsian approach places these comparisons in a wider context and clarifies their relations to other cases of conceptual variation and innovation.

Sellars also uses alternative logics to illustrate alternative formal concepts: "Classical negation and intuitionistic negation are varieties of negation" (CC 90; MFC 435). As a formal concept, negation is completely constituted by implications; the two forms of negation share many implications, although they differ in one key respect: intuitionistic negation does not allow an inference from not-not-$p$ to $p$. One consequence of this restriction is that excluded middle is not a logical truth in intuitionistic logic. This does not mean that excluded middle is false, but only that it cannot be assumed without independent proof. In any case in which excluded middle can be proven, all implications of classical logic are available. Sellars underlines the point that we are dealing with different forms of negation by comparing a possible evolution of chess:

> Suppose that at one point in the history of chess the piece which was checked and checkmated could capture like a knight as well as on adjacent

> squares. Suppose that shortly thereafter, following a period of controversy, the community of chess-players decided that the game would be improved in certain respects if this power to capture like a knight were dropped. Would we not be willing to say not only that the game has changed, but that the king has changed? It is not as though a dog vanished and a cat took its place.
>
> (CC 91–92)

There is no algorithm for determining when we should view two constructs as species of a common genus. Ultimately the decision depends on whether doing so deepens our understanding, and we may adopt different classifications for different purposes. Note also that Sellars' approach leads directly to the view that there are degrees of similarity between conceptual systems (e.g., MFC 434; SM 128–30). He holds that such degrees of similarity must be taken into account if we are to adopt a realistic philosophy of science in which our conceptual systems evolve towards a correct account of our subject matter (SM 95, n. 1).

## 4.6 Conclusion and Preview

There is a glaring problem with Sellars' account of the analogical introduction of new concepts: He tells us little about the content of these new concepts. We are told, for example, that color impressions resemble and differ among themselves in ways that are analogous to the ways in which facing surfaces of colored physical objects resemble and differ. But what are the actual resemblances and differences among impressions? If, as Sellars maintains, analogies provide full-blown concepts, we ought to be able to provide analyses of these concepts that go beyond vague general references to their models. A similar point applies to Sellars' remarks on comparisons of conceptual systems, which also include few substantive details. I think Sellars' theory of concepts has resources that will allow us to do considerably better, although Sellars never makes use of these resources. In Ch. 5 I will propose a theory of concepts that includes emendations that are largely elaborations of material already present in Sellars' writings, and that will provide the basis for more detailed accounts of conceptual change than Sellars provides. We will also find that the Sellarsian approach has considerably wider scope than the kinds of cases he discusses. While Sellars' discussions of conceptual change in science focus mainly on the introduction of new entities, this is not the only important case. For example, in the transition from Ptolemaic to Copernican astronomy reclassification of the earth as a planet and the sun as a star had revolutionary significance, but did not involve the introduction of new entities. In addition, sometimes items that are not properly thought of as entities or processes are reconceptualized; the reformulation of the concepts of space and time in special relativity is one example. A properly elaborated Sellarsian approach will provide the tools we

need for a systematic account of the ways in which these concepts are the same as, and different from, their predecessors.

The last point has wider application. The theory of concepts I develop in Ch. 5 provides a set of guidelines for comparing successive or competing conceptual systems from some stage in the development of science, as well as for comparing the conceptual systems of different philosophers. I will strengthen the basis for Sellars' account of how genuinely new concepts can be introduced by making systematic changes in available concepts, and of how we can move by a continuous process from one set of concepts to a new set that has nothing in common with its starting point. The approach can also provide a technique for teaching the conceptual systems of past science and philosophy to contemporary students by using analogies with familiar concepts as a bridge to unfamiliar concepts. In some cases this will amount simply to reversing the direction of the analogies that led from earlier to present concepts. The approach may even provide a bridge to learning the conceptual systems of other cultures as long as those concepts are not totally alien.

# 5 Reconstruction

> [N]ovel facts do not sport the concepts by which we grasp them, but . . . we must draw these concepts from our stock-in-trade, refurbishing them as needs be for their new jobs.
>
> (Torretti 1999: 431)

In this chapter I reconstruct the Sellarsian theory of concepts. In Ch. 6 I compare the resulting theory with other contemporary theories, defending and refining it as the need arises. We will have a complete formulation of the theory I am proposing after Ch. 6.

## 5.1 Concepts and Language II

In Sec. 1.4 I gave some initial reasons for avoiding the common practice of treating languages and conceptual systems as identical. Thus far I have not always kept the two apart because I have been discussing philosophers who do not regularly separate them. Since I am now moving towards a formulation of my own view, it becomes important to adhere to this distinction. I want to reinforce and extend my reasons for making the distinction, and then introduce terminology that will help keep the two subjects separate. Introduction of this terminology does not constitute a commitment to the view that concepts can exist without language; it just introduces terminology in which the question can be clearly raised.

Sellars provides an illustration of the need for such terminology since he often writes as if he holds that all thought is linguistic. Recall, however, that Jones used language as the *model* for introducing THOUGHT and treated language as an *analogy* for thought. Since Sellars also holds that genuine thinking takes place in overt language, linguistic utterances do not *always* report underlying thoughts. But Sellars denies that all thoughts are linguistic:

> I find that I am often construed as holding that mental events in the sense of thoughts, as contrasted with aches and pains, are linguistic

> events. This is a misunderstanding. What I have held is that the members of a certain class of linguistic events are thoughts. The misunderstanding is a simple case of illicit conversion, the move from "All *A* is *B*" to "All *B* is *A*".
>
> (ME 325)

Sellars is also open to the hypothesis that non-linguistic animals have concepts. In one discussion, after urging that "we take very seriously the view that a thought, in the sense in which thoughts occur to one, is the occurrence in the mind of sentences in the language of 'inner speech'" he adds:

> Before continuing, I must qualify the above remarks lest the animal lovers among us take them as libel and calumny. I count myself in their ranks and therefore hasten to add that of course there is a legitimate sense in which animals can be said to think. . . . Furthermore, the point is important in its own right and not simply a rhetorical maneuver. For if one ties thinking too closely to language, the acquisition of linguistic skills by children becomes puzzling in ways which generate talk about 'innate grammatical theories'.
>
> (SK 303)

Elsewhere Sellars maintains that animals have representational systems even though they do not have language (ME 328), and that if:

> there is a relevant degree of similarity between the functioning of a certain state, $\phi$, of an animal's representational system and the function of 'this is triangular' in our own representational system, then we can appropriately say $\phi$-states mean *this is triangular*. . . .
>
> (ME 331)

The practice of using "language" and "conceptual system" as synonyms also puts us in an unnecessary verbal straightjacket when discussing descriptive concepts; again Sellars will serve as an example. Sellars holds that beliefs about a domain are embodied in a system of descriptive concepts. Now language is one subject about which we have such beliefs, and the concepts we use for describing features of languages provide useful examples for some of our themes. When we compare the systems of grammatical concepts required to discuss different languages, we find clear examples of systems that are highly similar, but not identical: All languages may have nouns and verbs, but only some have separable prefixes, ablative absolutes, or split infinitives. Sellars is well aware of this point, but his tendency to write as if "language" and "conceptual system" are synonyms, along with his treatment of ETs and DTs as moves between language and the world, leads to unduly clumsy remarks. For example, in a discussion of language learning he notes that at a particular stage of development the learner

> is able to classify items into linguistic kinds, and to engage in theoretical and practical reasoning about linguistic behavior. Language entry transitions now include 'That is a "2 + 2 = 4"' as well as 'that is a table'. Language departure transitions, I will say "2 + 2 = 4" followed by a saying of '2 + 2 = 4', as well as 'I will raise my hand' followed by a raising of the hand. The trainee acquires the ability to language about languagings. . . .
>
> (RM 124–25, I have fixed some minor typos)

It is surely preferable to describe these as transitions between linguistic items and the conceptual systems we use to describe them.

For the remainder of this book I will *systematically* adopt the terminology and conventions that I have been using except where the views of a particular philosopher made them inappropriate. I will reserve the term "meaning" for cases in which I am explicitly discussing linguistic items and will put such items in quotes; I will talk about the *content* of concepts. In addition, since I have already argued that Sellars' ETs and DTs need to be replaced, I will drop this terminology and talk instead about relations between a conceptual system and its *extra-systemic domain* – terminology that underlines the point that a domain is (usually) distinct from the conceptual systems we use to think about it. I will develop replacements for ETs and DTs in this chapter. I will also replace Sellars' talk about "inferential moves" within a conceptual system with talk about *implications* among concepts in a system, or, as I will also call them, *intra-systemic relations*. Here, as in the case of the transitions, my aim is to eliminate all explicitly psychological features from the terminology I use for discussing conceptual content. We will usually be interested in concepts from an abstract perspective where part of the study of a conceptual system (including our own) consists of working out implications we had not noticed. Treating concepts abstractly does not negate Sellars' view that concepts exist only in individual minds – whatever these ultimately turn out to be. The point is only that an account of the content of a concept and an account of how it is embodied in organisms are distinct issues. As noted in Sec. 1.5, a full theory of concepts will require an interplay between the abstract, biological, and psychological perspectives.

I want to note some further advantages of clearly distinguishing conceptual systems from languages. As noted in Sec. 1.4, one unfortunate effect of this assimilation is that it encourages some philosophers (and others) to write as if each natural language embodies its own conceptual system, and then lose sight of two important points. A given concept may be expressible in different natural languages, and multiple, even competing, conceptual systems are expressible in a single natural language. Both points are central to Sellars' thought. He seeks to capture the first point in his dot-quotes, while the second point is fundamental to anyone who considers conceptual change a continuing feature of human thought. In a discussion of the concept of mass Sellars writes,

> the scientist in different contexts uses the term in different senses, according to different rules. In common sense contexts his language is of ancient vintage. Thus we can stick to English and yet be said to speak not one language but many.
>
> (LRB 311–12)

I prefer to speak of a natural language as being capable of expressing many different conceptual systems.

The new terminology will also help in discussing psychological phenomena from a realist perspective. Philosophers often describe realism as the thesis that the objects we study are mind-independent, but realism is not concerned only with studies of the physical world. A full-blown realism includes minds in its scope since we may seek a theory that gives a correct account of the nature of minds; this theory would not be "mind-independent." Alternatively, we could talk about items that exist independently of human beliefs, but this description also breaks down when we take human beliefs as the subject of our study. We can avoid such verbal tangles – and associated conceptual confusions – by talking of systems of concepts that describe items which are not part of that system. This will still leave some conceptual systems that are unavoidably self-referential; I will consider these in Sec. 5.9.

It is worth repeating that the decision to distinguish between conceptual systems and languages does not prejudge the issue of whether having concepts requires having a language; it only opens up linguistic space for independent consideration of the issue. I will consider some issues involved in attribution of concepts to non-linguistic animals in Sec. 5.10.3. Still, my main concern is with human concepts, and with sophisticated activities such as conceptual analysis and the intentional introduction of alternative conceptual systems. These are activities which, as far as we know, only humans pursue. Since these activities typically involve the use of language, I will continue to take linguistic practice as one source of *evidence* about underlying concepts.

## 5.2 Commentaries

Since the issues I am concerned with arise only among those who have achieved a high level of cognitive sophistication, in discussing these issues we are free to use all the linguistic and conceptual tools at our disposal. This is, of course, the normal procedure in philosophy; even those who would reduce all concepts to a set of simple, unanalyzable items argue their case in a metalanguage that is vastly richer than this base. An important example of this procedure is provided by Sellars' notion of a *commentary* that must accompany the analogical introduction of a new concept, and that explains how this concept is similar to, and different from, its model (Sec. 4.5). The cognitive resources employed in presenting a commentary, and the cognitive

resources that are assumed among those to whom a commentary is addressed, provide a paradigm example of what I mean by "cognitive sophistication."

Although Sellars' main use of commentaries occurs when he is discussing analogies, many passages suggest a wider role for this notion. In CDCM, for example, Sellars introduces a symbolism for discussing certain problems about causality, and notes that the correct understanding of this symbolism is not to be found either in the symbols or in the rules that govern their use. Rather, "the informal commentary with which we have been surrounding our use of logistical expressions is essential to their correct interpretation as a transcription of causal discourse" (CDCM 253). A bit earlier in CDCM Sellars ends a discussion of the view that universal generalizations should be considered inference tickets with the remark that this view is acceptable as long as it is accompanied with an appropriate commentary, and he offers the discussion that precedes this remark as an example (CDCM 242). Elsewhere Sellars writes of the role of a "philosophical commentary" in justifying a definition (EAE 432, n. 5); other extensions of this notion occur in other texts (e.g., SM 94, 199). In this more general usage, we provide a commentary whenever we engage in a metalinguistic discussion of our concepts – as opposed to simply using them to respond to some situation. Other metalinguistic discussions, such as proposals to modify a conceptual system or discussions of the range of application of a law, can also be viewed as commentaries on their subject matter. The label "commentary" is not important. What is important is the clear recognition that when we engage in metalinguistic discussions of some subject, we regularly use language, concepts, and other cognitive resources that are richer than those embodied in the subject being discussed.

I mention these obvious points in order to draw attention to a gap between official theory and actual practice that often occurs in epistemology. One common theme of naturalistic epistemology is the need to focus on the actual cognitive abilities of humans as epistemic agents. There are two sides to this requirement. First, we should not attribute to human knowers abilities that they do not have. This has been the main focus of the requirement in naturalistic epistemology where, for good historical reasons, a major aim has been to get away from treating human cognitive abilities as located in some entity (mind, soul, noumenal self) that is not part of the natural world. The second part of this requirement has not been as prominent in the literature: We must not be overly restrictive in the range of cognitive abilities we attribute to ourselves. This is an ever-present danger because contemporary naturalism is, for the most part, a development within the empiricist tradition. Historically, empiricists have been much more frugal in their assumptions about our cognitive abilities than rationalists. Contrast, for example, the minimal cognitive machinery that Hume attributes to us with, say, Spinoza's assumption that we can intuitively (which, for Spinoza, means, infallibly) grasp the fundamental truths that provide his axioms. The caution

we find among empiricists is, in many ways, the lesser of two errors, but it can also lead us seriously astray. If Hume had only the cognitive abilities embodied in his official doctrine, it would not have been possible for him to conceive and elaborate his own philosophy.

It is a familiar, if rarely mentioned, feature of conceptual analysis that analyses of a particular concept typically assume a large body of concepts that we take for granted in formulating and explaining the analysis. While those with foundationalist proclivities may object that an explicit analysis is no clearer than the language in which it is couched, I urge that this common procedure is both unavoidable and legitimate; two analogies may help pin down the point. The first is Neurath's familiar ship that is being repaired piecemeal while still afloat. Here is Sellars' version of this image:

> *Above all*, the picture [i.e., foundationalist empiricism] is misleading because of its static character. One seems forced to choose between the picture of an elephant which rests on a tortoise (What supports the tortoise?) and the picture of a great Hegelian serpent of knowledge with its tail in its mouth (Where does it begin?). Neither will do. For empirical knowledge, like its sophisticated extension, science, is rational, not because it has a *foundation* but because it is a self-correcting enterprise which can put *any* claim in jeopardy, though not *all* at once.
>
> (EPM 170)

The second image develops Sellars' point that the key problem with traditional empiricism is its static character. Assuming a set of beliefs B in the process of evaluating another set B*, and then using B* in a later re-evaluation of B may seem circular from a static perspective. But, as Hooker suggests (1987: 13), from a dynamic perspective we should replace the circle with a helix, in which we return to previously accepted beliefs and re-evaluate them from a richer perspective.

These images are usually invoked when we discuss the evaluation of propositions, but Sellars, in effect, extends them to reflection on the adequacy of our concepts. Sellars holds that we engage in an ongoing process of conceptual improvement, and that each stage of this process builds on whatever concepts we already have. One aim of the present chapter is to develop a more detailed account of this process – which we glimpsed in Sec. 4.5. A central theme of our discussion will be ways in which we manipulate available conceptual material in order to build new concepts, but this raises the question of how we are able to carry out these manipulations. The point I am driving at in the present section is that we clearly do have such abilities, and their recognition is built into Sellars' notion of a commentary. An account of the nature of these abilities should be a key part of a complete naturalistic epistemology, but such an account is beyond the scope of this book. The recognition that we have such abilities is sufficient for present purposes. I will take the extended notion of a Sellarsian commentary

as the model for discussing concepts and conceptual change. Indeed, I have already been doing this, although without explicit mention.

## 5.3 Descriptive Concepts II

Sellars' account of formal concepts is satisfactory given two clarifications: these concepts are constituted by implications – allowed inferences – not by actual inferences; and mastery of a formal system does not require the development of a tendency to actually make any of these inferences. A Sellarsian descriptive system combines a formal system with a set of connections between items in that system and items in the domain described, where these connections provide part of the content of descriptive concepts. But, we saw (Sec. 4.2.3) that Sellars' account of this connection must be replaced. Instead of habits that connect us to items in a domain, we need something that is internal to descriptive concepts, but that directs us to a concept's extra-systemic subject matter. Lewis' notion of *sense meaning* is a step in this direction, although it is inadequate for present purposes because it is limited to sensibles. We need something considerably richer.

One candidate for this element is a set of criteria for recognizing instances of the concept, and these will serve when available. This proposal is appropriate for physical objects that we can pick out by unaided perception. Sellars would classify these as observables, and we can give criteria for recognizing chairs, planets, trees, elms, and beeches. Depending on the particular case, criteria may include size, shape, and other typical properties, plus components, familiar examples, and functions. Criteria of this sort are included in dictionary definitions of the associated terms. Consider a definition of the noun "chair": "a seat, esp. for one person, usually having four legs for support and a rest for the back and often having rests for the arms."[1] This definition will not be much help to linguistic beginners, but neither dictionary definitions nor conceptual analyses are intended to teach language to beginners. Definitions and analyses are commentaries in which we may draw on all of the linguistic and conceptual resources at our disposal.

Typically, these criteria do not provide necessary and sufficient conditions for the application of the concept, and this is appropriate. Everyday concepts are our creations and are usually open-ended because we develop them only to the extent required for our practical concerns. Such concepts rarely elicit the level of theoretical interest that lead to formulations with the precision of mathematical definitions. When we encounter a new item that does not quite fit existing concepts – such as a beanbag chair – we must *decide* whether to adjust the concept to include this item. We do not settle the question by peering more accurately into an established concept. In this case we treat the item as a new kind of chair because we allow function to dominate considerations of shape and parts. Moreover, even when we have a complete set of necessary and sufficient conditions for a concept, the situation does not

change significantly. While the square root of a negative number may not be a number according to a once-prevailing concept, we are free to drop the old concept from our repertoire and substitute a new one – which we now associate with the word "number" – and which includes this case in its extension.

This approach applies to other cases. We can, for example, provide criteria for relations that are detectable with our unaided senses, such as TALLER THAN and TO THE LEFT OF. In a similar way, a description of the characteristic properties and activities of unicorns will be included in an analysis of UNICORN, and play a key role in our reasons for believing that no such animals exist. The approach also applies to many concepts that are wholly functional. Here is a dictionary definition of another sense of "chair": "to preside over; act as chairman of; *to chair a committee*." Note especially that the example at the end of the passage calls attention to a situation in which the reader might have encountered this term. This provides information that will help a competent language user recognize instances of the concept.

Now consider how this approach applies to *qualities*. These are of special interest because of the long history of treating some qualities as simples that can be introduced only by ostension, and that are not amenable to explicit analyses. On the view I am proposing there are no simple concepts, so we should consider how we provide criteria for identifying qualities. Again, dictionary definitions provide a useful guide. Here is a portion of a definition of the noun "red": "any of various colors resembling the color of blood; a color at the extreme end of the visible spectrum." The definition works by providing examples that may be familiar to the reader. These examples would be appropriate for those philosophers who would define "red" as (say) the color of blood in standard conditions. The examples also meet the requirement that Sellars invokes when he tells us that if someone does not classify *these* as *rot*, then *"rot"* does not express the same concept as our "red." Still, the definition provides more than just examples. By classifying red as a color the definition indicates that we are dealing with a visible item, and this point is relevant to both implications and to characteristic instances since it narrows the range of situations in which we should seek examples.

I turn next to theoretical concepts, which take us outside the range of unaided perception. Theoretical concepts run a gamut from ultraviolet light, which is barely beyond our normal visual range, to magnetism, which is easily detected even though we cannot perceive it, to neutrinos and quarks that can be detected only in a highly indirect manner. (See Brown 1987, 1995; Galison 1987, 1997, and Shapere 1982 for detailed discussions of detection processes in contemporary science.) This class also includes many familiar relations, such as FATHER or SISTER-IN-LAW, whose instances are not recognizable on the basis of unaided perception of the relata. In all such cases we cannot tie a formal structure to its domain just by describing observables; more complex criteria will be required. Still, we must allow for

a range of cases. A compass will allow us to detect the magnetic field of the earth in our immediate vicinity, and do so at a glance; a Geiger counter will do the parallel job for local radioactivity. But in some cases the identification may come long after the fact as a result of a painstaking analysis. The use of bubble chambers to identify processes involving specific particles provides an example. Bubble-chamber experiments produce hundreds of thousands of photographs that are analyzed over a substantial period of time.[2] Identification of a particular type of event may occur months after the experiment was completed as a result of analysis of photographs by a technician who was not present at the experiment. In many cases a specific instance in which the item of interest occurred will be picked out by means of a characteristic signature, and the time and place at which it occurred in the chamber will be established. We arrive, again, at criteria for identifying specific instances of the item in question, but criteria that are more complex and indirect than those used for identifying observables.

The top quark requires a further modification of our requirements: we have evidence that they exist, although we cannot specify a case in which the particle occurred. I want to emphasize that we should adapt our requirements, rather than challenge whether physicists actually have the concept TOP QUARK. Given the role that TOP QUARK plays in physics, and given that physicists were able to use this concept (in conjunction with available information and theory) to develop an experimental test for the existence of top quarks, we have adequate evidence that physicists have a fully developed concept of a top quark. No theory of concepts in the literature is sufficiently powerful and well supported to provide reasons for concluding that physicists do not understand what they are doing. If a theory of concepts yields that conclusion, it is the theory that must be reconsidered.

I propose that the situation illustrated by the top quark provides the requirement we need. In order to establish the relation between a formal system and a domain we must include *an account of the criteria for assessing if the concept is instantiated*. Where we can provide means of identifying specific cases in which instantiation occurs, these criteria should be included in the concept, but they are not *required*. Nor does the presence of these additional features somehow provide a "better" concept. There are, I think, no grounds for holding that the everyday concept of a chair is clearer, richer, more precise, or in any way better developed than the high-energy physicists' concept of a top quark. I will henceforth refer to this aspect of conceptual content as the *instantiation criterion* (IC).

The account I am proposing is intended to apply to all descriptive concepts, not just those of everyday experience and physical science. We use descriptive concepts whenever we seek to describe items in some domain, and the content of any such concept must include an IC. Even a set of theological or magical concepts may constitute a genuine system of descriptive concepts since the *aim* of such a system is to describe features of the universe – miracles, saints, deities, or whatever. An attempt at constructing

such a system can fail in either of two ways: It can fail to be sufficiently well elaborated – perhaps because it lacks criteria for assessing if the entities it postulates exist; in this case we have not produced genuine descriptive concepts. Or it can be adequately elaborated but not instantiated. We do not have to show that a system of descriptive concepts is incoherent to show that it is a failure.

Many concepts have multiple ICs, and this is desirable. Multiple independent criteria allow us to seek visual confirmation of something we hear, or tactile confirmation of something we see. A concept that describes something that can be seen but not touched (such as a hologram) differs significantly from the usual physical-object concepts. Scientists consider it highly desirable that items be detectable in more than one way. Note two different situations in which this occurs. In the case of weak neutral currents two different teams set out to determine if the phenomenon exists using different kinds of detection equipment. In the case of the Ψ/J particle two different groups engaged in different projects made the same unanticipated discovery using different equipment – thus the double name (Pais 1986: 605). The availability of distinct ways of detecting an item strengthens the claim that it exists. Suppose, by way of contrast, that someone asserts the existence in our environment of a class of items that can be detected *only* by those with a particular mutation. If this item does not interact at all with other items in ways that the rest of us can detect, there would seem to be no point to this postulate. If such interactions do occur, we have a basis for basis for developing alternative ICs.

A question now arises about cases in which ICs are added or deleted. In general, any change in the ICs associated with a concept results in a different concept. Sometimes the changes are minor; sometimes they are drastic. Recall that the discovery of isotopes led to the deletion of a characteristic weight as a criterion for identifying an element, while half-life provided a new criterion for identifying radioactive elements. In this case, the new concept associated with "element" was inconsistent with its predecessor, but still shared important features. In general, it is more informative to describe what has changed and what has remained unchanged than to simply note that a new concept has been introduced. We must also distinguish cases in which an IC is added from those in which it is *discovered* – that is, deduced from a concept in conjunction with available background knowledge. This does not involve conceptual change.[3] Such discoveries stand out clearly when we focus on concepts from an abstract perspective and consider implications – which are determinate independently of whether anyone has noticed them.

My approach includes another departure from Sellars' views. Sellars adopts the calculus-plus-correspondence-rules account of theoretical concepts in developing science. On this view it is not required that each concept in the system be individually related to observables, but Sellars goes further. He holds that the correspondence rules *must not* provide a

one-one connection between theoretical concepts and empirical evidence: "If they did, the 'theory', if successful, would simply be a representation of empirical generalizations in the form of a deductive system" (TE 148). This is a surprising remark since the role of implicit definition seems to have vanished. I suggest, instead, that if we have two non-equivalent axiom systems, each with the same one–one connection between concepts and data, we still have two different conceptual systems. The view I am proposing requires that each concept in a system of descriptive concepts be related to extra-systemic items by ICs, but the consequence that concerns Sellars does not follow since the two theories may still have different intra-systemic relations.

Our account can help clarify the status of a class of cases introduced by Putnam (1975), cases such as ELM and BEECH in the minds of a novice and an expert. Discussing linguistic meaning, Putnam used this example (among others) to argue that novices who do not know the appropriate criteria for identifying items must defer to experts to determine the meanings of the associated words – and thus that meaning transcends what occurs in an individual mind. My account of concepts concerns what occurs in individual minds. I urge that someone who makes no distinction between elms and beeches, or acknowledges that there is a difference but cannot recognize the two kinds of trees, or makes a distinction that differs from that of the expert, may still have a concept in mind, although one that differs from the expert's. If the novice confuses elms and beeches, but does not confuse these with pines or rocks, then some ICs are operating. Still, the various concepts are not *just different*. For biological purposes the expert's criteria are superior since they track similarities and differences not captured in the novice's criteria. These differences may even be important for the novice – for example, if she has a diseased tree in her yard and is seeking an effective treatment.

Note one further point. I have argued that our various conceptual systems are not completely isolated from each other, so that changes in one system can generate unanticipated changes in other systems. This point encompasses both ICs and implications. But this is just part of the process of developing knowledge. I will return to questions concerning the interaction of conceptual systems later in this chapter when I discuss the relation between conceptual systems and theories, and the individuation of conceptual systems.

## 5.4 Systemic Role

I am now going to propose a significant addition to Sellars' theory of concepts, although it is in the Sellarsian spirit and even hinted at by Sellars. I have argued, along with Sellars, that we generate our concepts and change or replace them as our experience and understanding develop. As a result, we should expect that each of our concepts has a role to play in our cognitive

economy (EPM 163, cf. SM 95, 128); an account of that role must be included in an analysis of the concept. I will refer to the role a concept plays as its *systemic role*; this provides a third dimension on which concepts can be analyzed and conceptual systems altered. Before developing this thesis I want to comment on my choice of terminology.

The Sellarsian theory of concepts can be viewed as a "conceptual role" theory, where the notion of a conceptual role is cashed out in his accounts of implications and transitions. The additional element I am introducing generates a minor difficulty concerning an appropriate label. I would be asking for misunderstanding if I were to use "conceptual role" for this purpose, even though that is a natural choice. Another reasonable candidate is "function," but this would also be confusing because some descriptive concepts include a function in their content (e.g., UMBRELLA). Analyses of functional concepts must include an account of this function, which would appear among the implications. In the sense that now concerns me, all concepts have a function, independently of whether they describe a function. I hope to avoid some confusion by using "systemic role."

The thesis that we create concepts to do specific jobs is clearly in the Sellarsian spirit. Many passages in Sellars suggest the extension I am proposing. Here is a fairly clear example: "A living language is a system of elements which play many different types of roles . . . "(LTC 513).[4] Another example occurs in Sellars' account of the how THOUGHT and SENSE IMPRESSION could be introduced into a behavioristic language: He holds that these concepts take on a "reporting role," and contrasts a reporting role with a theoretical role (EPM 189). Elsewhere, discussing Ramsey's proposal that the theoretical terms occurring in a set of axioms be replaced by variables, Sellars objects that, "Their role is not that of being substituted for or quantified over, but that of being available for connection with extra-linguistic fact" (SRII 162). Further hints include: The claim that words for thing-kinds should not be treated as analyzable into a conjunction of properties because thing-kind words "have quite a different role in discourse from that of expressions for properties . . . " (CDCM 259); and "A primitive predicate of a theory is meaningful if it does its theoretical job . . . " (P 104). In the case of referring expressions, "their sense is, at bottom, their job, and their job is to be linguistic representatives of objects" (SM 124). Sellars also notes that some concepts, but not all, play an explanatory role (CDCM 260–61), and distinguishes between terms playing a rhetorical and an adjectival role (GE 260–61).

These passages hint at the theme I want to introduce, although Sellars never develops the idea in the direction I will take. Considerations of systemic role become vital in a reflective context. Some concepts describe physical objects while others describe properties of physical objects, and this is a significant distinction. In a similar way, the distinction between observational and theoretical concepts is a vital feature of the logical-empiricist conceptual framework; no such distinction occurs in the Aristotelian framework.

Indeed, the absence of concepts that have a theoretical role (in the logical-empiricist sense) is an important feature of Aristotelian thought. Consideration of systemic role is especially important when we examine conceptual change since such changes may involve the introduction of a new systemic role or the elimination of a previously familiar role. Introduction of ISOTOPE involved the creation of a role that did not exist in the prior system of chemical concepts: describing varieties of a single chemical element that have different atomic weights. In fact, this role was precluded by the prior framework in which a unique weight is characteristic of each element. COMPLEX NUMBER plays, among others, the role of providing solutions to equations that, at an earlier stage in the development of mathematics, were viewed as having no solutions. On the other hand, one result of abandoning Aristotelian physics was elimination of the systemic role played by NATURAL PLACE. A more complex situation can be illustrated by some of the changes in the concept associated with "earth" as we move from Aristotelian to Copernican astronomy. In later astronomy EARTH no longer has the role of describing a unique object around which all other celestial bodies move. Instead, EARTH shares a role with other concepts that describe planets in our solar system. This case differs from NATURAL PLACE because EARTH has the same referent in both systems of astronomy. Thus we have a change in both implications and systemic role while the key IC – identification of the body on which we live – remains unchanged. Considerations of systemic role thus provide an additional dimension on which concepts can be compared.

I want to consider one much debated example to illustrate the importance of examining all three dimensions when comparing concepts across a scientific revolution: the relation between SPACE, TIME, and MASS in classical mechanics (CM) and special relativity (SR). How we deal with this question depends on our theory of conceptual content; one source of dispute is that different philosophers adopt different views. Two views have been prominent. On one side of the dispute, Kuhn and Feyerabend emphasize the differing implications associated with these concepts in the two systems. Here we do find radical differences: for example, in SR space and time are related in ways that make no sense in CM, while mass has a relation to velocity in SR that CM excludes. Other philosophers maintain that for scientific purposes the extensions of concepts are of central importance, and conclude that the change has been much less drastic (e.g., Kitcher 1978, Sankey 1994, Scheffler 1967; Field 1973 adopts an intermediate position). And, indeed, differences in the extensions of these concepts are less drastic – a point that shows up clearly when we compare the procedures for determining these parameters in the two frameworks. Spatial and temporal gaps are measured by rods and clocks, respectively, in both CM and SR, even though different inferences are allowed once we plug the measured values into the two theories. The rejection of simultaneity at a distance in SR places constraints on measurements of space and time that are not required by CM, although measurements that meet the relativistic constraints are classically acceptable. In addition,

measurements of mass made in a body's relativistic rest frame are acceptable as values of the mass in CM, although there are ways of determining mass that would be acceptable in CM but not acceptable as measures of rest mass in SR (e.g., measurement involving accelerating bodies).

I submit that these two lines of conceptual analysis are aspects of a more complete approach that includes one more part, consideration of systemic roles, which throws further light on these concepts. In both CM and SR mass provides a measure of resistance to acceleration; the need for this role and a concept that describes it was not recognized before Newton. In both theories the concepts of space and time continue to mark different kinds of gaps between physical items. Let us pursue this last point (see Sec. 10.4 for additional details). Calculation of the invariant spacetime interval in SR requires combining measurements of spatial and temporal distances into a single parameter, and the use of different procedures for measuring the two kinds of gap is reflected in their being expressed in different units. Combining the two into a single value requires that the units of one of these be converted. Standard practice is to convert units of time to units of distance: The time value is multiplied by the velocity of light in a vacuum, $c$, which is the appropriate conversion factor. Yet even after this conversion has been carried out, the spatial and temporal terms have different signs, so that space and time are not completely merged in SR; distinctions remain that embody a significant similarity with their roles in CM. Maintaining this distinction is crucial to SR. It determines the geometry of Minkowski spacetime, and is necessary for the key results that the spacetime interval is always zero on a light ray, and that the velocity of a light ray is the same whatever the observer's state of motion. Moreover, $c$ is not *just* a conversion factor. It has a pervasive and fundamental role in SR that it does not have in CM, where it is one physical quantity among many. This new role is reflected in a large variety of new intra-systemic relations, although the value of $c$, and the means by which it is measured, need not change across the revolution.

These examples indicate how SYSTEMIC ROLE functions in comparing conceptual systems.[5] An example from the realm of formal concepts will further underline the importance of considering systemic roles. As Sellars notes, one reason for considering INTUITIONISTIC NEGATION to be a kind of negation is that it plays the same role in intuitionistic logic that classical negation plays in classical logic (CC 90). Differences between the two negation concepts appear as differences in their implications. As the example indicates, two concepts can have the same role in different conceptual systems even though there is no one-one mapping of their implications (or, where relevant, extra-systemic relations). On the other hand, differences in systemic role require differences in either implications, ICs or both. Often differences between concepts on any of these dimensions will involve differences along other dimensions. We will encounter several examples of the importance of systemic roles when I develop some detailed conceptual analyses and studies of conceptual change in Chs 7–10. Note, in addition, that

specific conceptual systems are also introduced for a reason, and play a role in our overall conceptual economy. Sometimes understanding the systemic role of a concept will require understanding the role of the system in which it occurs.

I am proposing, then, that conceptual content is constituted along three dimensions. Two of these – implications and systemic roles – are relevant for all concepts; the third – extra-systemic relations – is relevant for all except formal concepts.

## 5.5 Prescriptive Concepts II

Moral and aesthetic concepts are the most familiar prescriptive concepts, but (for the most part) I will not enter into these complex topics. My concerns in this book are mainly epistemic, and this realm provides a rich array of prescriptive concepts: VALID (in deductive logic), EMPERICALLY WELL SUPPORTED, RATIONAL, and TRUE, among others. Recall from Sec. 4.4 that many such concepts have both descriptive and prescriptive aspects. For example, to describe an argument as valid is to say that a specific relation holds between its premises and conclusion, and also to express an evaluation of that argument. We saw that we must eliminate the tendencies to behave that Sellars incorporates into his account of prescriptive concepts; we must also integrate systemic role into our account.

Our previous discussion indicates the appropriate replacement for DTs: We must include a description of a required act in the content of the concept; actually being primed to act is not relevant. An injunction to act in a particular way directs us to action in an extra-systemic domain and thus provides the required tie to that domain. Moreover, once we have eliminated DTs as *the* way in which a prescriptive concept is tied to the extra-systemic world, we lose any motivation for reducing all prescriptive concepts to a single type. We are thus free to explore different kinds of norms on their own terms. I will illustrate the impact of these changes by considering some central normative concepts from epistemology; I return to this topic at greater length in Ch. 8.

Some prescriptive concepts specify ends. One major epistemic end is the acquisition of propositional knowledge. The exact conditions for propositional knowledge are currently a matter of some dispute, but this much is widely acknowledged: propositional knowledge requires a justified belief in a true proposition (JTB). Since Gettier (1963) much debate has focused on whether justification and truth are sufficient to elevate a belief to knowledge, and if not, what additional condition or conditions are needed. I will leave any further conditions aside for now and focus on KNOWLEDGE, JUSTIFICATION, and TRUTH. Note that there is an important asymmetry between knowledge, on the one hand, and justification and truth on the other hand: We can pursue knowledge by seeking justified true beliefs, but we do not have the option of seeking, say, justified beliefs by pursuing knowledge. A

contrast will bring out the point. In Euclidean geometry a triangle is equilateral if and only if it is equiangular, and both properties are equally accessible. As a result, we can assess whether a triangle is equilateral by measuring either the angles or the sides, and we can construct an equilateral triangle by constructing either appropriate angles or sides. Now one step in the pursuit of knowledge is to pursue justified beliefs, but we cannot reverse the procedure and pursue justified beliefs by pursuing knowledge directly.[6] Moreover, we can pursue justification without pursuing knowledge, but we cannot pursue knowledge without pursuing justification. From this perspective knowledge is an end only; justification is both an end and a means for the pursuit of another end. Thus KNOWLEDGE and JUSTIFICATION are different kinds of prescriptive concepts and part of this difference is captured in the different ways in which they point to an extra-systemic domain.

Since the appropriate domain is determined by the proposition in question, a feature of Sellars' account of "ought" – what we ought to do is determined by the specific proposition to which *ought* is attached – carries over to knowledge and justification. Moreover, the criteria for justification will depend on the domain. For example, criteria for justification in a deductive subject, such as mathematics, are different from those in an empirical subject, such as physics. But however these criteria vary, when we consider the prescriptive side of epistemic concepts we find that they direct our attention to some domain that is (usually) distinct from our epistemic conceptual system. Different prescriptive concepts establish this tie in different ways, and giving an account of this tie is part of an account of that concept.

Another complexity arises when we consider the role of truth in the pursuit of knowledge. In some cases it is not especially difficult to establish that a proposition is true – for example, short tautologies and that I have a headache. But in many important cases we can neither pursue nor assess truth directly. In such cases our reasons for holding that $p$ is true are just the reasons we have for holding that we are justified in believing $p$. Again we find that justification is at the center of the pursuit of knowledge: for the most part, we pursue truth by pursuing justification. For my purposes in the present section the only point I want to make is that when we focus on the extra-systemic ties involved in our epistemic concepts we find complexities that are not captured in the usual definitions. The reason for this is clear: the customary definitions state intra-systemic relations among concepts, but on the approach I am advocating these relations provide only one of the dimensions along which concepts must be studied.

I turn next to the place of *systemic role* in our account of prescriptive concepts; requirements, prohibitions, permissions, and ends are all instances of prescriptive roles. KNOWLEDGE, for example, specifies an especially desirable epistemic state. We can see the importance of this role by noting that it is embedded in Gettier's challenge to the view that JTB is sufficient for

knowledge. Gettier challenged this view by constructing examples in which we achieve JTB but it is intuitively clear that we do not have knowledge. In a typical case (adapted from Gettier 1963), Smith believes that Jones owns a Ford on the basis of strong evidence. (Perhaps Smith knows that Jones has always driven a Ford in the past and recently saw Jones come out of a Ford on the driver's side.) Smith has been studying logic, and notes that the proposition "Jones own a Ford" entails "Either Jones owns a Ford or Brown is in Barcelona," and Smith believes this proposition too; call it *p*. Although Smith has no idea where Brown is, the belief in *p* is justified. However, Jones is driving a rented car and does not currently own a Ford, but Brown is in Barcelona; so *p* is true. Smith has met the JTB requirement, but it seems clear that Smith does not *know p*.

A key feature of this case, and many others that have been discussed, is the loose connection between *p* and its justification. It is easy to imagine ways in which the believer could have a much better justification for the same true belief. So this justification does not yield knowledge because it does not yield an especially desirable epistemic state. This diagnosis is implicit in two of the main types of response to Gettier examples. One approach challenges the claim for justification in Gettier cases and seeks to tighten up the conditions for justification; the other approach lets justification stand, but seeks an additional condition for knowledge besides truth and justification. Both approaches thus take off from the recognition knowledge requires more than is provided in the usual examples. Attending to the role that KNOWLEDGE plays among our epistemic concepts makes clear why the grounds for belief provided in Gettier cases are not sufficient for knowledge.

I want to introduce an additional point about prescriptive concepts by considering a bit further the thesis that we need to tighten up the standards of justification. Gettier cases involve inductive justification and it is well-known that no adequate account of induction is currently available. How should we go about seeking an account? One response, typical of analytic philosophers, is that we already have the appropriate concepts, but need a better analysis of these concepts. I suggest that this approach is no more plausible than it would be to hold that "we" always had the concepts of a derivative and an integral, so that the long struggle that began some time before the seminal work of Leibniz and Newton, but was not completed until the late nineteenth century, involved the clarification of these available concepts. Rather, I submit, the process was one of constructing better concepts, always building on available concepts and modifying some features while keeping other features constant. In the same way, a solution to "the problem of induction" requires the creative construction of new concepts. Those who propose approaches to induction are typically involved in just this creative endeavor, even when a prevailing philosophical ethos leads them to describe their results as analyses of what we all always knew (cf. EPM 195). As this research continues we can expect proposals that may involve changes along any of the three dimensions on which the conceptual content

of prescriptive concepts is constituted. The following passage (parts of which have already been quoted) will underline the extent to which I am still working in the Sellarsian spirit:

> In a perfectly legitimate sense one language can change into another even though the noises and shapes employed remain the same. Indeed, modern man is not only constantly introducing new symbols governed by new rules, he is constantly changing the rules according to which old symbols are used. Thus, as science has progressed, the word "mass" as a class of visual and auditory events has remained, but the rules according to which it is used in the language of science have changed several times, and, strictly speaking, it is a new symbol with each change in rules, though each new implicit definition (conformation rule) has had enough in common with earlier implicit definitions so that the use of the same symbol has not seemed inappropriate. Indeed, the scientist in different contexts uses the term in different senses, according to different rules. In common sense contexts his language is of ancient vintage. Thus we can stick to English and yet be said to speak not one language but many.
>
> (LRB 311–12)

Sellars is discussing descriptive concepts in this passage but the point extends to prescriptive concepts as well.

## 5.6 Models, Analogies, and Conceptual Change II

In Sec. 4.5 I discussed Sellars' account of analogy and its role in comparing concepts and introducing new concepts. I am now ready to integrate that discussion into our modified Sellarsian framework. Conceptual *systems* should provide the main focus of such discussions since concepts occur in systems and it is these systems that are modified and compared. Conceptual systems are constituted along three dimensions: intra-systemic relations, systemic roles, and (except for formal systems) extra-systemic relations. Analogies consist of comparisons of conceptual systems along each of these dimensions. Sometimes we are interested in comparing specific concepts taken from two conceptual systems, but considerations of implications and conceptual roles will involve comparisons of entire systems. Such comparisons can be developed for successive conceptual systems in a science, alternative logics, the conceptual frameworks of different philosophers or cultures, and more. Strictly speaking, when we compare concepts from different systems we consider different concepts, although we often find that two concepts are sufficiently similar to justify viewing them as counterpart concepts. The theory of concepts I am proposing provides the tools for carrying out such comparisons. Comparisons of conceptual systems require at least a partial analysis of each system, and our theory of concepts

provides an account of what a conceptual analysis should include. This last point applies even when we engage in conceptual analysis without the aim of comparing conceptual systems.

Now consider the introduction of new concepts. No theory of concepts can replace the insight and creativity required for fruitful innovations, but an understanding of the sources of conceptual content can provide a general guide to the kinds of changes that are available, and an indication of how one can communicate proposed changes to others. The key point is that when we seek to introduce new concepts we begin with an existing system and produce an analogous system by making specific changes. Thus *conceptual systems are the models for new conceptual systems*. A new system will be similar to and different from its model in ways that can be explicitly specified. Moreover, once a new system has been constructed analyses of concepts in that system does not require any reference back to the model. To be clear on this point it is important to distinguish two different issues that Sellars tends to conflate: the *process* by which a new concept is introduced, and the *analysis* of the conceptual system that results from this process. The process of introducing new concepts proceeds by analogy and requires a previously existing conceptual system that serves as a model. But once a new system has been constructed, analysis proceeds by mapping out systemic roles, intra-systemic relations, and (where appropriate) extra-systemic relations. To be sure, we might find it useful to keep the analogies in mind both as a heuristic for developing the analysis and as a means of saving effort since many features of the model may have already been examined and given results that carry over to the new system. We may also find it useful to revert to the model in teaching the new system to others. But these pragmatic considerations do not alter the status of the new system as an autonomous system to be analyzed and evaluated in its own terms.

In Sec. 4.2.3 I discussed Sellars' distinction between conceptual systems in ongoing and completed science, and his consequent doctrine that the models we use to introduce new concepts provide a permanent part of the content of those concepts in ongoing science. We also saw that for Sellars all models ultimately take us back to observation concepts. So Sellars' point in holding that we cannot abandon our models in ongoing science is just his claim that observation concepts play a fundamental methodological role in ongoing science. Thus he insists that,

> There is a core of truth in the concept of "*the* observation framework" and, indeed, of the abstractionist approach to basic empirical concepts which survives the exorcizing of givenness. . . . [T]o reject the myth of the given is not to commit oneself to the idea that knowledge as it is now constituted has no rock bottom level of observation predicates proper. It is to commit oneself rather to the idea that even if it does have a rock bottom level, it is *still* in principle replaceable by another conceptual framework in which these predicates do not, *strictly speaking*, occur. It is

> in this sense, and in this sense *only* that I have rejected the dogma of givenness with respect to observation predicates.
>
> (SRII 187)

We have seen that Sellars does not provide much of an argument for the claim that we must, for now, hold onto the observation framework. I suggest that the root of this doctrine lies in the role that items available to our senses play in ETs; having abandoned ETs, we need not accept Sellars' view of the quasi-permanent role of models. Instead, we can recognize that although the introduction of new concepts requires prior concepts that serve as a model, once we have mastered the new concepts we can dispense with the model; the new system can become autonomous and its model can be forgotten.

## 5.7 Conceptual Systems and Theories

Thus far I have avoided the question (on which Sellars provides no help) of the relation between conceptual systems and theories. The term "theory" is used in many ways, some quite loose, and I will not survey even a significant subset of these usages. It will, however, be helpful to distinguish descriptive theories, which offer an account of some domain, from prescriptive theories which address issues such as what we ought to do and what ends we ought to pursue; we will also have to keep in mind that some theories have both descriptive and prescriptive aspects.

### *5.7.1 Descriptive Theories*

I will focus this discussion mainly on scientific theories; one major function of many theories is to describe entities, interactions, and processes that occur in a particular domain.[7] Any theory that has this descriptive function will embody a descriptive conceptual system. There are, however, two reasons for distinguishing between a theory and its conceptual system. First, we can construct and explore a conceptual system without considering it as a description of any actual domain. In practice this is rare outside of mathematics, but the fact that it can occur underlines the distinction between the content of a conceptual system and the claim that this system is instantiated in a specific domain. I will use "theory" to describe those cases in which it is asserted that a particular conceptual system (or set of conceptual systems) is instantiated in a given domain.[8] Indeed, I have been using "theory" in this sense throughout this book – in particular when describing my own project as developing a theory of conceptual content that applies to human cognition. The case studies at the basis of my account provide evidence that the theory does apply. Evidence that human concepts are not structured in the way I propose would count against this theory.

The second reason for the distinction arises because scientific theories often deploy multiple conceptual systems. Most scientists do research in limited domains that directly concern fairly narrow aspects of nature. But they do so in the context of wider theories that also apply. As a result, the conceptual system of the wider theory is assumed, and adoption of a theory for a specific domain includes adoption of any wider theories that are used. Newtonian mechanics, for example, provides a framework for analyzing physical processes, and was long assumed in studies of planetary orbits, terrestrial projectiles, fluids, and more. In a similar way, basic quantum theory is accepted and applied in fields such as high-energy and solid-state physics, quantum chemistry, and detailed calculations of the structure of DNA molecules. All these specific fields include additional concepts that supplement those of the wider theory. A similar situation arises in everyday life. If I set out to buy a piece of furniture or a computer, I assume a variety of widely applicable conceptual systems in addition to those that concern the specific item of interest. These may include my concepts for dealing with mass-manufactured objects (which might not come into play if I were seeking a work of art), and my concepts for thinking about the purchase process, shipping, and perhaps buying on credit. Moreover, in all of these cases I assume the applicability of my concepts for thinking about physical objects. Thus if my newly purchased computer is not in the room where I left it I will assume it has been moved, rather than taking it for granted that computers sometimes disappear. But if I return to the room in which I recently heard a symphony, I will not expect the music to be lurking ready to be heard again.

The fact that many cases involve embedding a relatively narrow framework in a wider framework has important consequences – especially for science – because changes in any of these theories *may* generate changes in others. New evidence that challenges an accepted theory in a narrow domain may affect only the concepts that have been introduced for that domain. For example, the newly-formed consensus that neutrinos have mass impacts several fields that deal with neutrinos, but raises no present challenges for relativity or quantum theory. Indeed, these wider theories play a central role in assessing the significance of the new view of neutrinos. But in other cases developments in a narrow domain can have a major impact on a wider theory that is implicated in the research – as happened with the impact of the orbit of Mercury on Newtonian gravitational theory. In addition, changes in a wider theory can generate changes in narrower domains that fall under that theory. Thus the changes in kinematics that were central to Einstein's first paper on SR led directly to new formulas for the Döppler effect and the aberration of light – changes that were not motivated by any empirical evidence from these domains. Exactly how a particular case will play out depends on the details of the way that the various theories interact. When change affects only a subset of the conceptual systems used in a specialized body of research, those systems that are not affected provide one source of continuity

through the transformation. On the other hand, when work in a specialized domain leads to revisions of a widely used theory, the results can have substantial impact on views in domains that are quite far removed from that in which the problem arose.

### *5.7.2 Prescriptive Theories*

I will focus again on prescriptive epistemological theories that specify the appropriate goals of our cognitive endeavors and provide norms indicating how we should pursue those goals. In this case I will distinguish between the content of a system of prescriptive concepts, and the injunction that we should implement this system in our epistemic endeavors. Here too we find different conceptual systems that specify different epistemic ends and recommend different norms for the pursuit of those ends.[9] The guiding thesis of the following discussion is familiar from work in naturalistic epistemology: We must learn the appropriate epistemic ends for beings with our cognitive abilities functioning in this world, and we must learn what means promote the pursuit of our ends. As a result, our conception of our epistemic ends and means is subject to revision as our understanding of the world and ourselves changes. Consider some examples.

For Descartes there is a clear epistemic ideal: omniscience. Infinite epistemic agents achieve this ideal, and do so without needing a methodology for constructing knowledge. Omniscience, however, is not an appropriate end for *us* because we are finite cognitive agents subject to error. We can achieve the highest possible quality of knowledge – certain knowledge of true propositions – in limited domains,[10] but even in these domains we are subject to error and thus need a methodology if we are to achieve our end. The claim that we are finite, fallible agents is a descriptive claim about human cognitive abilities, and Descartes recognizes that it provides constraints on the epistemic ends we should pursue, and the means by which we should pursue them. Moreover, his methodology of clear and distinct ideas depends on additional presumed facts about us: that we have the ability to analyze ideas into their simple components, intuitively grasp connections among these components, and withhold belief until the process of analysis has been completed. If we lack these abilities, we will not be able to pursue knowledge by the means that Descartes specifies. Descartes' theory of knowledge also requires the existence of innate ideas – another descriptive claim about our minds. If we lack such ideas, his prescriptions about how to pursue our cognitive ends will fail.

Consider another example. For most of the history of physics the discovery of causal relations has been a major goal, but this goal has been challenged by quantum theory, at least under a widely accepted interpretation which holds that, at the most fundamental level, nature is irreducibly statistical. If this theory-*cum*-interpretation is correct, then we should abandon this goal (at least in the quantum domain) because there are no

such causal relations to be found, and replace it with the different goal of seeking statistical laws. This alternative goal requires a different methodology from that appropriate to the pursuit of traditional causal laws.

A more recent discovery – deterministic chaos – impacts another traditional goal of research in physics: prediction of states of a system into the unlimited future. We now have reasons for believing that there are physical systems whose future states are rigorously determined, but cannot be predicted. This discovery does not entail a failure of *science* in these fields – the existence of deterministic chaos is a scientific discovery. Instead, this discovery indicates a need to rethink the appropriate ends of physical research in certain cases. We must also give up one traditional form of experimental research in these cases: keep all features of the relevant system constant except one which we vary, assuming that the magnitude of any changes in the system will be proportional to the magnitude of the change we have introduced. It is a characteristic feature of chaotic (and other nonlinear) systems that this proportionality does not always hold. For some inputs we may suddenly find a drastic, even a qualitative, change in the output.[11]

The last example underlines another complexity. In cases of deterministic chaos the long-term behavior of a system is determined, but unpredictable; we should abandon attempts to make such predictions because no methodology will yield them. In other words, lack of an appropriate methodology can provide a reason for abandoning an end – either temporarily or, as in the present case, permanently because an appropriate methodology is unavailable in principle. Arguments that we should reject an epistemic end because we lack any means of pursuing that end occur in other places as well. For example, Laudan (1984, Ch. 5) argues that we should reject the discovery of fundamental truths about the world as an aim of science because we have no means of assessing whether we have succeeded, or even whether we are making progress towards that end. Other value issues also enter our choice of epistemic ends and means. Popper, for example, held that scientists should seek deep truths about the world, but recognized that this quest is inherently risky, and considered the risks worth taking. Van Fraassen, on the other hand, urges the avoidance of such risks and the pursuit of relatively safe results (see Hooker 1985 for this analysis).

The upshot of this discussion is that there are alternative systems of normative epistemological concepts, and that at least some of these alternative systems develop out of our changing understanding of our epistemic situation. A prescriptive *theory* in this domain consists of the proposal that we adopt a particular system of such concepts as the basis for epistemic pursuits. I suggest that this approach also applies to other normative disciplines, although I will not argue the case here. Competing moral theories or aesthetic theories, for example, should be viewed as alternative systems of prescriptive concepts along with an injunction to adopt a particular one of these. Our theory of concepts provides the tools we need to analyze these

systems, where such analysis is surely a desirable step in deciding which we should adopt. The discussion also suggests that assessment of whether we should adopt a prescriptive framework should be carried out in terms of our best understanding of the relevant domain. A normative philosophy of art that makes no contact with anything that has counted as art in human life would be beside the point; a moral theory must speak to actual human abilities and concerns if it is to be a moral theory for us. This last example is governed by the widely accepted principle that ought implies can, when this principle is used in conjunction with *modus tollens*.

## 5.8 Individuating Conceptual Systems

I turn now to an important question that I have postponed: What constitutes a single conceptual system? In philosophical jargon: How are conceptual systems individuated? Fodor presses the parallel question for specific *concepts* and considers it a major challenge to any holistic theory of concepts (e.g., 1995: 76; 1998: 37; Fodor and Lepore 1992: 21, 23–26). For any concept C, he argues, we must either provide a principled basis for deciding which concepts contribute to C's content and which do not, or face the result that every concept is implicated in every other concept. The latter option yields a massive holism that Fodor considers clearly unacceptable.[12] It might seem that local holism avoids this problem by restricting concepts to specific systems. But, Fodor would no doubt argue, we need a principled basis for determining what constitutes a single system; otherwise talk of distinct conceptual systems loses significance. While one could respond by providing criteria, I am going to take a different tack. Fodor's challenge derives from a particular philosophical program; I am going to challenge that program and defend an alternative.

I propose a shift of perspective that will be familiar to those who have followed developments in philosophy of science during the second half of the twentieth century. One characteristic feature of logical empiricism was its restriction of philosophical studies of science to the *context of justification* and thus to the *product* of scientific research. This product was viewed as a set of propositions that can be evaluated independently of any consideration of the way they were produced. A major theme in the restructuring of philosophy of science begun by the work of Hanson (1958), Kuhn (1962), Toulmin (1961) and others, was an insistence that research in the *context of discovery* is a necessary part of any philosophy of science. This approach studies science as a *process*, not just as a product, so that understanding how scientists work becomes a central topic in philosophy of science. The two perspectives (process and product) are complementary, not antithetical; both are required for an adequate philosophy of science.

I propose that a similar shift of perspective has an important role to play in understanding concepts. As I have emphasized throughout this book, our concepts are cognitive tools we create and use as guides to thought and

action. We adjust our concepts as we accumulate experience and reflect on that experience. Many of our concepts are open-ended. This is apparent for scientific concepts generated in the course of research, where we often do not know exactly what we will want to include in a classification, or what the basis of a classification ultimately will be. (Recall the recognition of this point in Carnap's doctrine of reduction sentences and in Sellars' remark (EAE 438, n. 10) quoted in Sec. 4.2.) Moreover, even where a concept does have a clear set of necessary and sufficient conditions, research may lead us to replace that concept with a similar concept having somewhat different necessary and sufficient conditions.[13] We saw (Sec. 2.2) that this kind of change occurs in mathematics, which is often taken as the paradigm of static concepts. The point also applies to logic, where the correct account of LOGICAL CONSEQUENCE, arguably the central concept of the field, is currently a topic of wide debate (Goble 2001 provides a recent survey). While some view the various proposals as attempts to analyze a pre-existent concept (e.g., Etchemendy 1990: 6–7), our discussion in this book suggest another approach: The concept of logical consequence that emerged in earlier studies has turned out to be inadequate as limitations of early logic have been recognized, more powerful formalisms have been developed, and reflection has continued. The problem, then, is to *construct* an appropriate concept that reflects the present state of our understanding. We may even ask whether a single concept will do the job (cf. Goble 2001: 7–8).

On the perspective I am proposing, a theory of concepts must treat the process of conceptual change as an equal partner with the analysis of established concepts. Recall Sellars' remark "that human discourse is discourse for *finding things out* as well as for expressing, in textbook style, what we already know" (CDCM 250), and his view of science as a self-correcting enterprise (EPM 170). This self-correction includes reconsideration of the concepts we are using, not just re-evaluations of the truth-values of propositions expressed in a stable system of concepts.

With these points in mind, I suggest that Fodor's demand for criteria that will determine the limits of conceptual content is plausible if we accept two GAs:

> The only proper way to study concepts is to consider them as established apart from their use by cognitive agents. (F1)

> There is (at least) a set of stable concepts that is widely shared by very large communities. (F2)

I have already given reasons why I think that each of these is false; I will have more to say about F2 in subsequent chapters. For the moment I want to emphasize that while there is an important role for analyses of established conceptual systems, these are best viewed as snapshots taken at a moment in time. We require such analyses when, for example, we study an historical conceptual system, or examine a current system with an eye to modifications

that we might make. To be sure, some conceptual systems remain largely unchanged in cultures for long periods of time. But we do not get an adequate understanding of concepts and their role in human knowledge by focusing just on these stable concepts. My sole objection to F1, then, is to the claim that studying static systems of concepts is the *only proper way* to study concepts. Thus I offer the following alternative GAs for the study of concepts.

> In addition to studying static snapshots of conceptual systems, we must also consider their dynamic aspect as cognitive tools we use for dealing with various domains, and modify as our understanding develops. (C1)

> It is an empirical question how widespread particular concepts are among humanity, although it is clear that many concepts are creations of specific cultures and sub-cultures. (C2)

Once we include this dynamic perspective in our approach, we not only find that conceptual systems change over time, but also that the boundaries of a conceptual system are flexible as we respond to particular concerns by drawing on various parts of an available conceptual repertoire.

This returns us to the question of individuation, which still must be addressed. I will limit discussion to individuation of conceptual systems because if we can individuate these, then we can accept with equanimity the notion that every concept within a system is implicated in the content of every other concept in that system.[14] But instead of seeking a static principle of individuation, I will focus on the use of concepts by intelligent adults, where following explicit criteria is not a necessary condition for responsible cognitive behavior (cf. Brown 1988; 1992b; 1994b; 2000c; Sankey 1997; Stark 1995). When we are trying to understand some aspect of the world, or act effectively in a particular situation, we rely on our conceptual systems for guidance. But we often combine available systems in ways that make an appropriate response to the question of what constitutes a single system highly contextual. Many factors in our historical and social situations lead us to view different systems of concepts as distinct, but productive thinking can also lead us to combine some of these systems, to subdivide older systems, or to mix and match them in various ways. Doing this in a responsible manner requires skill and judgment, and maintaining appropriate distinctions between conceptual systems depends on this skill and judgment. To be sure, there are individuals who do not make appropriate distinctions, but allow every subject and every conceptual system to flow into every other. Often this is a failure of epistemic responsibility that shows up in the results of their thought and action; it would not be eliminated even if we could formulate explicit criteria for delimiting conceptual systems. I am urging that it is *not important* to be able to say in general *exactly* what constitutes a specific conceptual system.

What is important is that in using or studying a system in a particular context we be clear on the reasons for making the combinations and distinctions that we make.

My thesis goes sufficiently against the grain of many philosophers that it may be worthwhile comparing a similar approach from a related context: Giere's discussion of theory individuation in his account of the semantic view of theories. This is a formal account of theories that differs from that of the logical empiricists in a key respect: Given a formalism, the semantic view focuses on interpretations of that formalism, rather than on the language in which the theory is stated. In particular, the semantic view focuses on interpretations in which a formalism is true, which are called *models* in logic. Thus a formalism is viewed as a means of specifying a set of models, where different formalisms may use different languages to pick out the same models. We move to a scientific theory when we add the hypothesis that a model provides a description of some aspect of the world. Now, when we look at actual applications, we often find that several models of a formalism are connected to different domains or aspects of a domain. As a result, Giere suggests that "we understand a theory as comprising two elements: (1) a population of models, and (2) various hypotheses linking those models with systems in the real world" (1988: 85). Suppose we attempt to individuate *Newtonian mechanics* from this perspective. Giere notes that Newton's second law ($F = ma$) is a centerpiece of the application of Newtonian mechanics, but its use requires that we specify a force function. Different force functions are appropriate in different contexts, which generates different detailed versions of Newtonian mechanics. Other differences occur because of considerations of the level of idealization needed for a specific problem. Reflecting on this situation, Giere concludes that "a scientific theory turns out not to be a well-defined entity. That is, no necessary and sufficient conditions determine which models or which hypotheses are part of the theory" (1988: 86). What counts as a particular theory is a contextual matter, with the detailed choice motivated by the specific problem at hand, the degree of accuracy required, and even how much time one has to devote to the problem. Which features to use in a given case is "solely a matter to be decided by the judgments of members of the scientific community at the time" (1988: 86). Being able to make such judgments is a major distinguishing feature of competent professionals in a field.[15]

There is also an historical dimension to this contextuality. For example, modern Newtonian mechanics integrates Newton's laws, calculus, and Euclidean geometry into a seamless whole. But Newton never wrote down the differential equations that are standard in modern treatments. We also use other parts of mathematics that were not available to Newton. For example, differentiation and integration are now understood as linear operators, which means that we can draw on results from linear algebra in mechanics – although linear algebra can also be pursued as a distinct mathematical subject applicable in many fields of mathematics and mathematical science.

These remarks raise a further question about individuation: When we bring together two (or more) conceptual systems for a particular task, do we create a new system, altering the content of all of the systems involved? Examples of such situations range widely. For example, in order to buy a painting online, we bring together concepts we uses to think about art and those we use to think about online purchases. There are many people whose repertoires include one of these systems without the other. In a similar way, a physics student may learn relativity and quantum theory as distinct subjects, and then learn to synthesize the two when studying Dirac's equation for the electron. In such cases new inferences are generated but, we want to ask, does this constitute a new conceptual system? In response, I suggest that we move away from considering such questions as if they are important for their own sake. Deciding whether and when we have a single system will not provide any insight into cognition, or enhance our ability to use our conceptual repertoires effectively. The insistence that we answer this question is a consequence of a philosophical GA that, I urge, we should abandon. Instead, we can map out the implications embodied in each system individually, along with the new implications generated when they are used together.

My thesis, then, is that conceptual systems are *our cognitive tools* that we use to further our aims. In doing so, we adopt and adapt tools from our current repertoire and develop new tools as new situations arise. What constitutes a single conceptual system will thus vary with the circumstances. Deeper understanding requires a better understanding of human cognitive abilities, a subject that was considered outside the pale of epistemology for much of the twentieth century, but is central to contemporary naturalistic epistemology (cf. Kitcher 1992). I emphasize that this is a subject of current research, and while we have much to learn, I think that one point is clear: We should not be looking for a set of permanent, context-free criteria for the use of conceptual systems. I will return to this topic in the final chapter.

## 5.9 Self-reference, Circularity, and Reflexive Consistency

Typical descriptive concepts describe items other than themselves, but there are exceptions – including the topic of this book since CONCEPT is a descriptive concept. Self-reference is a red flag for many philosophers because it is at the root of several paradoxes; thus there are important arguments for banning self-referential concepts and theories. Properly speaking, paradoxes are inconsistencies, not just surprising or "counter-intuitive" results. Thus in considering problems generated by self-reference I will limit myself to cases in which inconsistencies occur.

A second red flag arises because of the method I am using in this study. I am seeking a general theory of concepts, with an understanding of conceptual change as a major aim. Yet my procedure is to begin with examples of conceptual change and use these as the basis for the construction and initial

defense of my account. How, it can be asked, have I selected these initial examples? I must have been assuming some theory of concepts in making my selection. Since the examples serve as evidence for the theory, if I was assuming the theory I wish to defend – even tacitly – I may have generated a biased body of evidence. I will reply to the circularity objection first, and then consider self-reference.

The circularity objection that concerns me arises in the context of justification. A paradigmatic example of a circular justification would occur if I claimed that every proposition asserted by Jones is true, and then sought to justify this general principle on the grounds that I heard it from Jones. As we say, I am assuming the very claim that is to be established.[16] The problem is that an argument which purports to assess the justification for $p$, but assumes $p$, seems to be biased in favor of $p$. But this is not always the case. In one of the basic modes of inductive justification we derive consequences from a theory and test those consequences. As long as the possibility of false consequences is not precluded there is no circularity, and it is legitimate to take cases in which the consequences are true as providing inductive support. (See Brown 1993, 1994a and 1995 for examples and discussion.) Refutations by *reductio* provide a limiting case of this procedure. In general, merely noting that the claim being evaluated is assumed in the evaluation procedure is not sufficient to establish circularity; convicting a justificatory argument of circularity requires a detailed analysis that picks out the specific way in which the outcome is biased. In the present case I can give a specific reason why there is no circularity in my procedure. I have been working from a large and varied set of cases in which it is widely agreed that conceptual change has occurred; this agreement stands independently of my account of conceptual content. Thus it is not likely that I have chosen a biased set. In addition, these examples provide only an initial test of the theory; in Chs 7–10 I consider further cases that will also test the theory. So, in the absence of a specific demonstration of bias, I will not concern myself further with this objection.

Now consider the self-reference involved in my goal of developing a theory of concepts that applies to all concepts, including itself. Since self-reference generates paradoxes in some cases, elimination of self-reference resolves those paradoxes. But there is no demonstration that all self-referential claims yield paradox, and considerable evidence to the contrary. For example, we regularly discuss English grammar using grammatical English sentences without generating paradoxes. As a result, my first response to the supposed dangers of self-reference is of the same variety as my response in the case of circularity: I am not going to worry about this issue unless it is shown that a specific paradox arises. But there is more to be said, because there are situations in which self-reference is *mandatory*. This thesis was defended by Fitch (1946), who considered cases such as a fully general theory of knowledge.[17] If that theory is to be accepted into our body of knowledge, it must meet its own requirements. The force of this point is especially

apparent for theories that specify limits to knowledge: If the development and defense of such a theory requires cognitive resources beyond those allowed by the theory, then the theory's limits are too narrow, and the theory is defective. Thus considerations of self-reference yield a constraint on our theories: they must be *reflexively consistent*. As with all consistency conditions, this is only a necessary condition for adequacy. The fact that a theory meets this condition is not a reason for thinking that the theory is true; but a failure to meet this condition is a reason for rejecting that theory. I intend to take this constraint seriously and give an account of the concept of a concept. In accord with the approach to concepts I have been defending, this will be an extended narrative, not a more-or-less compact if-and-only-if statement.

## 5.10 The Concept of a Concept

CONCEPT is a descriptive concept and must be a member of a system of concepts.[18] To give an account of this concept I must specify: 1) Its systemic role, along with the role of the conceptual system to which it belongs in our overall cognitive economy; 2) Other concepts in the system and the major implications among these; 3) The instantiation conditions for this concept. I will take these up in the order just stated.

### *5.10.1 Systemic Role*

We can approach this topic by recalling why Jones, in Sellars' myth, introduced the concept THOUGHT. Reflecting on overt linguistic behavior Jones noted that "a person's verbal propensities and dispositions change during periods of silence as they would have changed if he had been engaged in specific sequences of various types of candid linguistic behaviour called 'thinkings-out-loud' by our Ryleians . . . " (SM 151). Jones introduced THOUGHT to describe internal processes in the speaker that occur during these periods of silence, and that explain the observed changes. THOUGHT thus has both a descriptive and an explanatory role: It describes an item that occurs in our inner life, an item postulated to explain a specific phenomenon.[19] The justification of this postulate rests on its explanatory power.

Unlike Jones, we need not start from a Ryleian perspective. We can note that the phenomenon that caught Jones' attention is only one of a number of features of our behavior which suggest the existence of internal processes affecting overt behavior. Much behavior indicates that we store information about items and draw on that information in many situations. We typically have definite expectations about how a familiar object is likely to behave and respond to our behavior; often we are able to give verbal descriptions of these expectations. When we encounter an unfamiliar item we often explore it in ways that yield information which then affects our future behavior

towards it. This exploration is largely guided by similarities that we note between this item and others we have experienced in the past. Noting similarities between items leads us to generate classifications. We also think about items that are not present. Such thinking may lead, among other possibilities, to plans for future dealings with an item or type of item, or to new expectations. All these phenomena make sense if there are inner items that embody our beliefs about the objects, processes, and situations we encounter. These inner items, which we call "concepts," represent items in that they embody beliefs about those items. We also associate specific concepts with a word or phrase so that an encounter with that word in its verbal, written, signed, or Braille form elicits much of the same information as does an encounter with the item that the concept represents. When we construct a system of concepts to describe and theorize about our inner activities, we include CONCEPT in that system. This concept describes a class of inner items that serve as loci of beliefs about items in the world. CONCEPT may be considered a psychological concept since it occurs in the system of concepts we use for describing certain aspects of our psychology. The primary roles of CONCEPT are, thus, descriptive and explanatory: it represents a type of item that occurs in our mental lives, and it plays a key role in explanations of our thought and action. CONCEPT also plays an epistemic role because the items it describes embody beliefs – including beliefs about our epistemic lives.

On reflection we find differences worth noting among our concepts, and this leads to the distinction between formal, descriptive, and prescriptive concepts. It is important to be clear that in the system of concepts we use for describing our mental lives DESCRIPTIVE CONCEPT, PRESCRIPTIVE CONCEPT, and FORMAL CONCEPT are all descriptive concepts; they describe different kinds of concepts that serve different functions. In giving an account of each type of concept I am elaborating part of the conceptual system to which these concepts belong. The roles of these concepts are particularly clear when we consider that these concepts are included in a theory that seeks to explain aspects of our overt and inner lives. We include PRESCRIPTIVE CONCEPT in this system because we recognize obligations, think about them, apply them to ourselves and others, and distinguish between just describing an item and acknowledging an obligation that involves the item. According to this theory, people have concepts independently of whether they have the concept of a concept – just as they have neurons and neuroses without having the corresponding concepts.

The conceptual system in which CONCEPT occurs, then, is a system we use in describing and thinking about our inner life and its impact on behavior. The set of roles we include in this system depends on the views we hold about our psychology; we introduce adjustments in this framework as these beliefs develop. An example will illustrate the point. From one traditional perspective the conceptual system I have been discussing concerns cognitive features of our psychology.[20] We have a distinct conceptual system that we use for describing and theorizing about emotions. But some are now challenging

this distinction (e.g., Damasio 1994), and arguing that (in my terminology) these two systems should be integrated in a way that involves (among other changes) implications that would not be allowed on the traditional approach. We have already encountered other cases of conceptual change that involve integrating conceptual systems that were previously considered distinct.

### *5.10.2 Intra-systemic Relations*

We can now consider two topics simultaneously: the implications that are constitutive of CONCEPT and – since implications involve other concepts in a system – some of the other concepts that will be included. The implications I will consider are clearly a function of the theory of concepts I am proposing; those who defend different theories of concepts will disagree on the contents of this system.

Since I follow Sellars in considering implicational relations to other concepts to be a constitutive feature of all concepts, the conceptual system we use for thinking about concepts will include IMPLICATION. In this system *any* proposition of the form "C is a concept" will imply that there are implicational relations between C and other concepts. By way of contrast, the conceptual framework of classical empiricism includes the concept SIMPLE IDEA, where "*s* is a simple idea" implies that there are no implicational relations between *s* and other simple ideas. IMPLICATION also occurs in the empiricist framework – it plays a central role in the distinction between simple and complex ideas – and one characteristic feature of that framework is the principle that simple ideas are not loci of implications.

We find other intra-systemic relations when we consider each of the types of concepts I have distinguished, since each will have its characteristic set. DESCRIPTIVE CONCEPT implies the concept of an extra-systemic domain which includes the items that this concept describes. In addition, "D is a descriptive concept" implies that there are criteria for determining whether D is instantiated in that domain. Our conceptual system for thinking about concepts also includes SELF-REFERENTIAL CONCEPT. Prescriptive concepts involve claims about what agents ought to do or avoid, so PRESCRIPTIVE CONCEPT implies OBLIGATION. Since obligations involve actions, ACTION is also implied; and since actions are typically directed at items other than our concepts, PRESCRIPTIVE CONCEPT implies the concept of extra-systemic items. The concept ACTION brings along the concept AGENT, which is not redundant because it is part of the concept of a prescription that prescriptions apply only to agents capable of carrying them out. This point is captured in Kant's principle that OUGHT implies CAN; thus the concept ABILITY is included in the conceptual system we are exploring. At least since Kant, philosophical thought about prescriptions has included a distinction between hypothetical and categorical prescriptions, yielding two more concepts in our system. We have seen that we also require PERMISSION in

order to distinguish cases in which we are required to carry out (or avoid) an action, from those in which particular actions (or non-actions) are allowed, but not required. MIXED CONCEPT carries all of the implications carried by DESCRIPTIVE CONCEPT and PRESCRIPTIVE CONCEPT, and does not require any additional concepts. Since formal concepts are completely constituted by implications, "F is a formal concept" implies that F implies other concepts, but has no implicational relations to the concept of an extra-systemic domain. In addition, SYSTEMIC ROLE is a member of the proposed conceptual system for describing concepts; each of the types of concepts I have discussed implies a systemic role for that type.

### *5.10.3 Extra-systemic Relations*

Since CONCEPT is a descriptive concept, we must examine the kind of evidence that would be required to assess whether CONCEPT is instantiated in its domain. The primary domain of this concept is human psychology, and I will approach the question by considering evidence that individuals have specific concepts, since possession of specific concepts implies possession of concepts. Those who have a particular descriptive concept in their active repertoire should yield evidence that they classify items in accord with that concept and draw appropriate inferences from the concept. In the case of prescriptive concepts we would also look for evidence that they recognize obligations. In the case of formal concepts we would seek evidence of inferences alone. Indeed, when we are dealing with people the task is almost trivial since the use of language makes it easy to provide the needed evidence. When people describe their principles of classification, or state what follows from a particular classification, or what they ought to do and their excuses for not doing it, we have excellent evidence that they possess the relevant concepts. People can also tell us what constitutes relevant evidence that a concept has instances. For example, we know that some people understand the criteria for determining if top quarks exist because these criteria have been formulated and published, and the relevant experiments carried out. Behavioral evidence can support, and sometimes substitute for, linguistic evidence, but this is rarely needed. Occasionally behavioral evidence can undermine verbal evidence by indicating that a person is just parroting a script but does not actually possess the concept in question. Linguistic evidence will have to suffice for concepts possessed by scholars or theorists, but not in the repertoire they actually use for dealing with the world.

We have, then, both a clear understanding of the evidence relevant to determining if individuals possess concepts, and excellent evidence that people do possess concepts that play a central role in their thinking and behavior. Still, I think it will be illuminating to pursue the topic one step further. CONCEPT was created for application to people, but there are continuing inquiries about the applicability of this concept to other animals. An examination of the criteria used in these studies will provide additional

insight into how we assess whether CONCEPT is instantiated. Henceforth, for brevity, I will use "animal" to refer only to non-humans. At the present stage of this book it should be clear that exploration of this question might lead to more interesting conclusions than just a decision as to whether animals do or do not possess concepts. For example, we might find that CONCEPT does not apply in a straightforward way, but that a modified version does apply; and this may lead to new insight into the domain we are exploring, as well as to a better understanding of our own thinking. Another possibility is that we might find that we are dealing with an open-ended concept for which there is no clear answer in this new domain – which might, in turn, lead us to consider modifying our concept. This is all very messy, but no more messy than human thinking typically is as we attempt to extend our current frameworks into new areas. I cannot explore all of the possibilities here, and I will place a narrow limitation on the following discussion. I am going to look at some of the literature on animal concepts solely with an eye to sharpening our understanding of the *criteria that are relevant for deciding if non-linguistic animals have concepts*. In doing so I will take no stand on the question of whether any of these animals possess concepts.[21]

Given our general reliance on behavioral evidence when dealing with animals, concept possession is discussed in concrete situations involving items that we, and presumably our animal subjects, can observe. Thus most research deals with concepts that describe observables. This is appropriate since it is reasonably clear that animals do not possess theoretical concepts such as PHLOGISTON or BOSON, or philosophical concepts such as TRANSCENDENTAL ARGUMENT – indeed, few humans possess these concepts. If animals possess descriptive concepts they must behave in ways that are reasonably interpretable as indicating that they classify in accordance with these concepts, and make inferences from them.

Consider classification first. The following remark, from a long-time advocate of the view that animals have significant mental abilities, occurs in the context of a discussion of whether animals have concepts:

> many animals react not to stereotyped patterns of stimulation but to *objects* that they recognize despite wide variation in the detailed sensations transmitted to the central nervous system . . . a Thompson's gazelle recognizes a lion when it sees one. The lion's image may subtend a large or small visual angle on the retina, and it may fall anywhere within a wide visual field; the gazelle may see only a part of the lion from any angle of view. Yet to an alert tommy, a lion is a lion whether seen side or head on, whether distant or close, standing still or walking.
>
> (Griffin 1994: 122)

This ability to recognize different instances of a particular type on the basis of widely varied cues can be taken as evidence either that the animal has a large repertoire of distinct behavioral responses, or as evidence that these

behaviors are generated by a more unified psychological entity – a concept that allows it to classify items on the basis of a variety of evidence. This evidence need not all be visual. Griffin adds: "the ability to abstract salient features from a complex pattern of stimulation, often involving more than one sense, requires a refined ability to sort and evaluate sensory information so that only particular combinations lead to the appropriate response." Thus Griffin holds that gazelles are able to classify stimuli and arrive at conclusions about objects in their environment, and considers this relevant evidence for showing that they have concepts.

Griffin also discusses an experiment (due to Herrnstein and Loveland) in which pigeons were shown colored slides, and were sometimes fed when they pecked at those containing pictures of people (the "positive pictures").

> The positive pictures might show men, women, or children; the human figure might be large or small, dressed in different sorts of clothing or engaged in a variety of activities, sitting, standing, walking, with or without other people or animals present. In some pictures only part of a human figure such as the face was included. The negative pictures varied just as widely.
>
> (1994: 128–29)

Once the pigeons were pecking positive pictures at a significantly higher rate than negative pictures they were tested on a completely new set of pictures:

> Surprisingly, some of the pigeons mastered this task and pecked significantly more at the new pictures containing people. It is important to appreciate that the pigeons do not perform perfectly in these tests; typically they may peck at perhaps 70 to 80 percent of the positive pictures and only 20 to 30 percent of the negatives. But the numbers of pictures used in such experiments are so great that the differences are extremely unlikely to occur by chance.
>
> (1994: 129)

Herrnstein termed this concept learning, for the pigeons had learned not specific pictures or patterns, but categories.

Griffin notes that many psychologists dispute the interpretation of these results as evidence that pigeons form concepts (1994: 132–34). If, for example, they just respond to a single common feature in the pictures, then they have not learned a concept. But this disagreement underlines the only point I wish to make: Evidence of classification based on a variety of stimuli – as opposed to a set response to a specific stimulus – provides evidence for concept possession.

Some researchers have sought ways around the limitations imposed by the absence of linguistic evidence. For example, Pepperberg (1991, 1999) works with an African grey parrot, Alex, that she taught to use a variety of words

in appropriate circumstances – including giving verbal responses to verbal questions. Pepperberg is interested in Alex's cognitive abilities, not in language *per se*: "the techniques developed in the communication programs enabled researchers to examine those cognitive (and not necessarily linguistic) abilities in animals that were not observable using the more traditional paradigms . . ." (1991: 157). Teaching an animal some language "enables researchers to query their animal subjects in as direct a manner as they now query human participants in related studies . . . " (1991: 158). Her research also allows us to extend our discussion to more abstract concepts such as COLOR, SHAPE, SAME, and DIFFERENT.

Alex was trained to recognize a number of different colors, shapes, materials, and quantities, and to answer questions of the form "What color *X*?" and so forth. Evidence that Alex understands COLOR and SHAPE is provided by tests in which he must name, say, the color of objects that have both color and shape – and must answer just the specific question asked, not give all possibly pertinent information. "The test employed was rather strong, for it actually involved reclassification of objects; that is, Alex was required to classify the same object with respect to color at one time and shape at another" (1991: 167). In addition, the test questions were randomly inserted into strings of other questions, so that the bird could not just give a series of color or shape replies. Whether Alex grasped SAME and DIFFERENT was tested in another experiment (with precautions along the lines just noted):

> Alex was to be presented with two objects that could differ with respect to three categories: color, shape, or material (e.g., a blue wooden pentagon and a rose rawhide pentagon; a yellow wooden triangle and a grey wooden triangle). He would then be queried "What's same?" or "What's different?" The correct response would be the label of the appropriate category – not the specific color, shape, or material marker – that represented the correct response (e.g., "color", not "yellow"). Therefore, to be correct, Alex would have to (a) attend to multiple aspects of two different objects; (b) determine, from a vocal question, whether the response was to be on the basis of similarity or difference; (c) determine, based on the exemplars, what was same or different (e.g., were they both blue, or triangular, or made of wood), and (d) produce, vocally, the label for this particular category. Thus, the task required, at some level, that Alex perform a feature analysis of the two objects: Correct responses could not be made on the basis of total physical similarity or difference of the objects. . . .
>
> (1991: 169–70)

Now consider some attempts to assess if animals make the kinds of *inferences* that are characteristic of concept possession. Griffin maintains that evidence of the formation of expectations is evidence of inference, and provides examples from the literature in which it seems that animals form

expectations. In one experiment (by Tinklepaugh) monkeys were trained by watching the experimenter place a piece of banana under one of two distinguishable cups. Situations were then created in which the cups were out of reach and out of sight until a barrier was removed; at this point the monkey was free to overturn a cup and retrieve the food. Once the monkeys had learned to select the correct cup every time, the experimenters sometimes replaced the banana with a piece of lettuce while the cup was behind the barrier.

> As Tinklepaugh described the results, the moderately hungry monkey now "rushes to the proper container and picks it up. She extends her hand to seize the food. But the hand drops to the floor without touching it. She looks at the lettuce but (unless very hungry) does not touch it. She looks around the cup . . . [ellipses due to Griffin] stands up and looks under and around her. She picks up the cup and examines it thoroughly inside and out. She has on occasion turned toward the observers present in the room and shrieked at them in apparent anger.
>
> (Griffin 1994: 121)

The suggestion, then, is that the animal's behavior indicates that an expectation about an unobserved item had been formed on the basis of observation, and that this expectation was disappointed.

Implications are the sole defining feature of formal concepts, so the only behavioral evidence for possession of a formal concept would be evidence of appropriate inferences. The formal concept TRANSITIVITY has been the subject of animal research. For example, Gillan

> taught chimpanzees that container E had more food than container D, D had more food than C, C more than B, and B more than A. He then tested individuals on novel pairs like BD, BE, and CE. The animals consistently chose the container in each pair that was associated with the greater amount of food. In this and other tests, it seems possible that subjects inferred the relation *greater than* and solved test problems according to this relational rule rather than according to the prior association of particular stimuli. . . .
>
> (Cheney and Seyfarth 1992: 83–84)

As the passage indicates, the result is far from conclusive. Boysen replicated this experiment and is equally cautious, concluding only that "chimpanzees may be capable of employing transitive inference to determine the correct choice between two nonadjacent items in an ordered series . . . " (1993: 50). Again I stress a single point: whether these problems are solved by inference, or by some other means, is an appropriate question to ask when considering if the animals possess the formal concept TRANSITIVITY.

Now consider prescriptive concepts. These are difficult to assess on the basis of behavior alone since acting in accordance with a norm does not show that the action resulted from obeying that norm. But researchers in the field have argued that certain kinds of behavior indicate a role for norms. A well-known case is provided by chimps who learned to wash food before eating it. This behavior could indicate that the chimp is obeying a hypothetical norm; this is considered significant because it involves some delay in gratification, a rare occurrence among chimps when food is involved. An experiment by Boysen (1993: 53–57) explores whether chimps are able to learn hypothetical norms, as well as the limits of their ability to obey such norms. Boysen trained several chimps to recognize numerals in the range from 0 to 8 (the range of competence varied among chimps in the study). The chimps were able to select the correct numeral corresponding to a number of items, select the correct number of items corresponding to a numeral, and pass other relevant tests. Boysen then used this capacity to explore whether one of these chimps could learn to behave in accordance with a hypothetical imperative. In the experiment it is taken for granted that when given a choice between two quantities of food, chimps prefer the larger. The subject chimp was paired with a partner and required to select one of two dishes with unequal amounts of candy; the dishes were out of reach when the choice was made. The operative rule was that the candy in the chosen dish was given to the partner while the subject received the candy in the other dish. Thus, to get the larger amount of candy the chimp had to learn to choose the dish with the smaller amount. The experiment was carried out under two different conditions: in one the actual candy was in the dishes; in the other, the candy was replaced with a card containing the numeral indicating the amount of candy in the dish. When the card was present, the percentage of cases in which the chimp pointed to the numeral representing the smaller number ranged from 67 percent to 87 percent in different trials; when the candy was present, the chimp pointed to the smaller quantity only around 17 percent of the time. These results suggest that chimps have some ability to grasp hypothetical imperatives – and thus some prescriptive concepts – although their ability to obey these imperatives is severely limited because it is easily overridden by prevailing conditions. We have, then, another example of the kind of behavioral evidence that is relevant to assessing whether an individual possesses a particular type of concept.

Let me underline the conclusion I want to draw from these animal studies. I take it as given that the experimenters possess the concept of a concept, and that the content of this concept plays a guiding role in designing experiments to assess whether various animals possess concepts. Since CONCEPT (in all its varieties) is a descriptive concept, part of its content consists of ICs. In effect, the experimenters recognize this point: When they are considering whether animals have concepts they look for evidence that the animals display the behavior that is appropriate to that

concept. It is vital that we keep two issues distinct in this discussion: having a concept and attributing that concept to an individual. The point is particularly clear in the case of formal concepts. Specific formal concepts do not include ICs in their content, but FORMAL CONCEPT is a descriptive concept that does include ICs. These ICs determine the kind of evidence that is relevant for assessing whether to attribute a formal concept to an animal. In this case we look only at evidence that the animal makes particular inferences; which inferences are relevant is determined by the specific formal concept in question. In the case of descriptive concepts individuals should also exhibit evidence of the appropriate ICs – here, evidence that they recognize instances of the concept. In the case of prescriptive concepts we seek evidence that they recognizes the required actions.

## 5.11 Summary and Conclusion

I have now arrived at a nearly final formulation of the theory of concepts I am proposing; some refinements will be introduced in Ch. 6. I will henceforth refer to this theory as TC. In this section I will summarize the theory in its present state and underline some of its virtues.

TC holds that all concepts occur as members of conceptual systems in which concepts are related to each other by implications. In addition, concepts are included in a system because they play some specific role or set of roles in our thought and action. While these two features apply to all concepts, there are four types of concepts that are distinguished by the ways they relate to items outside the system. Formal concepts have no such extra-systemic relations. Descriptive concepts are used to describe items in the various domains that concern us, and are partly constituted by their relation to that domain. This relation consists of a set of criteria for determining if the concept is instantiated in that domain. Prescriptive concepts guide action and thought about these actions; they are related to their domains by specifications of actions that we ought to carry out or avoid, and by specifications of actions that are permitted. Many prescriptive concepts are quasi-formal (e.g., OUGHT and PROMISE): they operate on a proposition which determines the specific behavior that is required. Others (e.g., STOPLIGHT) require a specific behavior on their own. The fourth type consists of concepts that are both descriptive and prescriptive, and have all the features of both types of concepts. TC is offered as an account of some central aspects of human thought, and includes the theoretical hypothesis that human thought makes use of concepts of these types. The multi-dimensional constitution of conceptual content has implications for the nature of conceptual analysis: An analysis will consist of an extended narrative, rather than the (more or less) succinct statements of conditions for concept possession that are typical in the analytic literature.

A different formulation may help focus the proposed account of conceptual content. There are three questions we should ask in trying to understand

a concept: a) What are its relations to other concepts? b) How is it related to an extra-systemic subject? c) Why have we included this concept in our repertoire? Often this last question will reduce to asking why have we made a particular distinction (e.g., weight/mass or natural/violent motion). The status of (b) is different for different concepts: it has no role in the case of formal concepts; for descriptive concepts it requires an account of the considerations needed for deciding if such items exist; for prescriptive concepts it requires an account of the kind of behavior that is mandated in a domain.

Consider some of the strengths of TC. First, since TC is holistic – in the limited, local sense I have discussed – it shares the advantages that generally accrue to holistic theories of concepts. Such theories eliminate problems concerning the identification of basic concepts, since no such concepts are required. The traditional problem of theoretical terms also vanishes since that problem arises from the thesis that all conceptual content ultimately derives from basic concepts that derive their content from sensory experience. Nor do relational and higher-order concepts pose a special problem. Most importantly, TC provides a basis for understanding how fundamental conceptual innovation can occur while maintaining sufficient continuity with older concepts to make the changes intelligible. Such innovation is a central feature of the development of human knowledge; it is the major phenomenon that I want to account for in this book, and is not examined in detail by many advocates of competing theories of concepts. In Ch. 6 I will consider the relative strengths and weaknesses of TC and some of its current competitors.

For TC, as for Sellars' theory, the now-controversial analytic-synthetic distinction does not play a fundamental role in an account of conceptual content, although the distinction need not be rejected as either incoherent or uninstantiated. In the same way, the distinction between propositions that are true (or false) by virtue of the content of the concepts involved, and propositions that use these concepts to make claims whose truth-value depends on extra-conceptual matters, ceases to be a fundamental issue. The distinction remains, but how we treat a specific proposition can be quite variable. We saw that Sellars made this point when he considered a case in which we recognize an empirical generalization involving a concept and consider whether to alter the concept so as to include this generalization in its content (SRLG 357). If we make the change, we move to an altered conceptual system in which the counterpart proposition is a conceptual truth. In general, when beliefs provide a successful guide for thought and action we tend to treat them as conceptual truths. If further experience leads us to change our view of the adequacy of those beliefs, we tend to shift their status to that of empirical falsehoods.

Finally, I want to comment on one aspect of the relation between TC and Sellars' theory of concepts. In discussing the development of scientific theories Sellars rejects the logical empiricist view that scientists first establish empirical laws, and then introduce theories to explain these laws. Instead, Sellars

holds, empirical laws are often only approximately correct; theories correct these laws and explain "why observable things obey to the extent that they do, these empirical laws" (LT 121). My goal in building on Sellars' theory of concepts is to bring about progress in this Sellarsian fashion. I consider Sellars' theory of concepts the most successful and fruitful of the extant theories, although I have indicated specific ways in which it is inadequate. I have aimed to produce a theory that improves on Sellars' work, and explains why it has the successes that it does.

# 6 Clarifications, Responses, and Refinements

> Unfortunately he mislocates the truth of these conceptions, and, with a modesty forgivable in any but a philosopher, confuses his own creative enrichment of the framework of empirical knowledge, with an analysis of knowledge as it was.
>
> (EPM 195)

I turn now to some recent theories of conceptual content. The discussion falls into two clusters. Secs. 6.1–6.3 deal with some influential forms of externalism – theories holding that all or part of conceptual content occurs outside the individual mind.[1] TC is an internalist theory, and I will pursue two aims in this part of the discussion: arguing that we need an internalist account of conceptual content, and providing further clarification of some aspects of TC. In Sec. 6.4 I discuss work by several philosophers who study conceptual change in science using versions of what Nersessian calls "cognitive-historical analysis" (e.g., 1986, 1992) – an approach that integrates historical studies with recent work in cognitive psychology. In this part of the discussion I defend and further explain TC, but I also draw on some results of this work to enrich TC. In both clusters I will not discuss the work of everyone who has interesting things to say. My only defense is to plead finitude.

## 6.1 Natural Kinds

We can begin our discussion of externalism with an observation about the way we fix reference in ordinary language (Donnellan 1966): Sometimes we describe an item, and succeed in picking it out even though the claim embedded in the description is false. This point sustains a fair degree of generalization: We may succeed in picking out an item even though most of our beliefs about it are false. Building on this point, Kripke (1980, first published in 1972) developed a theory of proper names, which he and Putnam (1975, first published in 1973) extended to natural-kind terms; I will discuss this latter extension (henceforth KP).

The key idea is that nature is divided into natural kinds, where each kind is characterized by some underlying structure – an essence – that makes it the kind it is. All instances of a kind share this essence. We are often able to pick out a natural kind even though we cannot specify the essence; we can then make that kind a subject of scientific study that will presumably culminate in the discovery of its essence. The study may go on for a long time and require a great deal of innovation before we discover this essence, but the natural kind provides a stable item to which all our hypotheses refer, and against which all are tested. For example, water and gold are natural kinds that were recognized early in our cognitive history, and that have been the subject of research. In the course of this research many false claims have been made about these natural kinds, but the object of research has remained stable throughout and has provided the touchstone against which claims have been evaluated. We may even have reached the point at which we know the essences in these two cases; they are given by contemporary physical and chemical theory: gold is element 79 on the periodic table; the essence of water is captured in the chemical formula $H_2O$.

Semantic externalism is the view that these essences provide (at least part of) the meaning of natural-kind terms. The key argument for this claim is Putnam's twin-earth thought experiment (1975). As I noted in Sec. 1.4, Putnam is explicitly concerned with word meaning which he distinguishes from concepts.[2] Putnam asks us to consider a twin-earth: a world exactly like ours except that the chemical structure of water is XYZ rather than $H_2O$, although twin-earth water has the same easily observable superficial properties as our water. Putnam maintains that the process by which we learn "water" includes an association between the term and the typical stuff we refer to by this term. As a result, this referent is a permanent part of the meaning of "water." For folk brought up on twin-earth, XYZ functioned as the referent of their "water"; thus the meaning of their term is different from the meaning our term. This difference holds even if we consider two individuals, one from each planet, who are identical in every respect – including what occurs in their heads.[3] Thus, Putnam concludes, meaning is not completely in the head. Note that there is no actual argument for this conclusion. The conclusion follows from the story plus an appeal to intuitions about meaning – intuitions that not everyone shares.

The role of the referent in determining the meaning of "water" is one respect in which meaning has an external dimension; Putnam uses another example to introduce a second external aspect – one located in the linguistic community. He maintains that I can use "elm" correctly even though I do not know much about elms (for example, I cannot distinguish elms from beeches), because there is a "division of linguistic labor" (1975: 245–47): there are experts in my community who know about elms, and to whom I defer. Thus "extension is, in general, determined *socially* . . . " (1975: 245). Another item associated with a term also contributes to its meaning. This is a *stereotype* – an image of a typical member of the class. Stereotypes are

located in the individual mind, play a key role in communication, and (Putnam claims) provide the "sole element of truth in the 'concept' theory of meaning" (1975: 250).[4] I have nothing to add about this view of concepts beyond what is provided by the arguments already given for TC. In this section I focus on the two theses that an underlying essence provides the referent for natural-kind terms and that the discovery of this essence is a major aim of scientific research. In Sec. 6.2 I discuss a more developed form of social externalism due to Burge.

Given the role that natural kinds play in the KP account of scientific research, the claim that we are rather good at picking out natural kinds on the basis of unaided observation is central.[5] If this ability is significantly more limited than advocates of KP assume, we will encounter a good deal more conceptual change in the development of science than they recognize. I want to examine how successful we have been at picking out natural kinds.

Consider, again, the "elements" proposed in ancient Greece and China (Sec. 2.5). With the possible exception of water, none of these – air, earth, ether, fire, metal, and wood – are now considered natural kinds.[6] The three isotopes of hydrogen make chemically pure water an especially interesting case. If the hydrogen in our water is deuterium the water is toxic;[7] if it is tritium the water is radioactive. Zemach (1976: 120) notes that since hydrogen and oxygen each have three isotopes, there are eighteen different kinds of $H_2O$, each having some properties that distinguish it from the others. (Zemach also mentions some of the other examples I discuss below.) The existence of different isotopes of oxygen has additional consequences. For example, although oxygen-16 is vastly more common than oxygen-18, their ratio in the oceans has varied over time:

> When the world is in a cooler part of the global temperature cycle, water molecules containing oxygen-16 evaporate more easily than their heavier oxygen-18-containing counterparts and so the snow that falls at the polar ice-caps, and becomes locked up as ice, is very slightly richer in the former, and the water that is left behind in the oceans is very slightly richer in the latter. Marine creatures therefore lay down shells that have more oxygen-18 than expected and these are preserved in sediments. Analyzing the oxygen-18:oxygen-16 ratio in such deposits reveals the cycle of global cooling and warming that has characterized the past half million years with its five ice ages.
>
> (Emsley 2001: 304)

Emsley adds that this ratio of isotopes also varies geographically. This is reflected in people's bones and teeth, and has been used to show, for example, that some "people buried in England during the Roman period . . . come from southern parts of the Roman Empire." Most elements have several isotopes with varying properties. Thorium, for example, has more than twenty-five known isotopes, all radioactive, with half-lives running

from microseconds to billions of years. In addition, liquid oxygen has a magnetic field that gaseous oxygen does not have (2001: 303). Are liquid and gaseous oxygen the same natural kind?

Before considering more examples I want to emphasize the point I am after. Advocates of natural kinds characterized by essences can reply that essences are difficult to discover and that they are not committed to the view that we have actually discovered the essences of any natural kinds. I will consider this reply below; for the moment I want to stress two points. First, the claim that we are fairly good at recognizing natural kinds on the basis of naïve observation does not stand up. To be sure, we do not always fail. Gold has only one isotope, so in this case our distant ancestors may have succeeded in picking out a natural kind. But as Kuhn points out (1989: 79), gold is a special case, not the norm. Incessant citing of a few confirmatory instances while ignoring a large body of counter-examples is not good methodology. Second, the discovery of isotopes was required before we could recognize a major source of diversity. Rather than our being able to pick out natural kinds that provide stable points in the flow of conceptual change, conceptual innovation may drastically alter our understanding of where to look for natural kinds. This suggests that NATURAL KIND fails to fulfill one of the theoretical roles for which it has been deployed. Some additional examples will underline just how serious this failure is.

Consider *isomers*. These occur when we have two or more molecules with the same chemical formula, but with their atoms arranged differently in space. This can result in significantly different properties. Chemists recognize various types of isomers, but I will give just one example here. Many biologically important molecules occur in a twisted form, where the twist can be either in the left-hand or right-hand direction. Typically only the left-hand version is biologically active. Substitution of the right-hand version may result in eliminating a significant function, which may be unpleasant or even fatal. A similar situation occurs for those elements which occur in different *allotropic forms*. Ignoring spatial structure, diamond and graphite are both pure carbon. Nevertheless, they have different properties and play different industrial and social roles. In this case items that appear to be strikingly different on a superficial examination turn out to be "the same" on a deeper level of analysis. One reply is to include spatial organization of the constituents in the essence of natural kinds, but this underlines the point that natural kinds are not easily picked out early in our cognitive history. Moreover, if we include spatial organization of constituents as a feature that serves to distinguish natural kinds, then different excitation states of an element (which involves different distances between some orbital electrons and the nucleus) may count as distinct natural kinds. Perhaps we should also consider different ionization states of atoms to be distinct kinds since these are items with the same atomic number but different numbers of electrons, yielding different properties.

Emery, ruby, and sapphire provide another interesting example. They are all primarily composed of the same isomer of aluminum oxide ($\alpha$-$Al_2O_3$)

with different "impurities" determining their colors and other properties. If one holds that the impurities are not part of the essence, we have more cases in which dramatically different appearances lead us to mistakenly believe that we are dealing with different natural kinds. If such impurities result in distinct essences, the number of natural kinds grows substantially. We thus find further erosion in the value of NATURAL KIND as an explanatory and unifying concept. Moreover, as Kuhn points out (1989: 83–84), the distinction between superficial properties and underlying essence is misleading. In modern physical theory, the "superficial" properties are just as essential as the "deeper" properties. If gold looked blue to normal observers in standard conditions, it could not have atomic number 79. Similarly, if Putnam's XYZ had an elaborate chemical formula, it could not have the same apparent properties as water. For example, it would be too heavy to evaporate at normal earth temperatures (Kuhn 1989: 80).[8] Yet another set of problems arises from biology. Both Kripke and Putnam use biological species (e.g., tigers) as examples of natural kinds, but the view that species are distinguished by essences puts them seriously at odds with evolutionary biology: "a very traditional issue is whether there is some essential property defining membership of a species . . . it is currently rather uncontroversial that the acceptance of Darwinism forces the rejection of this aspect of essentialism" (Dupré 1993: 38).[9]

We can now consider the reply mentioned above on behalf of advocates of KP: KP is not committed to the claim that we have actually found the essence of any natural kind. Thus we can hold onto this claim against empirical failures, and continue to use it as a basis for scientific research. This suggests that the two theses – (1) the world divides into natural kinds characterized by essences that determine their properties, and (2) we can seek these essences through empirical research – are GAs. I urge that they are GAs whose time has passed. We regularly fail to pick out natural kinds, and our understanding of *where to look for them* shifts as a result of research – including research that involves major conceptual innovation. Instead of providing a stable focus around which conceptual change can flow, our understanding of where we are liable to find natural kinds depends on our current conceptual repertoire. When we add the point that KP is fundamentally at odds with contemporary biology, we have little reason for continuing to think of this view as a valuable guide to research, and little reason to continue treating ESSENCE and NATURAL KIND as significant explanatory concepts. Moreover, TC provides sufficient resources for understanding continuity through conceptual change without needing the sort of fixed points that essences and natural kinds are thought to provide.

## 6.2 Social Content

I turn next to work by Tyler Burge, who also rejects the view that conceptual content is completely determined by what occurs in an individual mind.

Instead, Burge argues, a social factor is involved in individuating content. This claim is put forth with limitations:

> Some mental states (for example, some perceptual states) depend for their identity on the nature of the physical environment, in complete independence of social practices. . . . [E]ven where social practices are deeply involved in individuating mental states, they are often not the final arbiter.
>
> (1986: 707)

Nevertheless, he holds that virtually all concepts have a social component: "Nearly anything . . . including technical and everyday natural-kind notions. . . . Concepts of ordinary objects and stuffs, which are not natural kinds. . . . Notions associated with common verbs. . . . " (1986: 709).[10] This social element enters into *mental* content: "Social context infects even the distinctively mental features of mentalistic attributions" (1979: 87). Burge rejects any attempt to reduce thought to language or language to thought. He notes that twentieth century philosophers have often attempted to explicate thought in terms of linguistic meaning, while a more traditional view attempts to move in the opposite direction; "A third view, which I regard as correct, is that the two notions are interwoven in complex ways which render it impossible fully to analyze one in terms of the other" (1986: 718).

I will focus on a central issue: whether it is appropriate to attribute a concept to someone who has an inadequate understanding of that concept. His main example (1979: 77–79) concerns a person who suffers from arthritis; one day, feeling pain in a thigh, this individual concludes that the arthritis has spread to the thigh. However, when informed by a doctor that arthritis occurs only in joints, the patient no longer describes the thigh pain as arthritis. Burge maintains that the patient's error is "conceptual or linguistic. . . . It is not an ordinary empirical error," (1979: 82); it is not a case in which the patient just has a false belief associated with ARTHRITIS. Still, Burge maintains, it is correct to use "arthritis" to describe the patient's mental content. Before speaking to the doctor, the patient had an *inadequate* notion of arthritis, but still had *the* notion of arthritis. Burge attempts to clarify this kind of conceptual error by considering a counter-factual situation in which "arthritis" is used to include certain cases of thigh pain. Except for this difference, every feature of the actual and counter-factual patients' experience and inner constitution are identical (until the conversation with a doctor). In the counter-factual case the patient's report is true; thus the two patients are expressing different mental contents in spite of being internally identical. It would not be correct to describe the counter-factual person's mental content using our concept ARTHRITIS, and since social context is the only difference between the two cases, social context plays a role in constituting conceptual content.

Burge considers a large number of additional examples – brisket, clavichord, contract, fortnight, mortgage, recession, and more – in pressing his claim that people can be correctly described as possessing a concept even while the understanding of that concept is in some ways inadequate. Inadequate understanding of common concepts is widespread: "One need only thumb through a dictionary for an hour or so to develop a sense of the extent to which one's beliefs are infected by incomplete understanding. The phenomenon is rampant in our pluralistic age" (1979: 79).[11] Burge discusses several alternative interpretations of his thought experiments but I will consider only one of these because it is the view I want to defend, although on different grounds than Burge considers: the patient has a different concept associated with the word "arthritis" before and after visiting the doctor. TC, then, is an example of the sort of individualistic theory that Burge rejects. I want to spell out how TC deals with the arthritis case.

According to TC we have a situation in which doctor and patient (initially) have different, but similar, concepts associated with "arthritis." We can compare these concepts along each of the three dimensions recognized by TC, although our comparisons will be limited because we are dealing with a fictional example that is only partially developed. Presumably the doctor's criteria for recognizing arthritis are reasonably clear and all instances that would be considered arthritis by the doctor will also be considered arthritis by the patient. Thus there is substantial overlap along this dimension. Patient and doctor would also likely agree on many implicational relations between descriptions of symptoms and ARTHRITIS. But the doctor may require tests in addition to symptoms, so that the conditions the patient considers sufficient might be considered relevant but not sufficient by the doctor. Indeed, in the case of thigh pain the patient infers arthritis while the doctor infers not arthritis. The systemic role of the doctor's concept is presumably determined by its place in a systematic classification of diseases. We are not told enough to be clear on the role of the patient's concept; it may have no role beyond serving as a label for a variety of pains. But we are given enough information to see that there are sufficient overlaps in the two concepts for the doctor to understand the patient, and for the patient to have little difficulty replacing an initial concept with one that is in closer conformity with the doctor's. TC also allows for cases in which someone possesses a concept but does not fully understand it because of unnoticed implications among the concepts in the system, or because the instantiation conditions (for descriptive concepts) or required actions (for prescriptive concepts) have unnoticed implications. Indeed, lack of complete understanding will be the norm. (Cantor and Frege did not have a complete understanding of early set theory since they did not recognize that it entailed an inconsistency.)

We have, then, alternative accounts of the phenomenon in question, so let us ask why it matters which account we accept, and how we should decide whether to favor one account or the other. The answer turns on differences

between Burge's larger project and my own. Burge is interested in clarifying common mentalistic concepts:

> My objective is to better understand our common mentalistic notions. Although such notions are subject to revision and refinement, I take it as evident that there is philosophical interest in theorizing about them as they are now. I assume that a primary way of achieving theoretical understanding is to concentrate on our *discourse* about mentalistic notions.
>
> (1979: 87)

The many examples Burge uses are a means to this end. The concepts of central interest are such mentalistic concepts as "misconception, incomplete understanding, conceptual or linguistic error, and ordinary empirical error" (1979: 88). Burge insists that discourse involving these notions should be taken literally unless there are specific reasons for doing otherwise, and that the touchstone for deciding how to deal with the various examples he explores is ordinary intuitive plausibility. He uses specific examples to elicit our non-theoretical intuitions about these mentalistic concepts. Two points are especially important. First, Burge's claim that doctor and patient associate the same concept with "arthritis" is *not* based on intuitions about ARTHRITIS; it is based on intuitions about CONCEPT. It is the content of this concept that is at the center of the discussion.[12] Second, Burge treats ARTHRITIS and CONCEPT differently in that he does not invoke experts about concepts to whom we should defer. Here he is concerned with everyday concepts. But Burge is doing more than just describing ordinary intuitions, since he takes these as normative in deciding how to think about ARTHRITIS and the other cases he considers. From my perspective the key question is whether Burge's approach is likely to provide insight into the nature of cognition. I want to consider the arthritis example with this question in mind.

As a result of the conversation with the doctor, the patient's epistemic position improves; let us ask why the patient should learn from the doctor rather than the reverse. It is not enough to say that the doctor is the expert – we want to understand the nature of that expertise. Consider two different answers. One is that the doctor is better informed about current medical concepts – that the doctor knows more than the patient about ARTHRITIS. The other reply is that the doctor knows more about arthritis. Presumably, the latter point is of greater interest to the patient who is mainly concerned with the causes of various pains and with treatment options; learning medical concepts is likely of only secondary concern. Before talking to the doctor, the patient should expect that the same treatment will relieve both the joint pain and the thigh pain. After talking to the doctor these expectations should change. This distinction between two kinds of answers does not apply to all of Burge's examples. BRISKET, for example, is a purely conventional

concept with no independent facts to be tapped in considering its content.[13] But CONCEPT is like ARTHRITIS: it is part of an account of cognition; we want the content associated with "concept" to accurately describe its subject matter.

Now Burge's counter-factual case is not especially far-fetched. Further research might reveal a previously unknown biological phenomenon that is at the root of both the thigh pain and the joint pain, and result in a single treatment for both. We have already encountered such cases in the history of science. Recall that the discovery of isotopes explained how samples of different atomic weights could be instances of the same element. In Burge's example such a discovery would also be an empirical result, not a result that comes from reflection on pre-existent concepts, and might well require conceptual innovation. If this occurred, the word "arthritis" might or might not continue to be used.[14] Let us consider how Burge's might deal with such cases.

Although Burge does not discuss such cases in any detail, here is one remark that he does offer:

> Dalton and his predecessors *defined* "atom" (and its translations) in terms of indivisibility. Major theoretical changes intervened. The definition was discarded. Despite the change, we want to say, Dalton wrongly thought that *atoms* were indivisible: despite his erroneous definition, he had the "concept" of atom (not merely the referent of "atom").
>
> (1986: 716)

But consider some of the empirically-motivated theoretical changes that occurred from Dalton's day until, say, the late 1930s – changes that included the discovery of isotopes, electrons, protons, neutrons, relativity, modern quantum theory, and more. TC views this case as involving a series of conceptual changes in which the concepts at the two ends of the series have little in common, although there is a great deal of overlap at each stage of the process. The strengths of TC include its ability to explain the interplay between continuity and change in such cases, and provide a guide to their detailed study. Burge does not offer an alternative account. All he offers is the claim that a single concept is associated with "atom" throughout this history. Presumably this claim is based on a shared intuition, although he gives no evidence for holding that this intuition is widespread (a theme that I will return to in Chs 7 and 8). CONCEPT, and related notions such as the difference between conceptual error and ordinary empirical error, were introduced in the course of reflective and empirical attempts to understand certain aspects of cognition. There is no guarantee that the first stabs at understanding these phenomena got them right, or that any versions that filtered down into common thought have normative force for future research.

Immediately after his discussion of ATOM Burge emphasizes that he is concerned with everyday concepts, which are different from theoretical concepts of science.

> It would be a mistake, however, to assimilate common sense notions to a theoretical paradigm. Although meaning-giving characterizations from ordinary terms or notions are vulnerable to theoretical change, they differ from theoretical definitions of terms whose original home is a systematic theory, not only in that they are more stable and in that sense less vulnerable to theoretical criticism. They also differ in the means by which they are known and checked and in the ways in which they are vulnerable.
>
> (1986: 716)

He then relates the distinction between common sense and theoretical terms to that between observational and theoretical terms. The meaning of common sense terms derives from

> reflection on perceived examples picked out by common indexical usage. By contrast, the natural sciences, whose methodology we best understand, do not expect to reach their normative characterizations through simple reflection on usage or common perceptual experience. Theoretical terms are not indexically applied to perceived objects.
>
> (1986: 716)

On this basis the main concept that concerns us here – CONCEPT – is a theoretical concept; so is ARTHRITIS. In neither case can we pick out instances in the way we pick out, say, sofas (a main example in Burge's 1986 discussion). In addition, we have already encountered many concepts that had their original home in ordinary discourse, but were are taken up into systematic thought where they are replaced by successor concepts. Moreover, Burge has nothing to say about the introduction of concepts for newly discovered phenomena such as radioactivity. The upshot, then, is that TC can give an account of the cases in which Burge's view applies, while Burge's approach has nothing to offer that helps us understand or study conceptual change.

## 6.3 Informational Atomism

Informational Atomism (IA), another influential form of externalism, rejects any internal or social contribution to conceptual content. IA is a combination of two distinct views. Conceptual atomism begins with the thesis that there is a set of primitive concepts out of which all other concepts are constructed, but adds the claim that primitive concepts do not have any structure. In particular, the content of a primitive concept is independent of the content of every other primitive concept, so that having one primitive

concept does not require having any other concept (e.g., Fodor 1998: 13–14, 22; all Fodor references in the present section are to this book). In principle, a conceptual repertoire may consist of just one concept. The term "information" is used here in the sense introduced into philosophy by Dretske (1981). The idea can be explained by considering tree rings, which carry information about the tree's age. They carry this information as a result of a causal process, and carry it independently of whether any human is aware of this fact. The key claim, then, is that *A* carries information about *B* whenever *A* was *appropriately* caused by *B*. *A* need not resemble *B* or have any particular features in order to carry this information. A descriptive concept's content is the information it carries about its cause. Now let us combine this view with atomism. We can analyze tree rings – e.g., we can count them to determine the tree's age. But if we could "look" at an atomistic concept we would find nothing to analyze. The concept is a mental entity, but its content is its extra-mental cause. In particular, no beliefs are part of the content of a primitive concept, although many beliefs may be associated with it.

Fodor gives the following formulation: "what bestows content on mental representations is something about their causal-cum-nomological relations to the things that fall under them: for example, what bestows upon a mental representation the content *dog* is something about its tokenings being caused by dogs" (12). But only certain kinds of causal relations will do. Fodor maintains that I acquire DOG when I become *nomologically locked* to doghood, where doghood is whatever property makes something a dog. The interaction must be law-governed (thus "nomological") and result in something that occurs in my mind (133). Exactly what must occur depends on the kind of mind I have (136–37, 139–40, 142–43). My mind must include an appropriate mechanism to mediate this interaction, but the mechanism does not contribute to the content of the concept. Fodor does not attempt to specify the details of what occurs in my mind when I acquire a concept, nor consider the range of possible mechanisms, although he does maintain that nomological locking typically occurs as a result of some perceptual experience of the items that become the content of a concept. The addition of nomological locking to IA generates what Fodor calls *supplemented informational atomism*. This supplement concerns the relation between a concept and its content, and the process by which an extra-mental item becomes the content of a concept. The supplement does not alter the IA view of conceptual content – which is my concern here. I will follow Fodor's recent discussion, but leave nomological locking in the background; thus my remarks apply to IA in general.

The thesis that conceptual content is independent of the mechanism that relates a concept to its content has some interesting consequences. First, it is possible to enter into the appropriate relation to a particular content by means of different sensory modalities. This eliminates any need to distinguish between, say, a tactile and a visual concept of dog.[15] "It's *that* your mental structures contrive to resonate to *doghood*, not *how* your mental

structures contrive to resonate to *doghood*, that is constitutive of concept possession according to the informational view" (76). Thus, Fodor concludes, he has the same concept DOG as Helen Keller: "For Helen Keller, it was *not* visual perception that sustained the meaning-making dog-DOG relation. Yet she and I, each in our way, can both satisfy the conditions for DOG-possession according to the present account of these conditions" (76, I have replaced Fodor's notation for concepts with my own and will continue to do so). As a result, conceptual content does not vary among individuals or cultures.

> It seems pretty clear that all sorts of concepts (for example, DOG, FATHER, TRIANGLE, HOUSE, TREE, AND, RED, and, surely, lots of others) are ones that all sorts of people, under all sorts of circumstances, have had and continue to have.
>
> (29)

For the same reason, Fodor notes, he has the same concept of food as Aristotle and the same concept of triangle as Einstein (29). Moreover, if the same causal relations produce WATER and $H_2O$, then these concepts have identical content (13). Since Fodor moves "back and forth pretty freely between concepts and word meanings . . . " (12) and regularly uses "concept" and "word" interchangeably, he concludes that "water" and "$H_2O$" have the same meaning. But Fodor also concedes that WATER and $H_2O$ are not the same concept, so "*content* individuation can't be all that there is to *concept* individuation" (15). I will not develop Fodor's account of the additional element needed for concept individuation, since it would require a long digression, and my concern here is with conceptual content.[16]

For IA concept possession is non-cognitive: possessing the concept C does not require knowing or believing anything about Cs. I may associate many beliefs with a concept, but none of these beliefs are part of the content of that concept; two people can have the identical concept C while not sharing a single belief about Cs. Although I have exactly the same concept DOG as Helen Keller, each of my beliefs about dogs may contradict one of her beliefs about dogs. Fodor considers this result a major attraction of IA since it offers a solution – albeit a radical solution – to the key problem he has pressed against conceptual role theories: how to provide a principled distinction between sentences that express conceptual content, and those that do not. According to IA, primitive concepts have no content that can be expressed as sentences, so the issue does not arise. Fodor also holds that most lexical concepts are primitive (121) and that the content of non-primitive concepts is determined compositionally from their lexical components. An analysis of a non-primitive concept will, therefore, consist of a description of its composition, but beyond this, there will be no sentences that describe the *content* of this concept either.

At this point we can see why IA is irrelevant to the project of this book. My concern is with the role of concepts in the development of knowledge,

and as guides to action and thought. According to IA, these epistemic and normative issues concern the beliefs associated with concepts, not the concepts themselves. The content of our descriptive concepts does constrain beliefs since that content consists of the items in the world and provides the subject matter of these beliefs. We test our beliefs by attempting to interact with that subject matter. If we adopt IA, the study of the development of knowledge would be concerned with the way these collateral beliefs are accepted, reconsidered, and changed. It is not clear that concepts would play any role in this study. But now the disagreement between IA and TC seems purely verbal: Fodor may just be using "concept" differently than I do. Yet the kind of theory I am defending draws on a use of "concept" that has a long and continuing history in psychology and philosophy. Prinz underlines one example of this disparity: For IA concepts play no role in categorization, "The atomist says that an explanation of categorization is not within the explanatory jurisdiction of a theory of concepts" (2002: 99). Yet most psychologists consider understanding categorization to be "the main motivation for postulating concepts; they implicitly define 'concepts' as the mechanisms by which we categorize. To say that concepts do not contribute to categorization is almost incoherent from this perspective" (Prinz 2002: 99). Fodor provides no reasons for abandoning this use of the term.

Suppose, however, that the disagreement is not purely verbal, that there is sufficient overlap between TC and IA to generate a genuine dispute. Theory choice is always a matter of balancing successes and failures. IA focuses on a set of problems that mainly derive from philosophy of language. (For discussions of problems solved and problems left unsolved by atomistic theories of concepts see Margolis and Laurence 1999: 59–71; Prinz 2002: 89–100, 241–49.) Fodor provides little discussion of the sophisticated concepts I am primarily concerned with, and little reason to think that IA can be extended to deal with them. Consider the theoretical concept PROTON. While PROTON is a lexical concept, Fodor acknowledges that IA will not work in this case, and suggests that PROTON may not be primitive (130, n. 9). This leaves the task of analyzing PROTON in terms of primitive concepts. In Sec. 3.5 I reviewed the best-developed attempts at such analysis and found them unsatisfactory. Of the attempts considered, only the first was compositional; the considerations that led the logical empiricists to reject it raise serious doubts about the prospects of any compositional account of theoretical concepts in terms of non-theoretical concepts. But whether we insist on compositionality or not, proponents of IA owe us an account of theoretical concepts. Historically, the failures of attempts to carry out this project in terms of primitives acceptable to empiricists were a major motivation for a move to holistic accounts. Sellars was one of the first philosophers to move in this direction with full knowledge of these prior attempts. Note especially that for IA two people could have the concept PROTON without sharing any beliefs about protons. This result is particularly implausible for theoretical concepts because these concepts raise two distinct questions:

What is the content of the concept? and Is the concept instantiated? It is thoroughly bewildering how we might attempt to answer the second question except on the basis of beliefs about protons. If no beliefs are included in the content of a concept, it is difficult to see what role a concept would play in our attempts to decide if it has instances.[17]

There are other concepts that IA must address in addition to theoretical concepts. These include mathematical concepts such as a LOGARITHM and DIFFERENTIAL OPERATOR, logical concepts such as ENTAILMENT, grammatical concepts such as a SPLIT INFINITIVE, legal concepts such as DRIVER'S LICENSE (understood as a specific right, not as a piece of paper), philosophical concepts such as ANALYTIC PROPOSITION and TRANSCENDENTAL ARGUMENT, concepts that specify ideals such as ABSOLUTE EQUALITY (cf. Keil and Wilson 2000: 316), and many more. IA will also have to provide an account of the many descriptive concepts that have been abandoned over the course of our cognitive history. Presumably, people cannot have beliefs about prepotency without having the concept PREPOTENCY, nor can they have beliefs about phlogiston without having the concept PHLOGISTON. These are clearly different concepts, and (according to IA) they are not primitive concepts, since there is nothing in the world to provide their content. It is far from clear how these concepts are to be constructed from concepts that have content. Nor do I see any reason for taking seriously the claim that two people could have identical concepts of PREPOTENCY without sharing a single belief about this supposed phenomenon. On balance IA has nothing to offer as an approach to the problems that motivate TC.

I want to introduce a general methodological issue at this point. The question of what concepts are, *really*, will not be answered by peering more carefully into minds, brains, or an essence. Rather, CONCEPT is a concept in a cognitive theory (which must ultimately be tied to a neurological account); the question of what concept we should associate with the word "concept" is to be decided by assessing competing theories. I have no fantasy that I will provide the final word on this topic. There will be open questions, and the balance in favor of one theory over another will be determined by comparative evaluation of problems solved, challenges remaining to be addressed, and our judgments of the fruitfulness of the research directions a theory supports. At the present stage of our knowledge it is probably desirable that more than one theory be pursued. As we learn more about cognition, and (hopefully) develop better theories, we should expect that the concept we associate with "concept" will change much as the concept we associate with "atom" has changed. It is even possible that at some time in the future we will drop the word "concept" from this endeavor.

## 6.4 Cognitive-Historical Analysis

In this section I examine work by several philosophers who seek to understand conceptual change by combining historical studies with recent work

from psychology and cognitive science. I begin with some work from psychology that they all accept.

Historically, philosophers and psychologists assumed that all concepts are constituted by necessary and sufficient conditions (NS). In philosophy this view was attacked by Wittgenstein (1953) in his account of word meaning in terms of family resemblances and overlapping strands.[18] In psychology the Wittgensteinian approach was put on an empirical foundation by the work of Rosch and her colleagues (e.g., Rosch 1973a, b, 1978; Rosch and Mervis 1975). The most influential outcome of this work is the discovery of *typicality effects* in people's classifications of familiar items. For example, subjects regularly respond that a robin is a better example of a bird than is a turkey, a car is a better example of a vehicle than is a raft, and a gun is a better example of a weapon than is a screwdriver. Such distinctions are inappropriate on the NS view which implies that an item is either a member of a class or not; there are no degrees of class membership, and no borderline cases in a properly constructed concept. A common response to this data has been to propose new accounts of how we store concepts and assess whether an item falls under a particular concept. For example, *prototype theory* holds that we store a concept by abstracting a typical instance and classify new items by comparing them with this prototype. A competing view holds that we store a set of typical members of a class without abstracting a prototype. These examples serve as *exemplars* of the class, and we assess new items by comparing them with the exemplars. An item may share different features with different exemplars, and advocates of this view have proposed several schemes for how we weight these identities in arriving at a classification. (For historical reviews and discussions of these and other approaches see Lakoff 1987; Medin 1989; Smith and Medin 1981.)

Both views have faced numerous criticisms, but I will consider only one objection here. A study by Armstrong, Gleitman, and Gleitman (1999) identified cases in which subjects exhibit typicality effects for concepts even though they can also state NS conditions for those concepts. For example, people who give a standard definition of "triangle" may still hold that some triangles are better instances than others. One consequence of this study is that typicality effects do not, by themselves, eliminate an NS account of how concepts are stored. But there is a more general point that Rosch has emphasized: there is a difference between acknowledging typicality effects (sometimes called "prototype effects") and proposing an account of how concepts are stored and applied: "prototypes only constrain but do not specify representation and process models" (1978: 41). Lakoff puts the point in particularly strong terms: "It is important to bear in mind that prototype effects are superficial. They may result from many factors" (1987: 45). Instead of constructing a theory that mirrors some of the data, we should seek a more fundamental theory that explains typicality effects along with other aspects of human concepts. One theory that meets this desideratum holds that concepts are represented by *frames*. Versions of this view have a

considerable history (Thagard 1984), and a recent version developed by Barsalou (1992) has been adopted and adapted by several philosophers to analyze conceptual change in science (see Barsalou and Hale 1993 for comparisons with other psychological accounts of concept representation). These applications will provide the main focus of my discussion.

A frame provides a convenient way of displaying the content of concepts that have a particular hierarchical structure: The concept is associated with a set of *attributes*, and each attribute has a set of mutually incompatible *values*. Consider BIRD: Attributes that characterize birds include body size, shape of beak, type of foot, and type of neck. Each kind of bird has one of a set of values for each attribute: body size may be small or large, the beak may be round or pointed, the foot may be webbed or unwebbed, and the neck may be short or long (Chen and Barker 2000: S210). Here are two other examples given by Barsalou (1992: 30); in each case I give the attribute in italics followed by a list of values. First, the attributes of CAR include: *fuel*: gasoline, diesel, gasahol; *engine*: four-cylinder, six-cylinder, eight-cylinder; *transmission*: standard, automatic; *wheels*: steel, alloy. Second, here is part of a frame for VACATION, which admits of more variability than the previous examples. Attributes may include *location*: mountains, woods, seaside; *distance*: near, far; *activities*: climbing, hiking, swimming, and more (33–34). Each attribute and value in a frame is itself a concept that can be represented by a frame (in this sense, frames are recursive). In the case of CAR, a gasoline engine has such attributes as *spark plugs* and *valves*; each of these has values that would be familiar to a mechanic. In general, a frame provides a means of mapping out the part of conceptual content that consists of relations to other concepts. A set of attributes and values generates a *conceptual field* – a set of contrasting concepts that describe different kinds of instances of the concept in question. It is not required either that all of these kinds have actual instances, or that they all describe concepts that anyone actually employs.

The attributes and values included in a frame range over a variety of different kinds of items with different relations to the frame's subject concept. In the case of BIRD we should distinguish the neck, which is part of the bird, from its size, which is a property but not a part. In the case of CAR the wheels are part of the car, but the values for wheels concern the material that makes up the wheels. In the case of VACATION, swimming and hiking are activities we may engage in; pursuit of an activity is a goal associated with the vacation. The key feature of attributes is their close association with the concept under consideration. Some proponents of frames hold that a concept entails its attributes, but Barsalou considers the relation to be probabilistic (1992: 5).

Frames also include connections among attributes and among values; these represent further beliefs about relations between items displayed on a frame. *Structural invariants* are "*relatively constant* [italics added] relations between a frame's attributes" (Barsalou 1992: 37); they embody both empirical

and conceptual connections among these attributes. For example, the frame for CAR includes the attributes DRIVER and ENGINE, but these are not merely juxtaposed – they are connected by our understanding that the driver operates the engine.

> Structural invariants capture a wide variety of relational concepts, including spatial relations (e.g., between *seat* and *back* in the frame for *chair*), temporal relations (e.g., between *eating* and *paying* in the frame for *dining out*), causal relations (e.g., between *fertilization* and *birth* in the frame for *reproduction*), and intentional relations (e.g., between *motive* and *attack* in the frame for *murder*).
>
> (1992: 35–36)

Connections among values are called *constraints*. They capture our beliefs about ways in which values limit each other, and are more variable across instances than structural invariants. Constraints come in several varieties; I will pick just a few of Barsalou's examples. In the frame for TRANSPORTATION there is a negative constraint between SPEED and DURATION since the duration of a trip varies inversely with speed; there is a positive constraint between SPEED and COST since cost tends to be higher for faster modes of travel. We also tend to associate faster modes of transportation with travel over greater distances. The activities that form our goals generate further constraints: If our vacation plans include surfing, then we must arrive at an ocean beach; if our plans include downhill skiing, we need snow and mountains (1992: 37–38). Note also that the inverse connection between speed and duration may be a reflection of a physical fact, or perhaps a consequence of the way we define "speed." The connection between speed and cost represents (perhaps) an economic relation; it is more variable than that between duration and speed. On a given day it may cost more to take a train between two points in the US than to fly. The relation between distance and speed is subject to personal preferences and other local considerations: "someone may want to travel slowly over a long distance to see beautiful scenery" (37). Barsalou classifies constraints into different types, but I will not follow that elaboration here since these distinctions play no important role in the applications that concern me.[19] Instead, I want to examine some attempts to use frames in the analysis of conceptual change in science.[20]

My first example is Barker's (2001) account of a key conceptual change as we move from Ptolemaic astronomy to Copernicus and then to Kepler. Barker argues that before Kepler astronomers sought to calculate the *path* of celestial objects against the background of the fixed stars, rather than an object's *orbit* understood as a real track through three-dimensional space. Ptolemaic astronomy divides celestial objects into three classes: the fixed stars, which have only a daily motion around the earth; the sun and the moon, which share this daily motion but also have an annual motion (known as "proper motion") around the earth; and the planets which exhibit

the above two motions plus, at times, a retrograde motion. Thus the frame for PATH includes the three attributes *daily motion*, *proper motion*, and *retrograde motion*. Barker notes that for Ptolemaics all these motions are circular, and he embeds a (simplified) frame for circular celestial motions into the frame for PATH at each of the attributes. CIRCULAR MOTION has the following attributes and values: *center*: center of cosmos, other; *radius*: large, medium, small; *speed*: 24-hour, other. This yields the following values for the three motions of Ptolemaic astronomy:

> Daily motion: *center*: center of cosmos, *radius*: large, *speed*: 24-hour;
> Proper motion: *center*: other, *radius*: medium, *speed*: other;
> Retrograde motion: *center*: other, *radius*: medium, *speed*: other.

Copernicus also computes paths and, Barker argues, as long as we confine ourselves to calculational astronomy (ignoring cosmology and physics), Copernicus introduces just two changes into this frame. In the Copernican account daily motions are generated by the earth's rotation on its own axis, and this radius is small. In other words, for calculational purposes the Ptolemaic frame for PATH is retained almost unchanged. Barker maintains that this allowed many astronomers to adopt Copernicus' approach for computing paths, while rejecting his cosmology (2001: 269). Kepler introduced a considerably more drastic change: for proper and retrograde motions he replaces paths with orbits – which are elliptical and governed by a force. The frame that results is the same as the Copernican frame for daily motion, but there is an entirely new set of values for proper and retrograde motions. As a result, resistance to the Keplerian view was considerably greater.

This is very interesting from the perspective of my project. The overall change from a Ptolemaic to a Copernican view involved major changes in the conception of the universe, but on Barker's analysis one aspect of the overall picture remained almost unchanged. Thus, using frames as the basis for his account, Barker has isolated one strand of continuity amidst major change. This is the kind of detailed analysis of conceptual change that I am advocating, and frames can provide a useful tool in carrying out such analyses.[21]

I turn now to work by Andersen and Nersessian who also pursue detailed analyses of conceptual change, and who introduce an elaboration of the frames approach. They argue that frames are adequate for the analysis of concepts such as DUCK, GOOSE, and PLANET whose individual instances can be picked out by ostension. But frames are not sufficient for concepts whose content is at least partly determined by the role they play in natural laws that involve several concepts: "for example, Newton's second law, $\mathbf{F} = m\mathbf{a}$ in which the concepts of 'force', 'mass', and 'acceleration' are simultaneously involved." In these cases we do not pick out instances of the individual concepts, but rather "complex *problem situations* to which a given law

applies" (Nersessian and Andersen 1997: 127, cf. Andersen and Nersessian 2000). Andersen and Nersessian argue that we must add an additional layer to frames in order to accommodate these concepts. This additional layer is derived from earlier work by Nersessian.

In her 1984 book and subsequent papers Nersessian studied the development of the concept of an electromagnetic field from Faraday, to Maxwell, to Lorentz, and to Einstein. The last of these is the current electromagnetic-field concept; Nersessian emphasizes that it may be replaced: it is "the *present* concept and not *the* concept" (1984: 183). In order to clarify the content (Nersessian says "meaning") of each of these concepts, and the relations between them, she introduces the notion of a *meaning schema*. This is a two-dimensional array where one dimension is representational: it provides an account of conceptual content that clarifies continuities and differences between different stages of the concept. The second dimension concerns the cognitive processes involved in constructing new versions. On this dimension versions of a concept are connected by *chains of reasoning*. I am interested here only in the first dimension. Nersessian has refined her account of meaning schemas somewhat since their initial presentation; I will look only at a recent version. Nersessian recognizes four features that should be included in an account of conceptual content: ontological status, function, mathematical structure, and causal power (2001: 282). These are illustrated in the table on p. 252.

I take it that the notions of *mathematical structure* and *ontology* are clear. *Function* is being used in the same sense as *systemic role* in TC; Nersessian notes that ELECTROMAGNETIC FIELD introduced a new function into physics. The notion of *causal power* is less clear. Nersessian tells us that "The *causal power* feature of a concept marks out the problem situations in which the referent of a concept comes into use in order to explain the situation (i.e., the situations that the concept is used to explain)" (2001: 282). The causal power of a referent also includes effects (Nersessian 1984: 157; Nersessian and Andersen 1997: 129). In the example given it seems that the causal power associated with ELECTROMAGNETIC FIELD consists of all causes and effects of the field.

Andersen and Nersessian propose an integration of the frame and meaning-schema accounts. Frames apply to all concepts, but a complete account of concepts that are essentially involved in laws requires the additional layer of analysis provided by the meaning schema. The upshot of this discussion, for present purposes, is that we have another approach to describing conceptual content in sufficient detail to permit analysis of which aspects are changed and which aspects are held constant as a subject develops.

I want to consider one more philosopher, Thagard, who develops an account of conceptual change in science that draws on work from cognitive science.[22] There is substantial overlap between Thagard's approach and the approach in terms of frames, but he stresses two features that add important

*Table 6.1* Changing Meaning of "Electromagnetic Field"

| | *Faraday* | *Maxwell* | *Lorentz* | *Einstein* |
|---|---|---|---|---|
| Ontological Status | Substance (preferred) or state of ether | State of mechanical ether | State of non-mechanical ether | State of space (on a par with matter) |
| Function | Transmit electric and magnetic charges continuously through region surrounding bodies and charges | Same as Faraday, but also transmits light | Same | Same |
| Mathematical Structure | Unknown | Maxwell's equations | Maxwell's equations plus Lorentz force and Lorentz transformation rules | Lorentz transformations with relativistic interpretation |
| Causal Power | All electric and magnetic effects plus charges | Same plus radiant heat, light, etc. | Same minus charge | Same |

Source: adapted from Nersessian 2001: 283

detail to the frames approach: "My proposal then is to think of concepts as complex structures akin to frames, but (1) giving special priority to kind and part-whole hierarchies and (2) expressing factual information in rules that can be more complex than simple slots" (1992: 29; Thagard references in this section are to this book unless otherwise noted). I will examine each of these additions, beginning with the two types of hierarchies that Thagard distinguishes.

Kind-hierarchies are exemplified by sets of subordinate and superordinate concepts. For example: Uranium 235 is a kind (isotope) of uranium, which is a kind of metal, and also a kind of radioactive material. Part-whole hierarchies can be typified by a nucleus, which is part of an atom, and a neutron, which is part of a nucleus. Each type of hierarchy embodies implications. Kind-hierarchies support implications from a concept to all the superordinate concepts in its hierarchy.

> For example, Tweety is a canary, which is a kind of bird, which a kind of animal, which is a kind of thing. . . . Part-hierarchies have different

> inferential properties from kind-hierarchies: because canaries are a kind of bird, and birds have feathers, you can generally infer that canaries have feathers, but you cannot infer that beaks have feathers because beaks are parts of birds.
>
> (7)

The implications supported by part-hierarchies are more difficult to specify because an item that occurs as a part can often exist independently of a particular whole. However, given a neutron that is part of a nucleus, we may infer, for example, that there is at least one proton in its immediate neighborhood, and that the two are bound together by the strong interaction. In a similar way, given an avian beak (that has not been amputated) we can infer that it is attached to a bird. Thagard maintains that the two types of hierarchies he distinguishes are pervasive: "Conceptual systems are primarily structured via kind-hierarchies and part-hierarchies" (7). As a result, Thagard holds, "all scientific revolutions involve transformations of kind-relations and/or part-relations" (7). However, Thagard does not claim either that these are the only types of relations among concepts, or that changes in these hierarchies are the only types of changes that occur in revolutions. Rather, he holds that "kind-hierarchies and part-hierarchies serve to structure most of our conceptual system, providing backbones off which other conceptual relations can hang" (28).

Thagard's second variation on frames concerns rules that license further inferences among concepts. He allows for a wide variety of rules, and does not propose an exhaustive list. We can see some of the kinds of rules that Thagard considers by looking at his discussion of the eighteenth-century revolution in chemistry. (In stating rules I will italicize the key term that Thagard uses to characterize the rule.) In Stahl's phlogiston theory compounds *with* phlogiston burn (41). At one stage in the development of Lavoisier's conceptual system, metals *become* calxes and gain weight when this occurs, while calxes *contain* common air (44). In Lavoisier's mature theory non-metallic substances combine with oxygen to *produce* caloric and light (47). With-rules, become-rules, contain-rules, and produce-rules are examples of rules that specify relations between concepts. In addition, consideration of specific instances of a concept leads to the inclusion of two further types of links: *instance links* typified by the case of Tweety who is an instance of canary, bird, and any other superordinate concept in this kind-hierarchy; and *property links* which relate a specific object to its properties – e.g., Tweety is yellow (31). Thagard also notes that conceptual relations based on relational and higher-order properties pose no special problems (31).

Thagard presents these hierarchies and rules as additions to the structure captured in frames, but it is clear that Barsalou and his followers would include them among the structural invariants. This disparity is probably explained by the publication dates of the relevant texts: Thagard appears to have earlier versions of frames in mind. Still, Thagard's claim that the two

types of hierarchies he distinguishes play a special role in conceptual systems is a substantive thesis, and although he defends it at length in his 1992 book, he also notes some apparent limitations (28). Put in linguistic terms, the two hierarchies apply generally to nouns, but not to adjectives, which are organized into contrast sets of the sort that characterize values in frames. Thagard also suggests that verbs are organized by relations of entailment (driving entails riding) and manner (nibbling is a manner of eating). We may add that mathematical theories involve relations that do not fit easily into Thagard's two hierarchies: consider the relation between force, mass, and acceleration in Newton's second law, or the pervasive relations between the speed of light and other concepts in relativity.

Thagard is especially interested in using his account of conceptual structure for the analysis of conceptual change in science. His discussion includes descriptions of different kinds of conceptual changes, and a response to the question of when we have conceptual change as opposed to change of belief. With regard to the latter issue, Thagard notes that we should not expect to find sharp criteria for conceptual identity:

> It would be futile to try to offer criteria for identity of concepts that attempt to specify when a concept ceases to be the concept that it was. We cannot even give such criteria for mundane objects like bicycles: if I change the tires on my bicycle is it the "same" bike? What if I change the wheels, or the frame, or all of the above?
>
> (34)

Attempts to provide such criteria have not been fruitful in that they have not yielded significant insight into the development of human thought. From the perspective of TC, Thagard can be viewed as proposing a change in the system of concepts we use for thinking about concepts: replacing the thesis that every change in a body of beliefs is either a change of belief or a conceptual change, with the view that some changes are more drastic than others. In many contexts it is more illuminating to think of the less drastic changes as changes of belief, and the more drastic cases as conceptual change. But the major goal is not to solve a (possibly artificial) philosophical problem, but to construct a system of concepts that will help us understand the development of human knowledge – a pursuit that may require changes in a traditional philosophic problematic.

Thagard considers the following sequence of "kinds of conceptual change, roughly ordered in terms of degrees of increasing severity" (34). (The list focuses on additions to a conceptual system but, as Thagard notes, deletions can easily be included).

1. Adding a new instance, for example that the blob in the distance is a whale.
2. Adding a new weak rule, for example that whales can be found in the Arctic ocean.

3. Adding a new strong rule that plays a frequent role in problem solving and explanation, for example that whales eat sardines.
4. Adding a new part-relation, for example that whales have spleens.
5. Adding a new kind-relation, for example that a dolphin is a kind of whale.
6. Adding a new concept, for example *narwhal*.
7. Collapsing part of a kind-hierarchy, abandoning a previous distinction.
8. Reorganizing hierarchies by *branch jumping*, that is, shifting a concept from one branch of a hierarchical tree to another.
9. *Tree switching*, that is, changing the organizing principle of a hier archical tree.

(35)

It will be useful to consider some examples of the more drastic kinds of change, beginning with the fourth; the examples that follow are taken partly from Thagard, partly from our database in Ch. 2. Early in the twentieth century, after Rutherford's discoveries of protons and the nuclear atom, nuclei were believed to consist of enough protons to account for the atom's weight, plus a sufficient number of electrons to cancel any excess charge. As atomic physics developed this led to a variety of problems that were not resolved until the discovery of the neutron in 1932. (See Anderson 1996 for discussion of some central parts of this development.) This resulted in two changes in previously accepted part-relations: inclusion of a new nuclear constituent, plus the banishment of electrons from the nucleus. New kind-relations, item 5, can be exemplified by the discovery of isotopes and isomers, as well as by the distinction between a genetic mother and a birth mother. New kind relations also occur when items once considered distinct are brought together under a single concept, although the old concepts continue to play a useful role; examples include electricity and magnetism in the hands of Maxwell, and mass and energy in special relativity. This is different from cases falling under 7, where a distinction is just abandoned – e.g., the distinction between the celestial and terrestrial realms from pre-Copernican astronomy. We have already encountered many examples of item 6, the introduction of new concepts, such as mass, complex number, and neutrino. Thagard illustrates 8, branch jumping, by the Copernican shift of the earth from the unique member of a special class into the class of planets, and the similar shift of the sun into the class of stars; each of these cases also involved the elimination of a concept (196–97). Other cases cut across Thagard's classification (a point that I do not think he would find objectionable). For example, mathematical generalizations, such as extension of the kinds of exponents and introduction of the gamma function, involve the introduction of a new concept, but in such a way that previously recognized items are now seen as species of the new class. Finally, Thagard illustrates 9, tree switching, by Darwin's change in the meaning of

the classification hierarchy from one based on similarity to one based on historical relations. Another example is provided by the developments in chemistry that led to the abandonment of weight as the key organizing principle of the periodic table of elements. All of these examples, along with others that I have discussed in this section, underline the many ways in which concepts within a system relate to each other, and to successor concepts that are produced as knowledge develops.

## 6.5 The Fine-Structure of Conceptual Content

As is the case with any general theory, TC works at a moderately high level of abstraction; its application in a specific case requires further attention to detail. The idea is familiar from cases such as Newton's second law, where the appropriate force function must be provided in order to apply it to a particular situation, and applications of Schrödinger's equation which require formulation of the appropriate Hamiltonian. An analogous situation obtains when we study a specific conceptual system. Every conceptual system is constituted, at least in part, by implications among its concepts, and a study of these implications is an integral part of any analysis. The discussion in Sec. 6.4 indicates that various implications may be included in a system for different reasons. These differences are captured in the differences between part-hierarchies and kind-hierarchies, the various types of rules that Thagard introduces, the various types of structural invariants that Barsalou recognizes, and other aspects of particular conceptual systems. Study of the underlying bases for the implications included in a system adds important detail to an analysis, and enhances our understanding of that system. This is actually an extension to the implicational dimension of a theme that was included in our discussion of the other two dimensions of TC. We have seen that different *kinds* of instantiation conditions are appropriate for different descriptive concepts, and these details constitute part of the content of those concepts. We have also seen that the content of prescriptive concepts includes such relations to extra-systemic items as injunctions to act in a specific way, prohibitions, and permissions. In addition, these requirements may involve physical acts, such as pressing a brake pedal, or cognitive acts, such as adding a column of numbers or refraining from making an inference. The third dimension of TC, systemic role, was included exactly because different concepts and conceptual systems are generated for different reasons; understanding those reasons is a necessary part of understanding particular concepts. We should also include such fine-structure in accounts of the implicational dimension of concepts. Another look at some examples will illustrate the point.

Suppose that a particular number is the value of a factorial; this implies that the number is also the value of a gamma function, and this implication holds because of the way GAMMA FUNCTION generalizes FACTORIAL. But factorials are not a kind of gamma function since factorials can be fully

understood without any reference to gamma functions. That canaries are birds is essential for an understanding of CANARY. One might suggest that factorials form a subset of the set of gamma functions, but this is not quite correct. Rather, the set of factorials is isomorphic to a subset of the gamma functions. Moreover, if we think of subsets as parts, they are parts in a different sense than that in which a beak is part of a bird.

Now consider two isotopes of an element. Their description as isotopes implies that their nuclei have the same electric charge but different atomic weights. In Soddy's 1913 account we have the further implication that the nuclei have different numbers of protons plus compensating electrons. In a post-1932 framework we have the implication that the nuclei are composed of the same number of protons, but different numbers of neutrons. In both cases the implications are based on part-whole relations, and the differences arise from changed views of the nuclear parts. The more recent change in MOTHER has a different basis. Immediately before the advent of the new reproductive technologies, "Mary gave birth to Pat" implies "Mary provided half of Pat's genes"; this implication no longer holds. To arrive at the same conclusion we require further information about the history of Pat's conception and gestation. Here the change is driven by rejection of a previous (typically unstated) assumption that what we now call the "birth mother" and the "genetic mother" are identical.

Our discussions of frames and meaning schemas illustrate further aspects of conceptual structure. Many conceptual systems are organized hierarchically, and recognizing this point is part of an analysis of concepts in that system. But some of Barsalou's examples illustrate different bases for these hierarchies. For example, while a skiing vacation is a kind of vacation, the need for mountains is an analytic consequence of SKING. The relation between a car and its fuel illustrates a different case. Fuel, valves, and wheels are all necessary for a car to operate, but fuel is not part of the car. The fuel that is in the car at a given time may be numbered among the car's contents, but not in the same sense in which a valise in the trunk is among the contents since the fuel is consumed as the car operates. An operating car also requires a driver who, like the fuel, is a transient part of the contents, but the driver is not used up as the car runs. Thus while CAR implies VALVES, FUEL, and DRIVER, the grounds for these implications are quite different. Moreover, the law-based conceptual systems that Andersen and Nersessian explore are not hierarchical at all.

Thagard suggests another consideration involved in structuring a conceptual system. He notes that all the implications he discusses can be expressed in predicate logic (1992: 31–32).[23] But, he adds, different ways of expressing the same body of information are not always computationally equivalent. Thagard illustrates the distinction in an earlier book (1988: 30–31): if we supplement Roman numerals with a zero we have a system for expressing integers that is expressively equivalent to the Arabic integers. Nevertheless, arithmetic operations can be carried out much more efficiently in the Arabic

system than in the augmented Roman system. Thus in some cases considerations of computations efficiency will help us understand why a conceptual system is built in one way rather than another.

We will encounter further examples of different ways in which implications are generated in the studies that comprise the next four chapters of this book. For the present I want to stress two points: that an account of the basis for these implications is part of a conceptual analysis, and that we should leave the list of possible bases open as we proceed.

## 6.6 Conclusion

We have now arrived at our full working version of TC. The account summarized in Sec. 5.11 stands, but is enriched in one respect: an account of the implications embodied in a conceptual system should include the basis for these implications in part/whole relations, kind hierarchies, and other structures. This is an extension of the earlier recognition – due to Sellars – that implications may be based on first-order or higher-order properties and relations. We have already seen that concepts and conceptual systems can play many different roles, and that different kinds of instantiation conditions are appropriate in different cases. We have now found a comparable richness and flexibility on the implicational dimension. TC provides a basis for rich studies of the contents of conceptual systems and the relations between such systems. Working in this framework we can acknowledge the full range of human conceptual resources in carrying out these studies. I turn now to a set of detailed studies that will both apply and test TC. In Chs 7 and 8 I apply TC to issues in conceptual analysis; in Chs 9 and 10 I turn to studies in the development of physics.

# 7 Conceptual Analysis I: Causation

> The attempt to "analyze" causation seems to have reached an impasse; the proposals on hand seem so widely divergent that one wonders whether they are all analyses of one and the same concept.
>
> (Kim 1995: 112)

## 7.1 Conceptual Analysis

One function of a theory of concepts is to guide the practice of conceptual analysis since our approach to this task depends on our view of conceptual content. We have seen, for example, that in the empiricist tradition only auxiliary concepts are subject to analysis, which consists of resolving these concepts into their basic constituents. C. I. Lewis offers an holistic view in which all concepts are subject to analysis, which consists of mapping out relations between concepts. Philosophers in either of these camps may adopt a necessary-and-sufficient-conditions view of concepts, and analyses will yield a statement of those conditions. Conceptual analyses guided by TC will not result in compact formulas, but in extended accounts whose details depend on the type of concept in question. Some concepts have specifiable necessary-and-sufficient conditions along one or more dimensions, and these conditions will be included in an analysis; for many formal concepts such a statement will constitute the entire analysis. For most concepts, however, the analysis will be open-ended.[1]

Conceptual analysis plays a central role in historical studies of conceptual change since we require accounts of the concepts in question; I will pursue this topic in Chs 9 and 10. Conceptual analysis also plays a central role in philosophy since, on any plausible account of the nature of philosophy, conceptual analysis is an important component. The pedigree of one common approach to conceptual analysis goes back to Plato: Analysis is a reflective, a priori endeavor that can be carried out in the privacy of one's own mind. We proceed by reflecting on instances of a concept, formulating an analysis, and testing the analysis by considering further instances. Often this last step yields counter-examples, and we seek to improve the analysis in a way that neutralizes the counter-examples. The process continues until no

more counter-examples can be found, at which point the analysis is ready for publication – which often results in other philosophers proposing new counter-examples that did not occur to the original analyst.[2] Ideally this process will continue until an analysis is achieved that elicits general agreement from the philosophical community. I want to consider two key assumptions that underlie this view.

First, it assumed that the concept being analyzed already exists in the analyst's mind and plays a role in generating candidate analyses and counter-examples. This view comes with two auxiliary assumptions: a) although analysts already possess the concept, it is difficult to formulate a correct analysis; this accounts for the many failures that fill the philosophical literature. b) We are better at recognizing counter-examples than at formulating correct analyses. Second, it is assumed that analysts who discuss a concept are all examining the same concept, which is why mutual criticism, and thus cooperative analyses, are possible. Thus criticisms of other philosophers' analyses are regularly extended to historical figures such as Descartes and Plato who are often interpreted as giving incorrect analyses of the same concepts that are currently under discussion.

However, once we begin checking our analyses with other people, it becomes unclear why we should think of analysis as an a priori endeavor. Graham and Horgan recognize this point and propose that we recast our understanding of analysis as a "broadly *empirical*, interdisciplinary, enterprise encompassing such fields as psychology, linguistics, social anthropology, and philosophy" (1998: 272). Our intuitions provide data that should be considered empirical, and our analyses are thus defeasible on the basis of evidence from other sources. As these authors recognize, it is debatable whether introspective evidence should be considered empirical (1998: 291, n. 5). Many philosophers who take counter-examples provided by others seriously still consider themselves engaged in an a priori endeavor as long as they retreat into their own minds to assess these proposals. This disagreement suggests that the concepts A PRIORI and EMPIRICAL are not all that clear, but I will not pursue these concepts here. For present purposes it is more important that even if this endeavor is considered empirical, it operates under the assumption that we share *the* concept being analyzed. Goldman and Pust (1998) defend the traditional view that the intuitions which provide the data for conceptual analysis are generated by concepts we already possess, but recognize the possibility of conceptual diversity among analysts, and discuss specific ways in which empirical psychology can contribute to analysis. Still, practitioners of conceptual analysis generally assume that all are discussing the same concept.

Let us ask how this presumed conceptual uniformity comes about. Plato recognized the issue and provided an answer in his doctrine of a pre-birth vision of the forms. While few now find this answer satisfactory, we should acknowledge that Plato attempted to answer an important question. Various doctrines of innate concepts also provide answers to this question, but most

contemporary practitioners of conceptual analysis work in the empiricist tradition, and reject that approach. There is, however, a serious lack of alternative accounts. Perhaps the only systematic attempt is due to Davidson who holds, in effect, that we learn our concepts when we learn our language, and that all languages must be mutually translatable (see Sec. 2.5). Thus, on a fundamental level, there is only one conceptual system. Yet the many examples I have already given, and further examples to be discussed in the rest of this book, undercut this claim of conceptual uniformity. It might be replied that many of the concepts I have considered are irrelevant; that the claim of universality holds only for a small subset of central concepts. (Recall *Parmenides* 130c–d where Socrates denies that everything we can distinguish – even hair, mud, and dirt – has a form.) The obvious rejoinder is to request a characterization of that set, but instead of pursuing this challenge, in this chapter and the next I will focus on concepts that have been the subject of many attempts at analysis. These, if any, are members of the core set. Moreover, since debates on these concepts typically assume that all are engaged in analysis of the same concept, it remains appropriate to ask for an account of how even this limited uniformity comes about.

Disagreement on the correct analysis of philosophically interesting concepts is a pervasive feature of the philosophical literature (see Brown 1999 for an extended discussion). While disagreement keeps conceptual analysis alive as a research endeavor, it also indicates that analysis is a task at which the vast majority of professional philosophers fail most of the time. But the existence of pervasive disagreement is susceptible to a different explanation: Perhaps the philosophers who disagree were not all discussing the same concept, and successful analyses are more common than they seem on the usual view. How should we decide between these two views? I submit that the assumptions identified above serve as GAs of the standard approach: These assumptions are synthetic claims, and failures of analysis could be interpreted as evidence against them; instead, the assumptions are maintained, and guide research by imposing the task of finding a new analysis that is immune to the counterexamples. I want to emphasize that this is *not* dogmatism; it is the customary mode in which we carry out research. Still, such research sometimes leads to the conclusion that a set of GAs has outlived its usefulness (and may even be false), and that its replacement by another set is in order. One of my aims in this chapter and the next is to argue for such a replacement. I will pursue this aim by examining examples of analyses of some key concepts – the causal relation in this chapter, and the cluster of concepts that are central to epistemic analysis in the next – and arguing for the fruitfulness of a different way of thinking about conceptual analysis. One effect of the alternative approach will be to bring a new issue into focus: If different people have different concepts for thinking about some subject matter, it is reasonable to ask if some of these concepts are preferable, so that some of us should abandon current concepts and adopt different ones. Perhaps none of the currently available concepts are adequate. I will develop and defend this proposal as we proceed.

## 7.2 The Causal Relation

I will be concerned here with the concept of a causal *relation*, not with the concept of a cause or an effect, although study of this relation will require consideration of its relata. However, I am not going to propose an account of this relation. Instead, I will pursue three other aims. I will pursue two of these aims simultaneously: documenting the massive disagreement among analysts who have addressed this topic, and using TC as a guide for organizing the accounts and clarifying some of the concepts that philosophers have associated with the term "causal relation." In addition, I will consider what is at stake in choosing among these accounts. I will limit discussion to causal relations between items in the physical world, leaving open whether other kinds of items exist. Moreover, I will discuss only a selection of the massive literature on causation and of the disagreements among those who claim to be analyzing a single concept.

### *7.2.1 Implications*

In this section I am going to examine various views of what is implied by propositions of the form "*x* causes *y*," which I will symbolize *xCy*. These intra-systemic relations depend, in part, on which concepts are included in the system we are examining – a contested issue. I postpone discussion of this topic until Sec. F, although I will not be able to avoid making some assumptions as we proceed. I will also have to make some working assumptions about the nature of the causal relata, although this is another subject of dispute. I postpone systematic discussion of these relata until Sec. D. In discussing implications we will encounter cases in which *x* (*y*, *z*, etc.) stand-alone as a premise or conclusion of an argument. Since premises and conclusions of arguments are propositions, *x* standing alone should be read as "*x* occurs," in a timeless sense. I will use "*not*-" to negate propositions: "not-*x*" standing alone reads "*x* does not occur"; "not-(*xCy*)" reads "It is not the case that *x* causes *y*"; and so forth. We will also encounter cases in which it is asserted that a causal relatum does not occur; I will use "-" for this purpose. For example, "*xC*-*y*" says, "*x* causes the absence of *y*." The expression "*x*&*y*" standing as premise or conclusion reads "*x* and *y* both occur," while "*x*&*yCz*" says "*x* and *y* together cause *z*." I will omit quotation marks except where required for clarity. Following Mackie (1980: 51) I will describe a cause as *causally prior* to its effect, leaving consideration of temporal relations between cause and effect open for further discussion.

#### *A.* Sufficient Condition

According to one common view, *xCy* implies that *x* is a sufficient condition for *y*, although this implication holds only *ceteris paribus*. For example, dropping a brick on my naked toe is sufficient to cause pain under typical

conditions, but not in special circumstances, such as when my toe is anesthetized. This restricted notion of a sufficient condition is captured in the combination of the *validity* of C1, and *invalidity* of C2:

$$xCy, x \therefore y \tag{C1}$$

$$xCy \therefore x\&zCy. \tag{C2}$$

Several features of this notion of a sufficient condition need clarification.

Consider Mackie's account of a cause as an INUS condition: "an *insufficient* but *non-redundant* part of an *unnecessary* but *sufficient* condition" (1980: 62). Mackie's main concern is different from mine: he is proposing an account of what we typically pick out as the cause of some outcome (1980: 64). In doing so, he takes for granted the notion of a complete set of conditions that will guarantee an outcome (cf. Mill 1868: 365–73), and argues that we typically select only a specially important part of this condition as the cause. (See Hanson 1958, Ch.3 and Miller 1987: 86–98 for similar views.) In my usage $x$ stands for the complete cause.

The details of what we should include in $x$ is a complex matter that must be determined empirically. A sufficient condition for a fire is more than just the presence of flame, flammable material, and oxygen; it also requires appropriate relations among them – especially spatial and temporal relations. A flame that occurred yesterday, or in a distant part of the galaxy, will presumably not cause the paper before me to ignite. But how hot the flame must be, and how close it must be to the paper, depend on the nature of the paper. If the flame is hot enough, and the paper has a low enough ignition point, the flame may be several inches from the paper and still ignite it. If the flame is provided by an ordinary match and I am considering heavy construction paper, ignition may require that the paper be in the flame. Some hold that causation always requires spatial and temporal proximity, so that a flame held at a distance from the paper is not the actual cause of ignition, but a prior step in a causal chain; however, this is a contested issue. Classical mechanics – under one interpretation – involves action at a spatial distance (cf. Salmon 1984: 209–10; Suppes 1970: 84–86), and some argue that recent developments in quantum theory *require* action-at-a-distance (e.g., Salmon 1984: 245–50). In addition, unless cause and effect are simultaneous there will be some time gap between them (cf., Tooley 1987: 210–12). The possibility of simultaneous causation is another contested topic; it is discussed in Sec. C. One might require that an effect occur in the "next instant" after its cause, but the acceptability of this claim depends on one's view of time. For example, if time is continuous (or even compact) in the modern mathematical sense, then the phrase "next instant" has no meaning. Tooley (1987: 235) insists that our causal concept does not require either spatial or temporal contiguity. Suppes maintains that while the concept requires temporal continuity, when we consider "the framework of fundamental

beliefs about the general character of the universe . . . " (1970: 31), there are many contexts in which we make use of a causal concept that does not include this requirement. "The concept of remote direct causation is a tool . . . [that] is essential for practical and scientific analysis of many sorts. Its usefulness will not disappear in the foreseeable future in disciplines ranging from political history to meteorology" (1970: 32).

The invalidity of C2 captures the *ceteris paribus* clause: once we have established that $x$ is sufficient for $y$, any addition to $x$ blocks the *implication*. In the fire example it is an empirical matter whether a specific addition – say, water, carbon dioxide, or alien intervention – actually prevents ignition. Fire may still occur, but we now need additional premises to justify inferring the conclusion. The conjunction in C2 is also worthy of further discussion, but I will note just two points. First, when we add water or carbon dioxide to the mix, there is some interaction among the elements that prevents ignition, but interaction is not always required. Placing a two-pound block on a scale causes the scale to read 2, but if we add a three-pound block beside the original block, the pointer will no longer read 2 even if the blocks do not interact. Second, causal conjunctions are not commutative. For example, the result of adding water to sulphuric acid is different from that of adding sulphuric acid to water.

Use of a *ceteris paribus* clause eliminates any need to describe negative conditions, such as the absence of water, which must be met in specifying a sufficient condition. Burks disagrees: "By 'sufficient conditions' we mean a set of conditions, complete with respect to negative properties as well as positive ones (i.e., counteracting causes must be explicitly mentioned) sufficient to cause the state of affairs expressed by the consequent" (1951: 368). Yet there is no limit to the range of possibly relevant negative conditions, so it seems preferable to establish a set of factors that is sufficient for $y$ to occur *ceteribus paribus*, and recognize that any alteration in this set blocks the implication.[3] Ducasse has a different objection. He holds that we must sharply distinguish between a cause and the conditions in which it occurs. However, since both are required, Ducasse concludes that causation is a triadic relation with "circumstances" providing the third term (1926: 58–59, 1951: 145–46).

The status of SUFFICIENT CONDITION as part of the content of CASUAL RELATION has been challenged by advocates of probabilistic accounts of causation who hold that this requirement is the hallmark of DETERMINISTIC CAUSATION, which is just a special case. Suppes, for example, argues that identification of causality and determinism was a result of the success of Newtonian physics:

> The overwhelming empirical success of Newtonian mechanics, particularly in accounting for the motions of the solar system, inevitably yoked the notions of causality and determinism. In the heyday of classical mechanics in the nineteenth century, it was impossible to talk about causes without thinking of them as deterministic in character.
>
> (1970: 6)

But, Suppes continues, this is "a mistaken notion of causality." In everyday conversation we use "cause" in a "rough and ready sense" to describe "partial relations." When we say, "His reckless driving is bound to lead to an accident" (1970: 7), we mean only that there is a high probability of his having an accident in which his driving will be a part cause. We do not mean that an accident is inevitable every time he drives. Suppes adds several examples of such everyday talk which he takes to express the proper meaning of "cause" – the meaning that he formalizes in the text that follows.[4] Suppes maintains that *positive statistical relevance* (PSR) – an increase in the probability of the effect as a result of the cause – is required for a causal relation. That is, $xCy$ implies:

$$Pr(y|x) > Pr(y). \qquad \text{(PSR)}$$

Determinism is the special case in which $Pr(y|x) = 1$.[5] However, PSR is not a complete account of $xCy$ since it follows by probability theory alone that PSR implies that $Pr(x|y) > Pr(x)$. Without an additional condition we would have $xCy$ if and only if $yCx$. Suppes adopts the additional condition that the cause must precede the effect in time. Humphreys (2000: 35) mentions another version of probabilistic causation that requires a boost in probability across a wide variety of situations: we must have $Pr(y|x\&z) > Pr(y|\text{-}x\&z)$. Among other issues, however, Humphreys notes that what we should include in $z$ depends on which interpretation of probability is adopted.

Salmon, who defends a different probabilistic account of causation, also cites examples from everyday experience and science to show that a causal relation does not require a sufficient condition: "it seems altogether unnecessary to burden our common sense concept of causality with the dubious metaphysical thesis of determinism" (1984: 189).[6] Consider one example that Salmon uses to illustrate a situation that is both statistical and causal (1984: 186–88): The atoms of a laser are in an excited state. When a photon of the correct frequency impinges on the laser these atoms drop to their ground state and emit a burst of light. Salmon takes it as clear that the impinging photon caused the laser burst, but did not act as a sufficient condition because under identical conditions the same photon need not have been followed by the burst. Whether the burst occurred is irreducibly statistical; the impinging photon greatly increased the probability of the burst, but did not guarantee it. Still, there are cases in which emission of a laser burst was the result of the impinging photon, and in these cases, Salmon maintains, it is appropriate to say that the photon caused the burst.

Salmon agrees with Suppes that probabilistic causation is our basic concept and that the situation in which the cause provides a sufficient condition is a limiting case (1984: 190). Discussing an example from the molecular theory of gases Salmon writes: "To most nineteenth century kinetic theorists, the causal interactions [between gas molecules] were strictly deterministic, but we can cheerfully admit that they may actually be irreducibly statistical.

Our causal concepts admit irreducibly statistical features without any strain" (1984: 228). But Salmon rejects Suppes' detailed account because (Salmon argues) that account allows some inappropriate sequences to count as causal while eliminating some genuine causal relations from the class (see 1984: 192–94 for details). In particular, Salmon rejects PSR. One of Salmon's examples concerns a golfer who gets a hole-in-one as a result of the ball hitting a tree. In general, hitting a branch will not increase the probability of a hole-in-one. After discussing three alternative ways of analyzing this case (1984: 193–202) Salmon concludes that "we must give serious consideration to the idea that a probabilistic cause need not bear the relation of positive statistical relevance to its effect" (1984: 202). At present there is a substantial literature on whether PSR is required for causation; Dowe (2000: 33–40) provides a recent review. Some even argue that there are conditions in which a cause may lower the probability of an outcome – that is, $Pr(y|x) < Pr(y)$. This can occur, for example, when there are multiple causes with different probabilities of producing the effect, and the actual cause blocks the occurrence of a more effective cause (Davis 1988: 140–41; Dowe 2000: 33–40). Salmon proposes an alternative approach: "If positive statistical relevance is not the essential ingredient in a theory of probabilistic causality, then what is the fundamental notion? The answer, it seems to me, lies in the transmission of probabilistic causal influence" (Salmon 1984: 202). Tooley (1987: 251) holds a similar view: "causation is that theoretical relation that determines the direction of the logical transmission of probabilities." Keep in mind that this is offered as an explication of what we ordinarily mean by "causation."

Miller puts a somewhat different spin on the probabilistic approach when he contrast a common understanding of a cause as a *trigger* with determinism which he describes as a philosophical add-on:

> Together with the everyday implication that a cause is a trigger, there is a philosophical assumption that needs to be cancelled – that if something causes an event, then it made the event inevitable under the actual circumstances; given the cause and its actual background, the sequel could not have been otherwise. This assumption that all causes are deterministic was never part of the everyday causal analysis, where the turn of the honest croupier's hand causes red to come up on a roulette wheel that stops at red by chance. . . . Also, this deterministic assumption is no longer part of physics, where a dynamical event is typically attributed to an antecedent total state that need not have had the event as its sequel.
>
> (1987: 61)

While Miller, Suppes, and Salmon spend a good deal of effort arguing that probabilistic causation is not only coherent, but our basic causal notion, some advocates of probabilistic causation do not feel the need for this kind of justification. Eells, for example, begins his study thus:

> In the past 30 years or so, philosophers have become increasingly interested in developing and understanding probabilistic conceptions of causality – conceptions of causality according to which causes need not necessitate their effects, but only, to put it very roughly, raise the probabilities of their effects.[7]
>
> (1991: 1)

Although the case of deterministic causation is mentioned in the book, the topic is not considered sufficiently important to rate an entry in the index.

### *B.* Necessary Condition

Some hold that a cause is a necessary condition of its effect, but we must be careful about our terminology. First, we must distinguish this claim from the claim that the *causal relation* is a kind of necessary relation (Sec. 7.3). Consider the material conditional, $p \supset q$. Material conditionals do not express a necessary connection between the $p$ and $q$, but $q$ is a necessary condition for $p$ since the additional premise not-$q$ allows us to infer not-$p$. Second, when Eells writes that "causes need not necessitate their effects" he is discussing a sufficient condition – a condition that guarantees the effect – and his point is that causes need not be sufficient conditions.

It will be useful to consider another feature of the material conditional: $p$ is a sufficient condition for $q$ since the additional premise $p$ allows us to infer $q$. In this case the claims "$p$ is sufficient for $q$" and "$q$ is necessary for $p$" are equivalent; this equivalence also holds for logical implication, which is the paradigm case of a necessary connection. But Sanford argues that this equivalence does not hold in causal cases. For example, while light is causally necessary for grass to grow, grass growing is not casually sufficient for producing light (1995: 82). To be sure, if grass is growing we may *infer* that light has played its usual role, and the fact that light is causally necessary for grass to grow provides a required premise for this inference. But the conclusion that light has occurred in the neighborhood is not a causal claim.

Sanford's example illustrates a class of cases in which there is an item that must be included in any causally sufficient condition for a particular outcome, but is not itself sufficient for that outcome. Consider the disease *shingles*: Presence of the virus herpes zoster is necessary for the occurrence of shingles – no one gets shingles without harboring this virus. But many people harbor the virus without suffering the disease, so the virus is not causally sufficient for shingles. Using a circumflex to indicate a condition that is necessary in this sense, and $x$ to indicate any other elements in a sufficient condition for $y$, we have the following implication:

$$x\&\hat{z}Cy, -\hat{z} \therefore -y. \qquad \text{(C3)}$$

This implication also covers cases in which there are multiple sufficient conditions for an outcome, but one or more elements that must be included in every sufficient condition.

C3 captures a special case, but some philosophers hold that a cause is always a necessary condition. They maintain, that is, the validity of:

$$xCy, -x \therefore -y. \tag{C4}$$

Mackie provides one example: He considers cases in which there are multiple sufficient conditions for an effect, none of which are necessary. He calls the disjunction of all these sufficient conditions the "full cause," which provides a necessary condition necessary for the effect in the sense of C4 (1980: 64).[8]

Many view causes as both necessary and sufficient for their effects (without going Mackie's disjunction route). Sometimes Hume seems to hold this view. In the first *Enquiry* he defines a cause as "*an object, followed by another, and where all the objects similar to the first are followed by objects similar to the second*. Or in other words *where, if the first object had not been, the second never had existed*" (1975: 76). On this view, both C1 and C4 are accepted as valid. Blanshard adopts a version of this view when he defines a cause as "the sum of conditions given which the effect will occur; and in the absence of any of which it will not occur" (1962: 457). Typically, for Blanshard, a cause is a complex sufficient condition in which C4 holds for each element of the complex. Taylor (1963: 296–303, 1966: 26–31) defends a similar view. However, Blanshard and Taylor both identify the view that a cause is a necessary condition with the view that a causal relation is a necessary connection. A Humean can accept the validity of C4 while holding that it just expresses a regularity.

In spite of the considerations mentioned earlier in this section, some philosophers defend the equivalence of "$x$ is causally necessary for $y$" and "$y$ is causally sufficient for $x$." In discussing this view it will be help to distinguish two arguments:

$$xCy, -y \therefore -x \tag{C5}$$

and

$$xCy \therefore -yC-x. \tag{C6}$$

Presumably, any view that takes a cause to be a sufficient condition of its effects will consider C5 valid, but the conclusion of this argument is not a causal claim. Rather, the argument expresses a point about *evidence*: Given $xCy$ and the absence of $y$, we may conclude that $x$ did not occur. The conclusion of C6 is a causal claim, and this argument seems to be invalid. In addition to Sanford's example, consider a case in which there is an uncrushed box of crackers on the table in front of me. An elephant sitting on the box would crush it, but the uncrushed box did not cause the absence of any

sitting elephants – or of any sitting kangaroos, meteorite impacts, momentary large increases in the local gravitational field, and so on, *ad nauseam*. It seems, then, that the absence of an item allows us to infer the absence of any of its sufficient conditions, but does not allow us to infer that this absence *caused* the absence of those sufficient conditions.[9] Burks (1951: 369) and von Wright (1993: 113–14) disagree. Each of these philosophers constructs a formal account of causation in which C6 is a theorem. In Suppes' formal account probability considerations alone yield C6: given that $Pr(y|x) > Pr(y)$ it follows that $Pr(\text{-}x|\text{-}y) > Pr(\text{-}x)$; but Suppes rejects C6 on the basis of temporal considerations (1970: 53–54). Burks and von Wright do not consider temporal considerations to be determinative. Burks seeks a logic of causation that does not include any temporal characteristics since he intends that it apply to causal laws, such as Ohm's law (369) – although he notes that his account seems to clash with ordinary usage. Discussing the inference from rain causing someone to wear a raincoat, to the claim that absence of a raincoat causes the absence of rain, Burks acknowledges that not wearing a raincoat has no causal influence on the weather. But, he adds, "if he does not wear a raincoat we can infer on causal grounds from the given premise that it won't rain" (369). Yet this is just to shift from the causal inference C6 to the evidential inference C5. I will postpone von Wright's reasons for rejecting a temporal condition until Sec. E.

Suppes accepts the validity of:

$$xCy, \therefore -xC-y \qquad \text{(C7)}$$

If $Pr(y|x) > Pr(y)$, it follows from probability calculus that $Pr(\text{-}y|\text{-}x) > Pr(\text{-y})$ (1970: 53–54), and Suppes' temporal condition is met.[10] Suppes notes that the interpretation of this result "may bother some" (1970: 55) but illustrates the point by considering a case in which exposure to measles causes children to become infected, and we explain the fact that a particular child did not get measles by the lack of exposure (1970: 54). This may seem a special case, similar to that of shingles, rather than a general feature of causation, but according to Suppes' account, such discomfort should be overridden. Alternatively, one could use this case to challenge either the probabilistic approach to causation, or Suppes' version of that approach. Salmon, we have seen, takes the latter tack, holding that a probabilistic view of causation need not require PSR. Salmon also argues that causation does not involve a necessary condition. In effect, Salmon rejects the validity of C4 and C7 (we saw in Sec. A that he rejects C5 and C6): In a laser a burst of light sometimes occurs spontaneously, without an impinging photon, even though a photon causes the burst in other cases.

### *C.* Temporal Implications

We encounter major controversy when we consider which temporal relations, if any, are implied by causal claims; I will consider a few examples

from a large literature. Hume (2001: 54), Suppes (1970), and many others, hold that a cause must precede its effect in time, but this has been challenged for many reasons.[11] One common challenge is from those who hold that there are cases in which a cause and its effect are simultaneous; these include Gasking (1955: 479), Papineau (1985: 273) and Salmon (1984: 182). Brand (1980: 137–53) goes further, arguing that cause and effect must be simultaneous (see Tooley 1987: 210–12 for a critique of Brand's arguments). This view was once defended by Taylor (1963: 305, 311), although he later concluded that it is impossible for *all* causes and effects to be simultaneous (1966: 35–39). Kline (1980) criticizes several purported examples of simultaneous causation and provides reasons for doubting that simultaneous causation occurs. Kant is especially interesting since he requires temporal sequence for causation, but also holds that "The great majority of efficient natural causes are simultaneous with their effects, and the sequence in time of the latter is due only to the fact that the cause cannot achieve its complete effect in one moment" (1963: 228). The passage suggests that Kant can reconcile these views because he considers the causal relata to be extended in time: The complete cause is temporally prior to its effect, but the final phase of the cause is simultaneous with the initial phase of the effect. I will examine disputes about the causal relata in Sec. D.

Many adopt a third alternative: $xCy$ at least implies that $y$ does not precede $x$ in time. However, this view has also been challenged on several grounds. For example, Mackie (1980: xiv, 161–66) holds that, as a matter of fact, our causal concept does not include any temporal implications. Thus it is conceptually possible for a cause to follow its effect, although Mackie holds that this never actually occurs. Cartwright (1983: 32–33) holds a similar view in the context of a probabilistic account. Dummett (1954, 1964) holds that while our ordinary concept of causation does not admit of backwards causation, we can introduce a closely analogous concept that he dubs "quasi-causation" which does allow for causes that follow effects in time. It is then an empirical question whether quasi-causation has instances. Tooley holds that no temporal implication *should be* built into the concept of causation, and introduces another consideration: He defends a causal analysis of time, and thus requires that we define causal relations before we introduce temporal relations (1987: 178–81, 190–94, *et passim*).[12] Many philosophers, including Reichenbach, Salmon, and Papineau, agree. David Lewis (1973: 566) rejects the claim that a cause must precede its effect on the multiple grounds that backwards and simultaneous causation are "legitimate physical hypotheses" and that including a temporal condition in a our concept of causation will trivialize causal analyses of time.

Some argue that backward causation actually occurs. For example, von Wright maintains that certain kinds of actions – raising my arm is an example – "may have necessary, and also sufficient, conditions in antecedent neural events (processes) regulating muscular activity" (1971: 76). But these neural items are caused by my arm raising, which follows them in time. "This

is causation operating from the present towards the past. It must, I think, be accepted as such" (1971: 77). The last sentence of this passage overrides the "may" of the previous quote. More recently, Dowe defends backwards causation as providing the best explanation of the surprising correlations that result from quantum entanglement (2000: Ch. 8). All of this is in spite of Taylor's insistence that "there is something metaphysically absurd . . . in supposing that efficient causes might work backwards" (1963: 306), and that:

> This metaphysical way of conceiving these relationships seems, moreover, to be the way all men do think of causes and effects, and it explains the enormous absurdity in the assumption that causes might act so as to alter things already in the past.
>
> (1963: 308)

Those who deny that a cause must precede its effect in time require some other criterion for causal priority. Naturally, there are several proposals in the literature.[13] For example, von Wright gives an account in terms of our ability to use the cause as a means of bringing about or preventing the effect:

> I now propose the following way of distinguishing between cause and effect by means of the notion of action: $p$ is a cause relative to $q$, and $q$ an effect relative to $p$, if and only if by doing $p$ we could bring about $q$ or by suppressing $p$ we could remove $q$ or prevent it from happening.
>
> (1971: 70)

Mackie views von Wright's account as a first approximation to an adequate account of causal priority, but is uncomfortable with its dependence on the concept of agency (1980: 190). His own proposal depends on the point that one item may be *fixed* when another item is not. Given two items, $X$ and $Y$, such that $X$ is an INUS condition for $Y$, "the basic requirement for the judgement that $X$ caused $Y$ is met" (1980: 190), but we still must establish the causal priority of $X$ with respect to $Y$; this requires that there was a time at which $X$ was fixed but $Y$ was not fixed. Note that no particular temporal ordering is required. After adding a few minor refinements, Mackie concludes: "This, then, or something like this, is our concept of causal priority" (1980: 190–91).

Salmon suggests another approach, which is implicit in his account of the difference between genuine causal interactions and interactions that only appear to be causal. Consider two bright spots on a wall that are produced by a pair of spotlights, with spot $A$ to left of spot $B$. If we move $A$ to the right until it touches $B$, and then immediately move $B$, it might seem that $A$ caused $B$ to move – especially if the sequence is repeated. Yet there is no causal interaction between $A$ and $B$. Building on work by Reichenbach,

Salmon proposes the following criterion as a basis for the distinction (1984: 147–57): A causal process is capable of transmitting a mark – that is, a structural alteration – from cause to effect. For example, if we put a red filter on the light that causes *A*, then spot *A* becomes red and stays red as it moves around the wall. The direction of mark-transmission is from the light to the spot, and the beam will be found to be red at every point between the filter and the wall. It is an immediate consequence that the direction of causal influence is the same as the direction of mark transmission. This criterion is completely independent of time. If an effect precedes its cause in time, then the mark would be transferred from the present to the past. Salmon's discussion has elicited a number of related proposals that I will not describe; Dowe (2000: Ch. 8) discusses several of these.

PSR accounts of causation require an additional condition to distinguish cause from effect since, as noted in Sec. A, if $x$ raises the probability of $y$, then $y$ raises the probability of $x$. Suppes invokes time for this additional role, so that on his account $xCy$ implies that $x$ precedes $y$ in time. Dowe, Salmon, and Tooley (among others) develop probabilistic accounts of causation that reject both the PSR requirement and the appeal to time as a means of determining causal priority, for which they all need some additional criterion. I have already described how Salmon meets this challenge. Tooley provides three accounts (that I will not describe in detail); which account we accept depends "on what view one takes on the relation between causal laws and causal relations among states of affairs" (1987: 254).

It will be useful at this point to pause and consider just what is at stake in this debate. If we are analyzing a common concept, then it is a matter of fact whether this concept includes any temporal implications. If our common concept does include such implications, but scientific research identifies cases that are best explained by invoking different temporal relations, we can withhold the label "causation," but this will not make the cases go away. Under these circumstances we should conclude that causation is not universal. We might then introduce a new concept, such as Dummett's quasi-causation, to cover the problematic cases, or perhaps introduce a drastically different concept to cover *all* cases. These are questions for ongoing research and theory development. At one point Dowe argues that we should not include a temporal implication in our causal-relation concept because the occurrence of backwards causation is a scientific matter that should not be ruled out a priori (1995: 322). But refusing to apply a concept is the only sense in which our current causal concept may make backwards causation impossible a priori; this no more constrains nature than does the common concept of a ghost. Recall van Fraassen's remark (Sec. 3.6) about the course of science resulting in a particular concept of matter becoming irrelevant; the same may occur for any other concept. In a similar way, if philosophical reflection leads to the conclusion that the direction of time is best specified in terms of the direction of causation, then we will have a significant argument for adopting an atemporal causal concept, whatever concept currently exists in

common sense. Salmon is proceeding along these lines when he writes, "Our ordinary causal language is infused with temporal asymmetry, but we should be careful in applying it to basic causal concepts" (1984: 176), and adds "I am trying to develop causal concepts that will fit harmoniously with a causal theory of time" (1984: 176, n. 14). We must be clear that this is a different task from analyzing concepts that are already available. It is an open possibility that developments may take us so far from the historical use of causal language that continued use of this terminology becomes misleading.

### *D.* Causal Relata

Views on the causal relata are often closely related to views on other matters, and provide another major area of disagreement. Consider the widely held view that the causal relata are events. There is dispute on whether these are event *tokens* or event *types*; views on this question are often tied to views concerning the relative priority of causal laws and causal interactions between specific items. One common view is that causal laws are conceptually prior and that we attribute causal relations to individual items only insofar as they exemplify causal laws. But laws hold fundamentally between event types, so claims about causal relations between specific events presuppose claims about these types. Others maintain that causal relations hold between event tokens, that causal laws are generalizations over these cases, and that laws are not required for causal attributions. Ducasse, for example, describes causation as:

> a relation between two concrete, individual events and a set of circumstances: the definition of the relation does not employ the notion of collections or kinds of events.
>
> Accordingly, if the requirements specified by the definition are really met by the relation between two concrete events in a given case, then the two events concerned really are cause and effect even if each of them should happen to be completely unique in the history of the universe.
>
> (1951: 150)

Tooley (1990) also defends a singularist view with events as the relata;[14] so does David Lewis – although Lewis denies that events are the only things that enter into causal relations (1973: 558). Others deny that we must choose between event types and event tokens as the causal relata. Suppes account involves "a deliberate equivocation in reference between events and kinds of events" (1970: 79), so that his formalism should apply in both cases. Eells argues that both type and token causation occur, but they are distinct subjects: "the explication of probabilistic causation at the two levels requires quite different kinds of theories, involving quite different concepts, though there are interesting analogies between the two theories . . . " (1991: 16).

Hume's view seems to be a hybrid in which causal relations hold between specific events, but we can attribute causation only insofar as these events exemplify regularities; yet we learn the regularities from the specific instances. On one hand, 'there are no objects, which by the mere survey, without consulting experience, we can determine to be the causes of any other; and no objects, which we can certainly determine in the same manner not to be the causes" (2001: 116).[15] The relevant experience is the familiar constant conjunction: "There must be a constant union betwixt the cause and effect. 'Tis chiefly this quality, that constitutes the relation" (2001: 116). In the absence of a constant conjunction, no causal claim can be justified, but tokens are transient; only types can be *constantly* conjoined. Still, our awareness of a constant conjunction arises from observation of individual cases, and once we are aware of a constant conjunction we attribute causal relations to individual events.

For those who include event tokens among the causal relata, detailed positions vary with varying accounts of events. Ducasse defines an event as "either a change or an absence of a change (whether qualitative or relational) of an object" (1926: 58). Lewis gives the following partial list of examples of events: "flashes, battles, conversations, impacts, strolls, deaths, touchdowns, falls, kisses, and the like" (1973: 558). Davidson, who has an especially rich notion of event-tokens, tells us that "we must distinguish firmly between causes and the features we hit on for describing them . . . " (1974: 194). We must not make the mistake of "thinking we have not specified the whole cause of an event when we have not wholly specified it" (1974: 195). In response to the claim that striking a match is only a part cause of its lighting, Davidson writes:

> It cannot be that the striking of this match was only part of the cause, for this match was in fact dry, in adequate oxygen, and the striking was hard enough. What is partial in the sentence "The cause of the match's lighting is that it was struck" is the *description* of the cause; as we add to the description of the cause, we may approach the point where we can deduce, from this description and laws, that an effect of the kind described would follow.
>
> (1974: 195)

Davidson then adds another example: "If there was an event that was a drying by Flora of herself and that was done with a towel, on the beach, at noon, then clearly there was an event that was a drying by Flora of herself – and so on."

Blanshard denies that events are causal relata, holding that causal relations occur only between kinds (1962: 441). Menzies takes it as clear that type-level causation occurs between properties (1989: 59) and examines various views of the relata at the level of token causation. His list includes propositions, physical objects, events, event aspects, facts, states of affairs, and situations (1989: 59–62). In a list of candidates for the causal relata that

partially overlaps with Menzies', Sanford includes substances and persons (1995: 79, see also Humphreys 2000: 31–32). Salmon maintains that any account of causation in terms of events, whether tokens or types, is "profoundly mistaken" (1984: 138–39), and gives an alternative account in terms of items that are not on either Menzies' or Sanford's lists. According to Salmon, a proper account requires two basic concepts, both familiar to common sense: causal propagation (i.e., transmission over space and time) and causal production. Underlying both of these is the concept of a process, and Salmon takes "processes rather than events as the basic entities" (1984: 139).[16]

> The main difference between events and processes is that events are relatively localized in space and time, while processes have a much greater temporal duration, and in many cases, much greater spatial extent. In space-time diagrams, events are represented by points while processes are represented by lines. A baseball colliding with a window would count as an event; the baseball, traveling from the bat to the window, would constitute a process. The activation of a photocell by a pulse of light would be an event; the pulse of light, traveling, perhaps from a distant star, would be a process. A sneeze is an event. The shadow of a cloud moving across the landscape is a process.
>
> (1984: 139–40)

Salmon includes physical objects persisting through time – including a material object at rest (1984: 140) – among processes, but notes that not all processes are capable of entering into causal relations – as the shadow example shows. Salmon also rejects any attempt to analyze processes as chains of events (1984: 156–57): processes are continuous, events are time-slices of process. We should, Salmon tells us, follow Venn and replace the image of a causal chain made up of distinct events with the image of rope (1984: 183). While it is widely held that relativity theory requires an event ontology, Salmon rejects this view and argues that relativity can also be developed taking processes as the appropriate relata (1984: 140–41). Salmon notes that it might be possible to rework his account of causation in terms of events, but asks why we should bother since taking processes as basic eliminates a central problem of the causal chain view: determination of what connects the events on a chain (1984: 183, also 56–57, 147, 155). Salmon also suggests that classical physics requires a physical-thing ontology (1984: 140). Unfortunately, his remarks on this topic are brief and it is not clear if this is a special case of a process ontology, or an entirely different view. If the latter, Salmon's account implies that it is a view that has been superseded by twentieth-century physics.

Token-level causation plays a fundamental role in Salmon's account:

> A causal process is an individual entity, and such entities can transmit causal influence. An individual process can sustain a causal connection

> between an individual cause and an individual effect. Statements about such relations need not be construed as disguised generalizations.
>
> (1984: 182)

But this is not quite the entire story because of *conjunctive forks*. These arise when two distinct items, $x$ and $y$, have a common cause $z$. In such cases there will be a high correlation between occurrences of $x$ and $y$, although there is no causal relation between them. When we explain this correlation by means of a common cause, "we are implicitly making assertions concerning statistical generalizations. Causal relations, it seems to me, have both particular and general aspects" (1984: 182).

Views on the number of causal relata also vary. It is commonly held that causal relations are binary – say, between two events or two states of affairs – and I have assumed this in my notation. But I noted in Sec. A that Ducasse considers the causal relation to be triadic. Eells argues that type-level causation requires a four-term relation holding among "a cause factor, an effect factor, a token population within which the first is some kind of cause of the second, and, finally, a kind (of population) that is associated with the given token population" (1991: 22). At least, this is required if we are to tie causation to increase in probability. If we leave out the last two factors we will find cases in which causation is not associated with an increase in probability (see the examples that Eells discusses on 2–3). I have already noted that Salmon and Dowe, among others, defend probabilistic accounts of causation that reject the any tie between causation and probability increase.

There are also those who would eliminate the entire debate over the causal relata because they deny that causation is a relation – e.g., Achinstein (1979), Mellor (1995: Ch. 13). For present purposes there is no need to pursue these arguments.

### *E.* Second-order Implications

I turn now to some familiar properties of relations that, when they obtain, license characteristic inferences. Consider *reflexivity* first. If causation is totally reflexive, every item causes itself and we could write down $xCx$ for any item $x$. It should come as no surprise that is not a serious candidate. Most philosophers consider causation irreflexive:

> not-($xCx$), for any $x$. (C8)

Some philosophers, such as Spinoza, reject C8 because they maintain that there is one entity – God – for which C8 does not hold, but Spinoza clearly associates a different concept with $xCy$ than do most philosophers and other folk. This is not objectionable in itself, but it provides an occasion for reflection on issues that arises in deciding what terminology to adopt. Spinoza would have us use causal language in a way that eliminates a property

commonly attributed to this relation in order to include a special case in its scope. This special case is an entity whose existence (and properties if it exists) is more controversial than most of the items to which causal thinking is commonly applied. This is doubly so because Spinoza's God-concept differs substantially from that of most theists. These considerations suggest that following Spinoza's route will do more to confuse our thinking about causation than to clarify it. Moreover, Spinoza takes his description of God as "self-caused" to be equivalent to saying that God's essence includes existence; but we can understand and discuss this claim without introducing causal terminology. As a result, I will consider only causal concepts that include C8 among their implications.

Many philosophers consider causation to be *transitive* and thus accept the validity of the argument:

$$xCy, yCz \therefore xCz. \tag{C9}$$

Here is a recent description of this view: "That causation is, necessarily, a transitive relation on events seems to many a bedrock datum, one of the few indisputable a priori insights we have into the workings of the concept" (Hall 2000: 198). But even here disagreements occur, with some philosophers proposing examples that intuitively seem to fall under the causation rubric, but are non-transitive (e.g., Davis 1988; Lee 1988). Hall examines several arguments against transitivity, and concludes that transitivity is incompatible with counterfactual dependence.[17] This leads Hall to conclude that we have two causal-relation concepts. One of these is defined by counter-factual dependence and is not transitive: the other – which is more central to our causal thinking – is transitive but violates counterfactual dependence (2000: 219).

Probabilistic accounts of causation raise problems about transitivity since it is fairly easy to construct cases in which $Pr(y|x) > Pr(y)$ and $Pr(z|y) > Pr(z)$, but $Pr(z|x) < Pr(z)$; Suppes (1970: 58–59) gives one example. But Suppes notes that deterministic causation, for which $Pr(y|x) = 1$, is transitive on his account. Eells (1991: Ch. 4) discusses several cases in which transitivity fails for probabilistic causation, and establishes a condition that is sufficient, but not necessary, for transitivity to hold.

Now consider *symmetry*. No one I know of considers causation to be symmetric – that is, no one holds that $xCy$ implies $yCx$, but there is substantial debate about whether causation is asymmetric or non-symmetric. In other words, there is debate over the validity of:

$$xCy \therefore \text{not-}(yCx).[18] \tag{C10}$$

If causation is non-symmetric and $xCy$ true, then we must determine on a case-by-case basis whether $yCx$ is true.

Views on the symmetry properties of causation can depend on views of the causal relata. For example, if one maintains that causal relations hold among event-tokens, then the asymmetry of causation immediately follows since event tokens occur only once. If causal relations hold between event-types, then there is a straightforward sense in which the same relata may occur at different times, and this may yield cases in which we have both $xCy$ and $yCx$. Davis gives this example: "Suppose Jack and Jill regularly give each other colds. Then Jack's getting a cold causes Jill to get one, and Jill's getting a cold causes Jack to get one" (1988: 146). In this example the cause precedes the effect in time, but consider two ladders supporting each other. This would seem to be a case of simultaneous symmetric causation, although the exact analysis will depend, again, on one's view of the relata. One could, for example, argue that the relata are event tokens and that we have an asymmetric series of very rapid successive events consisting of ladder *A*'s standing causing ladder *B* to stand, followed by *B*'s standing causing *A* to stand, and so on. Alternatively, if the relata are physical objects, then we have a symmetric relation in which each standing physical object causes the other to stand.

As the last example suggests, we must not forget the distinction between temporal and causal priority. Some accounts of the causal relata (e.g., event types) can be combined with the requirement that a cause precede its effect in time to yield an asymmetric causal relation, but this will not work if causal relations are independent of temporal relations. In this case, the detailed account will depend on other views. For example (without going into the details), Mackie (1980: Ch. 7) holds that causation is non-symmetric while Tooley holds that it is asymmetric (1987:178–80). Torretti maintains that symmetry considerations yield another difference between determinism and causation:

> the binary relation '*x* determines *y*', where *x* and *y* are different states in the evolution of a physical system subject to differential equations, is a *symmetric* relation; whereas '*x* causes *y*' is *antisymmetric*: If *x* causes *y*, it is certainly false that *y* causes *x*.
>
> (1999: 133)

Other philosophers, we have seen, do not share this certainty.

### *F.* Causal-Relation Systems

We turn now to the membership of the conceptual system that includes CAUSATION. Some accounts require that this system include temporal concepts (and perhaps spatial concepts); others require that temporal concepts be excluded. Some accounts require EVENT TOKEN or EVENT TYPE. Other versions require LAW OF NATURE, and further details will depend on one's view of laws. Probabilistic accounts of causation require concepts from

probability theory – a theory that has not been mastered by most people. Salmon's view requires PROCESS and accounts of PHYSICAL OBJECT and EVENT in terms of PROCESS. More recently Salmon (1994) has changed his account in response to criticisms and mostly accepted a view proposed by Dowe (1992: 210–15) which holds that causation is to be analyzed in terms of the transmission of *conserved* quantities; Salmon disagrees mainly in preferring *invariant* quantities (Salmon 1994: 305, see Dowe 1995 for a reply). These proposals require the introduction of additional concepts from physics into the conceptual system in which CAUSATION is located. Heathcote (1989) carries this tendency further, requiring the machinery of quantum field theory for an account of the physical basis of causation. These moves are in accord with Salmon's insistence that he is seeking an account of causation that describes *this* world, not any possible world (1984: 239–42). This project requires that we rework our causal concept as we learn more about this world, and is completely in accord with the spirit of a Sellarsian view of descriptive concepts.

### *7.2.2 Extra-systemic Relations*

We turn next to the kinds of evidence that is considered relevant for assessing the existence of causal relations. As Hume noted, this issue arises on two different logical levels: We may ask why we believe that there are any causal relations at all, or why believe that a specific causal relation holds (2001: 55). Kant recognized the same distinction and attempted to provide an a priori basis for the general belief in causation while insisting that the justification of specific causal claims is an empirical matter. From the naturalistic perspective adopted here both questions are ultimately empirical in spite of their different levels of generality. The general thesis that every event has a cause has long served as a GA in everyday life and scientific research, although it has often been questioned in the realm of human action and has been challenged more recently in microphysics. As in the case of other GAs, it is accepted as a general guide for research because of its success in specific cases; its continued acceptance largely depends on its continuing success in guiding research. For this reason I will focus on the kinds of evidence relevant for evaluating specific causal claims. Many philosophers who propose analyses of causation offer accounts of this evidence, either explicitly or implicitly. I will consider some key cases, beginning with accounts of deterministic causation.

Many agree that an observed constant conjunction of two types of items provides one important kind of evidence for causal claims, but any moderately sophisticated account of causation recognizes that this is not definitive evidence. Constant conjunctions may arise through coincidence, as joint effects of a single cause, or from spurious causes in Suppes' sense. Mill's methods capture some of the considerations that support or undermine causal claims. These are useful methods as long as we use them intelligently

and understand that they do not give indubitable results – which are not to be expected in an empirical inquiry in any case. They are especially useful when we can control the occurrence of the putative cause and observe whether the presumed effect occurs, or whether the degree of the effect varies with changes in the degree of the cause.

While Hume insisted that a constant conjunction is required for a causal attribution, advocates of a singularist view reject this requirement. Hume also required that the constant conjunction be observed, but he wrote before science taught us that much of nature is not available to our unaided senses. This led to the wholesale postulation of causal relations in which we cannot perceive one or both of the relata. The role of viruses and bacteria as causes of disease symptoms, and of radioactivity in causing a variety of effects, illustrate cases in which we introduce an unperceived cause for perceived effects. Cases in which we administer antibiotics to kill bacteria, or change the underlying structure of an object by heating or crushing it, are cases in which we recognize an unobserved effect of an observed cause. Detailed accounts of what occurs in these cases – such as how an antibiotic molecule interacts with a particular bacterium – illustrate cases in which neither cause nor effect are available to our senses. In these cases we ultimately rely on evidence that passes through our senses, but this is often evidence for a theory or cluster of theories from which we derive a causal relation. These scientific developments have generated an ongoing re-examination of the nature of empirical evidence; examples include Bogen and Woodward 1988; Brown 1985, 1987, 1995; Kosso 1989, and Shapere 1982. Changing views of the nature of empirical evidence may alter our understanding of the evidence for causal relations; according to TC, such changes involve changes in our causal-relation concepts. A similar impact derives from varying views of theoretical entities and of the concepts we use to describe them. As noted above, Tooley (1987) considers causation to be a theoretical relation, and this claim is a substantive part of his account. Tooley sets this claim in the context of a detailed defense of a realist view of theoretical entities, while also seeking to retain as much traditional empiricism as he can. As a result, those who reject Tooley's views on theoretical entities and empirical evidence will advocate different causal-relation concepts than Tooley does.

In mentioning Tooley I have moved into the realm of probabilistic accounts of causation. These views require only statistical evidence for a causal relation. Suppes describes his approach as a generalization of Hume's constant-conjunction requirement:

> The notion of frequent co-occurrence is at the very heart of the idea of causality. . . . It is not a matter of presenting evidence for causality by offering probabilistic considerations but it is part of the concept itself to claim relative frequency of co-occurrence of cause and effect.
>
> (1970: 45)

Suppes holds that the required evidence consists of positive statistical relevance plus temporal priority of the cause; other advocates of a probabilistic view reject one or both of these requirements. Salmon considers the use of statistical evidence for causal claims as a "striking failure to fit the Humean picture of constant conjunction" (1984: 184), rather than as a generalization of that picture.

Allowing probabilistic evidence does not entail a probabilistic account of causation; such evidence may be considered relevant to a deterministic account. Whether probabilistic evidence can support a deterministic claim will depend on how one deals with other issues. For example, recognition that we do not know all the relevant confounding factors may lead one to hold both that the same cause always yields the same effect *ceteris paribus*, and that probabilistic correlations provide relevant evidence for deterministic claims.[19] On a Bayesian approach, which is most clearly applicable to probabilistic hypotheses, we use evidence to assess the relative probability of competing hypotheses, but there is no direct mapping from probabilities found in the evidence to the probabilities of various hypotheses. A sample containing equal numbers of *x*s and *y*s may provide grounds for preferring the hypothesis that, overall, 55 percent of *x*s are *y*s to the hypothesis that 60 percent of *x*s are *y*s. We may even use a sample in which all *x*s are *y*s as evidence for a probabilistic claim.

The examples given in this section are far from exhaustive, but they will suffice for the point of current interest. Varying accounts of causation may include varying accounts of the appropriate kind of evidence for causal claims, and as these are mixed and matched with other features we have considered, the range of alternative causal-relation concepts expands.

### *7.2.3 Systemic Role*

Now consider the aim or aims that CAUSATION embodies in our repertoire. I will discuss four roles that have been proposed: control, explanation, understanding, and prediction.

According to Gasking control is the essence of causation: "the notion of causation is essentially connected with our manipulative techniques for producing results" (1955: 483). When we properly claim that events of type *x* cause events of type *y*, "it is always the case that people can produce events of the first sort as a means to producing events of the second sort" (1955: 483). Thus we explain "the 'cause-effect' relation in terms of the 'producing-by-means-of' relation" (1955: 485). Gasking acknowledges that "cause" is used in other senses, but "The notion of 'cause' elucidated here is the fundamental or primitive one" (1955: 486). For example, when we say that gravity causes unsupported bodies to fall our use of "cause" is "a sophisticated extension from its more primitive and fundamental meaning" (1955: 487).

Von Wright defends a similar view:

> I now propose the following way of distinguishing between cause and effect by means of the notion of action: $p$ is a cause relative to $q$, and $q$ an effect relative to $p$, if and only if by doing $p$ we could bring about $q$ or by suppressing $p$ we could remove $q$ or prevent it from happening.
>
> (1971: 70)

The notion of what *we* can do is central to this account: there is an "essential connection between causation and (human) action" (1993: 119). Even in a case such as the destruction of Pompeii by the eruption of Vesuvius, which we could not bring about, we break complex events into a series of simpler events that we can bring about. As a result we still think of the cause as something that we could do: "*that p* is the cause of $q$, I have endeavored to say here, *means* that I could bring about $q$, if I could do (so that) $p$". This shows, von Wright adds, that "to think of a relation between events as causal is to think of it under the relation of (possible) action" (1971: 74). More recently Pearl has taken up this theme, although he allows for other roles as well: "the very essence of causality [is] the ability to predict the consequence of abnormal eventualities and new manipulations" (2001: 345). Pearl adds that even in cases such as celestial motions, where we have no prospect of control, "the theory of gravitation gives us a feeling of understanding and control, because it provides a blueprint for hypothetical control."[20]

Note one point of disagreement between von Wright and Gasking: Gasking holds that it is logically impossible for an effect to precede its cause in time. This follows from his analysis of causation in terms of control plus the claim that "It is a logical truth that one cannot alter the past" (1955: 483). We have already seen that von Wright allows for effects that precede their causes. Mackie offers another option: He holds that backwards causation is logically possible, and that we cannot alter the past, but does not identify causation with control. Thus: "If there were such a thing as backward causation, it would somehow have to stop short of offering us a means of bringing about the past" (1980: 168).

I now want to consider a general point before looking at other roles that have been attributed to the causal relation. Often philosophers undertaking an analysis seek a single feature that captures the essence of a concept – as do Gasking and von Wright. This quest is at odds with the TC account of descriptive concepts in which systemic role is just one facet of a full account. From this perspective the accounts given by Gasking and von Wright of the "essence" of causation are incomplete. Moreover, a concept may play multiple roles in its conceptual system. When this occurs debates about which is *the* basic or fundamental role will often be misconceived. Mackie recognizes this point, attributing both an explanatory (1980: 164) and a control (1980: 168–71) function to CAUSATION. He also maintains that the relation has different relata associated with the two roles: "We need, then, to

recognize both kinds of cause, producing causes and explanatory causes, events and facts, and at the same time to distinguish them, in order to understand what we think and say about causal relations" (1980: 265). Blanshard (1962: 445) includes understanding and control as functions of causation, and identifies understanding with explanation.

The view that causation has an explanatory role is common among those who adopt a deterministic account. On this view the concept of a causal relation implies the concept of a sufficient condition, so that *xCy* conjoined with *x* explains why *y* occurred. Those who adopt a probabilistic causal concept may also hold that causation is an explanatory concept (e.g., Salmon 1984: 19–20, 113, 120–21, 132–33), although this will typically require a different account of explanation. The role of causal knowledge as a means of prediction is too familiar to require documentation, but much discussion of the relation between explanation and prediction derived from Hempel and Oppenheim's thesis that explanation and prediction have the same logical form – which led many to identify the two roles. This view has been vigorously opposed by those who note that we can often predict in cases where we cannot explain (e.g., Toulmin 1961) and that we can sometimes explain where we cannot predict (e.g., Scriven 1962). Cases in which we can predict but not control are clear enough: any constant conjunction can provide the basis for prediction without control. We can also control without explaining since we may learn to regularly bring about some outcome without an explanation of the connection. This is common in contemporary medicine where the phrase "mechanism of action not understood" occurs frequently in descriptions of medications in *The Physicians' Desk Reference*. Cases in which we can explain a phenomenon without being able to control it are also myriad. It does seem, however, that the ability to control an outcome brings along the ability to predict its occurrence, whether the prediction and control are deterministic or probabilistic.

The point of these remarks is that explanation, prediction, understanding, and control are all functions commonly attributed to causation, although conceptual analysts disagree on which of these are proper or basic roles of the causal relation. Those who advocate different roles for causation provide another example of analysts who may well be discussing different causal-relation concepts. How one develops these concepts may depend on views about related issues. Mackie, for example, agrees with von Wright that the control function of causation is the source of our awareness of causal asymmetries, but rejects the attempt to analyze causal asymmetry in terms of control (1980: 172–73). In addition, one's views of the nature of causal explanation, and of the relation between explanation and other roles, will depend on the associated view of explanation. Salmon, for example, maintains that all explanation is causal explanation – which provides some of the motivation for trying to interpret probabilistic physics as causal – and also identifies explanation with understanding (1984: 9, 132, 259–63). But the relation between explanation and understanding has been a subject of extensive

debate in philosophy of science; I will cite just one contrasting view. Cushing (1991) agrees with Salmon that understanding requires causation, and that irreducibly probabilistic physics explains, but rejects probabilistic accounts of causation. Thus Cushing holds that quantum mechanics, interpreted probabilistically, explains but does not yield understanding.

## 7.3 Is Causation a Kind of Necessary Connection?

In Sec. B I noted the distinction between holding that causation involves a necessary condition and that it is a necessary connection. I turn now to varying views on the latter thesis. Following Hume many hold that there is no necessary connection between a cause and its effects, but Hume's arguments show, at most, that the relation between cause and effect is not one of logical necessity.[21] In this part of his argument Hume assumes that logical necessity is the only intelligible kind of necessity, and many philosophers agree. Of course, other philosophers disagree. Among those who hold that causation involves a necessary connection there are at least three different views of the nature of this necessity: some hold that causal necessity is logical necessity (entailment); some hold that causal necessity is different from, but analogous to, entailment; and some hold that causal necessity is *sui generis*. I will explore examples of each view.

I want to approach the first view by comparing some of the features most commonly attributed to the causal relation with those of entailment, as commonly understood.[22] Consider some characteristic implications involving the entailment relation. First, entailment involves the notion of a sufficient condition:

$$pEq, p \therefore q, \tag{E1}$$

which parallels C1. However, if an entailment relation holds between two propositions, it continues to hold no matter what other considerations apply:

$$pEq \therefore p\&rEq, \tag{E2}$$

is valid for any $r$. No *ceteris paribus* clause is required for entailment.[23] The parallel argument C2 is, we have seen, typically held to be invalid. An analogous situation holds for the notion of a necessary condition. In the case of entailment, the following two arguments are valid:

$$pEq, -q \therefore -p \tag{E3}$$

$$pEq \therefore -qE-p. \tag{E4}$$

We have seen that while C3 is valid, C4 is typically rejected. In addition, entailment is totally reflexive while causation is irreflexive on any reasonable

account; and entailment is transitive while the transitivity of causation is a subject of dispute.

These examples mostly involve cases in which the consequences of *xCy* are a subset of the parallel consequences of *pEq*, so one might try to model causation as a kind of restricted entailment. But the irreflexivity of causation is an exception. Moreover, many of the causal concepts we have examined license implications that have no parallel for entailment. For example, entailment is clearly non-symmetric: Given *pEq*, nothing follows about *qEp*; those who hold that causality is asymmetric maintain that *xCy* entails not-(*yCx*). In addition, entailment is atemporal: *pEq* has no temporal consequences, while on some accounts *xCy* does have temporal consequences. It is also widely held the relata of the entailment relation are propositions, while typical accounts of causation reject this view of the causal relata.

There is, then, not much motivation for thinking of causation and entailment as substantially the same relation, and attempts to assimilate the causal relation to entailment are misguided. I submit that consideration of instantiation conditions and systemic roles support this conclusion.

Blanshard will take us towards the second approach – that causation is a kind of necessity analogous to, but not identical with, entailment. Sometimes Blanshard asserts that causal necessity is identical with entailment (e.g., 1939 vol. I: 513), but he also presents a weaker view:

> we are not suggesting, of course, that causality *reduces* to logical necessity. What we hold is that when one passes in reasoning from ground to consequent the fact that the ground entails the consequent is one of the conditions determining the appearance of this consequent rather than something else in the thinker's mind.
>
> (1939 vol. II: 496)

And a bit later, "It seems probable that the causal relation is everywhere complex, and that the relation of necessity is but one of its strands" (1939 vol II: 502, cf. 503). Elsewhere, after considering the role of entailment in thought Blanshard concludes: "it is legitimate to surmise something like it in the sequence of physical events" (1962: 454). Note also that Blanshard's conception of entailment is rather different from that of most contemporary logicians since he maintains, for example, that necessity come in degrees (1939 vol. II: 499). Thus, from the perspective of the most common contemporary view of entailment, Blanshard holds, at most, that causal necessity is analogous to entailment.

Sellars presents a clear account of what this second approach involves. Asking whether causation can be considered a kind of "physical entailment" he notes that to justify such an approach, "one must make plausible the idea that these entailments play a role in causal reasoning analogous to the role of 'formal' entailments in less problematic forms of inference" (CDCM

270). I submit that our discussion indicates that the analogy is limited, and that it is not illuminating to think of causation as a species of entailment.

This leaves the third possibility – that the causal relation is a kind of necessity, but a different kind of necessity from that found in logic. We find this view in the Aristotelian tradition; it has also been defended by Burks (1951), Ducasse (1951), and Taylor (1963, 1966: 27–28). This approach easily handles Hume's point that we can consistently conceive of a putative cause without the usual effect, which holds for logical necessity but may not be relevant to necessary connections of a different species. I am going to evade any further consideration of this issue on the grounds that a proper discussion will require a study of necessary-relation concepts of the sort that I am currently engaged in for causal-relation concepts. I have no intention of adding this study here, but I submit that our discussion in this chapter indicates what would be required to pursue this question.

## 7.4 Conclusion

The discussion in this chapter leaves us with a large number of concepts that various philosophers hold to be "the" causal concept. If there is a single, unambiguous causal-relation concept already embedded in ordinary, non-philosophical, non-scientific thinking, then one of these accounts may accurately capture that concept. If, on the other hand, there are various more or less similar causal-relation concepts in ordinary discourse, different philosophers may be providing accounts of different members of this set. Moreover, there is no reason why philosophers should limit themselves to seeking accurate accounts of pre-existing concepts. Introduction of variant concepts is part of the normal process of seeking to understand a domain. Considerations of communication and coherent thought require that alternative concepts have a basis in existing concepts but, we have seen, continuity is compatible with genuine novelty. Once a new concept has been formulated, we can then ask such questions as whether it is embedded in ordinary language or a particular theory, whether that concept is instantiated in the world, or whether it is the appropriate concept for carrying out some scientific or philosophical research program.[24]

Analytic philosophers typically reject this view of the properly philosophical concerns about concepts. It has long been a standard view in analytic philosophy that there is a common sense framework, and that philosophers should seek to specify the content of concepts in that framework with precision. Most of the philosophers discussed in this chapter claim to be engaged in that project. Here is a relatively early statement of the view:

> The problem of giving a "correct" definition of the causal relation is that of making analytically explicit the meaning which the term "cause" has in actual concrete phrases that our language intuition acknowledges as proper and typical cases of its use.
>
> (Ducasse 1926: 57)

Some expressions of the project seem to be more ambitious. Tooley, for example, is concerned with causation as it occurs in the actual world, and beyond:

> For the goal is to set out an analysis of the *concept* of causation, and not merely to offer an account that is true of causation as it is in the actual world. The theory must be true of causation as it is in all possible worlds. So none of the statements in the theory can be merely contingently true.
>
> (1990: 292)

However, Tooley is working from a traditional empiricist perspective which holds that all non-contingent truths are analytic, so his view is not that different from that of Ducasse. Yet this view raises crucial questions that are rarely discussed in the current philosophical literature. Who has this concept of causation? Is it an innate concept possessed by all humans everywhere and at all times? Is it a concept that is learned as children become socialized but that is, nevertheless, learned in all societies at all time periods? Is it a concept that is peculiar only to certain cultures or groups? If the latter, we may ask just how important that concept is. The concept of a witch is widespread across cultures and eras, but most contemporary educated people discount this concept on the grounds that is it not instantiated, and philosophical analyses of the concept are hard to find. What guarantee do we have that "our" concept of causation has instances? Moreover, if the concept is not universal, but peculiar to a particular culture, shouldn't we consider the possibility that another culture may have a concept that plays a role in their thinking similar to the role our causal concept plays in ours, but that their concept has advantages? Sometimes culture *A* will adopt tools, such as a steel ax, from culture *B* because members of *A* conclude that *B*'s axes are superior for *A*'s own goals. Short of an argument for a universal conceptual framework, we should acknowledge the possibility that concepts are like tools in this respect. In addition, work within our own culture can improve our tools and sometimes lead to their replacement by different tools that we judge superior. Given our cognitive history the conclusion that our conceptual repertoire develops and changes in this manner is unavoidable. The discussion in the present chapter underlines the significance of the questions I have just raised.

Science is an especially important source of new concepts, and science is concerned with discovering concepts that are instantiated in this world. It should come as no surprise, then, to find philosophers who keep a close eye on science taking a different view of the aim of an account of causation than the one expressed by Ducasse and Tooley. Salmon, for example, is quite clear that he is concerned to describe a concept that is appropriate to the actual world, not to any possible world. Thus Salmon rejects teleological causes because "A world in which teleological causation operates is not

logically impossible, but our world does not seem, as a matter of fact, to be of such kind" (1984: 164). As we have seen, Salmon's discussion of the causal relata proceeds in terms of what is required by current scientific theories, and he includes concepts from recent science in his account. Salmon also notes that he does not know if we can give causal explanations of quantum phenomena, but suggests that this will be possible "only if the concept of causality is fundamentally revised" (1984: 254). We found a similar approach in Dowe, and my remarks (at the end of the discussion of temporal implications) about a priori constraints on causation apply to all features of all versions of that concept. Adoption of a concept – even universal adoption of a concept – does not constrain nature. At most, a concept may constrain the development of science by serving as the basis for a GA. This will give the concept a certain tenacity in our thinking, but GAs are open to reconsideration, and adopting a GA does not prevent researchers from recognizing evidence that could count as a challenge.

I want to press this point in another direction. Consider again Salmon's remark that a causal explanation of quantum phenomena will require a fundamentally revised concept of causation. It is a common feature of human conceptual development that when we replace a concept with one that is fairly similar, we tend to associate the same word with this new concept. A series of such changes can leave us with little similarity between the concept currently associated with a word and concepts that were associated with it at an earlier stage. Proposals to revise our concept of causation are, I submit, examples of this process. But we may also encounter domains that we cannot successfully understand as causal in any sense in which we have previously thought about causation, and in which no cognitive gain will be achieved by retaining causal language. In such cases we should seriously consider dropping causal language, rather than retaining the language while we shift the associated concept.

From this perspective it is worth taking another look at the claim we found in Salmon and Suppes that the common understanding of causation is probabilistic rather than deterministic, and that the identification of causation with determinism is a result of the success of Newtonian physics. I suggest that the success of Newtonian physics provided an excellent reason for adopting a deterministic concept of causation in physical theory, and that quantum mechanics provides a reason for reconsidering the scope of that concept. Neither of these options is eliminated by the use of causal language in non-scientific discourse.[25] There are fields, such as anthropology and linguistics, in which it is important to understand how members of a group usually speak, but attempts to understand domains of the physical or biological world are not among these fields. Even in the social sciences theoretical concepts introduced in the attempt to understand group behavior need not be in the repertoire of those groups. In the case of causation, some of the proposals we have seen, such as eliminating any temporal constraint on causes, may strike many as "counter-intuitive," but the history of science

is full of counter-intuitive conceptual innovations. In this regard the case is no different from the introduction of curved space or fractal dimensions, from recognizing that there are infinite sets of different sizes, or introducing a time concept that makes simultaneity at a distance conceptually impossible.

I want to end this discussion by returning to the beginning of Mackie's book where he distinguishes three different kinds of analysis of causation: epistemic, conceptual, and factual:

> It is one thing to ask what causation is 'in the objects', as a feature of a world that is wholly objective and independent of our thoughts, another to ask what concept (or concepts) of causation we have, and yet another to ask what causation is in the objects so far as we know it and how we know what we do about it.
>
> (1980: ix)

As is common in the analytic tradition, Mackie focuses on the second question. I have been arguing that all three questions are within the scope of philosophical reflection – although the first and third questions require attention to subjects other than philosophy.

Questions of the same type as Mackie's arise for other concepts besides causation. The importance of the first and third questions is particularly clear when we consider the development of a new field of study, such as radioactivity, where researchers do not begin with an appropriate set of concepts, but must work these out as they discover new features of the world. These questions are also central when new developments lead us to rethink familiar areas of discourse. I submit that properly informed philosophers can contribute to the ongoing process of understanding our world – and even, at least in social domains, improving it – by exploring alternative conceptual structures, rather than limiting ourselves to seeking clear formulations of what we already think.

# 8 Conceptual Analysis II: Epistemic Concepts

> The essential point is that in characterizing an episode or a state as that of *knowing*, we are not giving an empirical description of that episode or state; we are placing it in the logical space of reasons, of justifying and being able to justify what one says.
>
> (EPM 169)

A system of epistemic concepts provides the basis for thinking about, and acting with respect to, cognitive matters. An epistemic system will typically include such notions as knowledge, belief, evidence, confirmation, truth, and others – although here too the exact membership of the system is subject to dispute. Analytic epistemologists offer analyses of these concepts and I will consider examples as we proceed. But I will also argue that the contents of this system change with changes in our overall epistemic situation and our understanding of that situation. I will consider some changes that have been discussed in the literature, and will make some recommendations about what we should include in this system given our present perspective. This will amount to a fragment of a descriptive epistemic theory. However, the central concepts in this system have both descriptive and prescriptive aspects.[1] Although I will begin with the descriptive side, the prescriptive side will soon make its way into our discussion. This chapter will, then, provide the most detailed account of a prescriptive theory in this book.

## 8.1 The Analytic-Synthetic Distinction II

I want to return to the analytic-synthetic distinction since this will bring into focus some key issues to be addressed by a system of epistemic concepts. Recall that for Kant the analytic-synthetic distinction is not an *epistemic* distinction. It is a logical distinction: analytic propositions have inconsistent negations while synthetic propositions have consistent negations. Kant considered it crucial to separate this distinction from the epistemic distinction between a priori and a posteriori knowledge. Yet Kant's way of setting up these distinctions is just one phase in the debate over how to think about

epistemic matters. Twentieth century empiricists typically held that there are only two types of propositions: analytic a priori and synthetic a posteriori. As a result, the claims that a proposition is analytic and that it is a priori are (at least) materially equivalent, and it was standard practice to infer one from the other; a parallel point holds for synthetic and a posteriori. Thus in an empiricist framework only two concepts are required – as in Hume: "All the objects of human reason or enquiry may naturally be divided into two kinds, to wit, *Relations of Ideas*, and *Matters of Fact*," (1975: 25). Each type of proposition is jointly specified by an epistemological and a logical feature: Relations of ideas are knowable by thought alone *and* their negations are inconsistent. Matters of fact have consistent negations and are (therefore) not knowable by thought alone. Kant drives a wedge between these epistemic and logical characterizations in order to make room for a third type of proposition. Indeed, Kant's distinctions were developed partly to eliminate problems he found in Hume and Leibniz; it will be useful to look briefly at Leibniz.

Leibniz also distinguishes two basic kinds of propositions, *truths of reason* and *truths of fact*. This is an epistemological distinction that is relevant only to creatures of limited cognitive ability, such as ourselves. For Leibniz, all true propositions have inconsistent negations and are knowable by thought alone for a sufficiently powerful intellect. Truths of reason are propositions that *we* can establish by reason alone; truths of fact are propositions that *we* can discover only on the basis of experience. From a logical perspective there is only one kind of proposition. There is no room for synthetic a priori propositions in either the Humean or Leibnizian framework.

Kant maintains that synthetic a priori propositions must exist if we are to avoid both Humean skepticism and Leibnizian dogmatism. Thus Kant alters his predecessors' frameworks. His separation of the analytic-synthetic distinction from the a priori/a posteriori distinction provides the required conceptual space. This results in four possible combinations, although Kant immediately drops one of these – analytic a posteriori propositions – from consideration: "For it would be absurd to found an analytic judgment on experience. Since, in framing the judgment, I must not go outside my concept, there is no need to appeal to the testimony of experience in its support" (B11, 1963: 49; Kant is using "judgment" as a synonym for "proposition"). But although the predicate of a proposition does not go beyond the subject, this may not be obvious; disputes among analytic philosophers about whether specific propositions are analytic illustrate the difficulties we encounter. In such cases we may have empirical grounds for believing that a proposition is analytic (such as the behavior of members of a community when we question its truth) and thus seek an appropriate analysis. As another example, consider an alien anthropologist who arrives on earth with limited knowledge of English. After interviewing a few English speakers our anthropologist might conclude that all aunts are female on

empirical grounds. Later, after learning more English, the anthropologist realizes that the empirical procedure was unnecessary; "All aunts are female" is analytic and its truth *can* be discovered by reflection alone. Still, we can describe the anthropologist as initially learning an analytic truth empirically.

We can view Kant's new conceptual structure as a result of adapting and synthesizing elements from Hume and Leibniz. From a Humean perspective we have the addition of a third class of propositions that is generated by distinguishing the logical from the epistemological aspects of his account. From a Leibnizian perspective the new class can be seen as the result of adopting the distinction between truths of reason and truths of fact while insisting that, considered logically, there really are two distinct types of propositions.

Now let us recall Quine's critique of the analytic-synthetic distinction. Quine worked in the empiricist tradition that does not consider the notion of a synthetic a priori proposition to be a live option; if any propositions can be known a priori, they are analytic. But Quine's theses that all our beliefs form a web, and that any belief in the web can be modified in response to experience, implies that no propositions can be known a priori. Thus there is no reason to have either ANLYTIC or A PRIORI in our active epistemic system. From the perspective of TC, Quine is proposing to change our epistemic system by eliminating two concepts that do no work. We can retain those concepts for purposes of historical studies – much as historians can reconstruct PHLOGISTON – but those engaged in active epistemological research have no more need for ANALYTIC and A PRIORI than a contemporary chemists has for PHLOGISTON.

TC also provides a new perspective on a much-criticized feature of Quine's argument: his move from the failure of various attempts to define "analytic" to the conclusion that there is no such concept. Naturally, many philosophers rejected the move from a failure to provide a definition to the conclusion that there is no coherent concept. The history of philosophical analysis largely consists of failed attempts to define concepts that we seem to possess. Yet there is a point to Quine's strategy: Analyticity is a descriptive concept, and an account of this concept must include criteria for determining if there are such propositions. Moreover, ANALYTIC is supposed to describe many propositions that regularly occur in everyday thought. We should, then, be able to specify conditions that will allow us to identify analytic propositions, if there are any. The lack of such criteria is a genuine problem for those who would use this concept, and TC suggests that the concept is not well developed. In the context of analytic philosophy this is a major problem because of the central role that analyticity plays in the enterprise. Quine's challenge could provide the impetus for seeking a better-developed concept – which is not the same as seeking a more accurate analysis – but I will not pursue this option. Instead, I want to return to Putnam's alternative critique of the analytic-synthetic distinction and the introduction of GUIDING ASSUMPTION into our epistemic framework (Sec. 3.7).

Recall that on Putnam's account we retain the analytic-synthetic distinction but it does not play the fundamental theoretical role that it does for many philosophers.

Introduction of GAs into an empiricist framework has clear analogies with Kant's move when viewed as a response to Hume. Still, GAs do not fit neatly among Kant's distinctions, and we can now see that the major differences between GAs and synthetic a priori propositions are located in the associated implications. GAs meet Kant's criteria for synthetic propositions even though they are protected from empirical refutation. But they are not a priori propositions because they are not *permanently* immune from rejection in response to empirical outcomes. Moreover, propositions become GAs as a result of human decisions; they are not imposed on us by transcendental philosophy. In other words, GA does not imply A PRIORI, PERMANENT, or TRASCENDENTAL. Still, the instantiation criteria and systemic roles of GAs are very close to those for Kant's synthetic a priori propositions. Consider ICs first. Propositions of both types have a key characteristic that allows us to identify them: we recognize the kind of evidence that, from a purely logical point of view,[2] could stand as an empirical refutation; but when such evidence occurs we direct the force of *modus tollens* elsewhere (as Lakatos (1970) would put it) and protect these propositions. Kant's main examples of synthetic a priori propositions can now be seen as GAs of an earlier stage of science. Quine was only a small step from recognizing the importance of GAs when he noted that we treat some propositions in this fashion, but he failed to develop the positive side that this practice plays in guiding research.

In the case of systemic role, GAs have both a descriptive and a normative side: GAs such as Newton's laws and conservation of energy state what we take to be basic features of the physical world, and also direct us how to proceed in dealing with empirical outcomes. Synthetic a priori propositions share these roles.[3] While we can view the introduction of GAs as a modification of either the Kantian framework or the empiricist framework of the 1950s, historical considerations suggest that the second perspective is more illuminating. Philosophers such as Kuhn, Putnam, Sellars, and Toulmin, who gave a central role to GAs, were working in the context of logical empiricism and were aware of its problems; at least some of them were familiar with Kant. We can, then, view them as proposing a modification of the empiricist framework they inherited through the introduction of Kantian themes.

Now consider the various roles that analytic propositions play in twentieth century empiricism. One role is as transformation principles that allow us to reduce claims in the auxiliary language to claims in the observation language; this kind of reduction is the second dogma that Quine attacks. Analytic propositions also provide definitions that connect terms in the auxiliary language, such as when we define "aunt" as "sister of a parent." But analytic propositions were forced to play another role for which, I will argue, they are not suited: The norms that provide the epistemic basis for

science must be either analytic or grounded on analytic propositions. This follows directly from two claims: knowledge of analytic propositions is the only a priori knowledge we can have; and these norms must be known a priori. I want to examine this view.

Twentieth century empiricists typically held that norms and propositions are different in kind, and that "analytic" applies only to propositions. In ethics, empiricists typically adopted either an emotivist view, which rejects any cognitive account of ethical norms, or confined themselves to meta-ethics – a descriptive project – and left questions of ethical norms aside. In philosophy of science, logical empiricists did not adopt either of these approaches. They sought norms – especially norms that would guide the epistemic evaluation of universal generalizations on the basis of a finite body of evidence. A standard argument leads to the conclusion that knowledge of such norms must be a priori (e.g., Siegel 1989; Stroud 1985):[4] These norms must be known either a priori or empirically. Now we are seeking criteria for the epistemic evaluation of empirical claims, and these criteria must be justified; call these criteria CE. If CE consists of empirical claims, then their justification will have to be based on CE – we would have to use the criteria we are seeking to justify for their own justification. Yet if criteria of justification can be used to justify themselves, we could easily find a variety of mutually incompatible criteria that are all equally justifiable, each on their own grounds, and we would lack any sound basis for carrying out empirical research. But if these criteria cannot be known empirically, the only alternative (short of skepticism) is that they must be known a priori.[5]

An additional consideration supports the view that such norms can be known a priori: Deductive logic exists and provides a paradigm of a priori normative knowledge. When we acquire evidence against a universal generalization the relevant norm is provided by *modus tollens*, a valid deductive argument. This suggests that a set of a priori norms for evaluating supporting evidence could be found if they are logical norms, somehow modeled on deductive logic (e.g., in an early twentieth century version, if they are purely syntactic). I will not pursue the vagaries of this project – confirmation theory – here. Instead I want to press the question of the basis for this a priori knowledge of norms – whether logical norms or some other sort. The answer currently on offer is that this knowledge must derive from analytic propositions. The discussion in Sec. 4.2.1 of the correspondence between propositions and rules of inference – where the rules of inference provide the norms – indicates one way in which this might occur. But now we must be careful. A key feature of norms is that they can be violated; let us ask what this involves in the present case. Consider a Sellarsian material rule, understood as a norm that licenses an inference. We have the proposition "All *A* are *B*," and a rule that allows us to infer "*x* is *B*" from "*x* is *A*." We violate the norm if we identify an item as an *A* and conclude that it is not a *B*. Doing so amounts to acknowledging an *A* that is not a *B*, and this is equivalent to accepting a counter-instance to the generalization. This makes

sense if "All *A* are *B*" is synthetic, but not if it is analytic: analytic propositions do not have counter-examples. In other words, given that the possibility of violation is a constitutive feature of norms, and if norms are related to propositions by the kind of mapping we have considered, then analytic propositions cannot provide the basis for norms. If all a priori knowledge is encompassed in analytic propositions, then there are no a priori *norms*.[6]

We can see more clearly what is at issue here by considering a situation in which analytic propositions do seem to provide norms. When we teach the meanings of words in our language to children we take the standard definitions of these words as normative; a child who does not use a term in accordance with the accepted definition has made a mistake and is corrected. Yet definitions are a paradigm case of analytic propositions. But the normativity in this case does not arise from the definitions; it arises from a *decision* (whether individual or social) to use a term in a particular way. In the absence of such a decision there is nothing normative about a definition. This parallels the situation for GAs: their normativity derives from the decision to use specific synthetic propositions in a particular way.[7]

The key point I want to draw out of this discussion is that the debate over the analytic-synthetic distinction is a debate over what we should include in our system of epistemic concepts. We propose alterations in our epistemic concepts as we come to understand more about the contents of established concepts, learn more about our epistemic situation, and as that situation changes. One difference between Kant's situation and ours is that we know much more about science than Kant did – both because a great deal of science was created only after Kant's death, and because we know more about the early history of science than was known during Kant's lifetime. This increased knowledge – including transformations in science that Kant did not foresee – are part of our changing epistemic situation. As this situation changes we should expect to find that some epistemic concepts developed in earlier contexts are not satisfactory, and must be replaced.

## 8.2 Propositional Knowledge

I am now going to use TC to explore a central epistemic concept: KNOWLEDGE. This is clearly a concept with both normative and descriptive dimensions; I will consider the descriptive side first, beginning with the concept's systemic role. However, since the actual use of the word "knowledge" is controversial, I am going to avoid this term in the present section, except when citing views that use it. I am concerned here with an epistemic state that is both especially valuable and pursuable by human beings. Since much of my argument will involve comparisons of the relative preferability of various epistemic states (e.g., belief and true belief), I will refer to the state of interest as a "superior epistemic state," which I will abbreviate $E_{sup}$. This state will not be the best possible epistemic state that we can imagine,

but one that is highly desirable. By way of contrast recall that in *Theaetetus* Plato begins his discussion of knowledge by introducing two criteria that a state must achieve to count as knowledge: it must be infallible and "of the real." As I read the dialogue, these were intended as necessary and sufficient conditions for knowledge, although for present purposes it will suffice to treat them as necessary conditions. There is some ambiguity in just what Plato intends by the second criterion. Since my aim is to illustrate a point, not to defend an interpretation of the dialogue, I will consider only infallibility.[8]

We can conceive of an epistemic state that is infallible – being able to conceive of such a state is a prerequisite for recognizing that various states are not infallible. It also seems clear that, with the exception of some fairly trivial cases, infallibility is not in the reach of human beings. Let me press this point about triviality. Suppose we grant that some cases, such as my awareness that I have a headache (cf. Sosa 1980) are infallible. Cases of this sort do not provide a basis for building the kind of epistemic structure I am exploring in this book. We would be pretty pathetic cognitive beings if our major epistemic achievements consisted of the ability to recognize a headache, or to be indubitably conscious of a pink sense datum, while discounting as comparatively insignificant all we have learned about planetary systems, anatomy and physiology, the causes of disease, the chemistry and physics of elements and compounds, and so on, on the grounds that these beliefs are fallible. As Popper argued long ago, fallibility is the price we pay for pursuit of significant beliefs, and I will continue to take examples of the sort discussed in Ch. 2 as our major epistemic achievements. In these cases infallibility is not a relevant option; it is not an end we can seriously pursue.

Consider another example of a cognitive state that we can conceive of but that is not relevant to our actual epistemic pursuits. Descartes was clear that the best possible epistemic state is omniscience, and recognized that it is not in our grasp. If we were omniscient none of the familiar epistemological problems, such as finding a methodology for certifying knowledge, or assessing the limits of human knowledge, would arise. Again, the fact that we can make these points indicates that we have a concept of omniscience, but it would be a mistake to conclude that we are epistemic failures if we must settle for anything less. My aim here is to explore a high-level epistemic state that we can pursue, and I am using $E_{sup}$ to label this state. I turn now to some candidates from the literature. I will focus discussion on the view that $E_{sup}$ is justified true belief (henceforth JTB), which was long accepted by epistemologists as an analysis of knowledge, but has become a subject of debate in recent decades. For now I will consider only candidates for $E_{sup}$ that can be expressed in propositions; I will discuss non-propositional candidates in Sec. 8.5.

The notions of justification and truth are central to the status of JTB as a superior epistemic state.[9] We all have many beliefs, but having a belief about

some subject matter is not a particularly significant epistemic achievement. A bare belief may be either true or false, and true beliefs are epistemically preferable to false beliefs. If I have beliefs about whether it is safe to cross a particular bridge, or about the rate of interest that the money in my bank account will receive, or about whether my friend will keep our appointment, I prefer true beliefs. So we include truth in our description of $E_{sup}$. We add justification as a further condition because I am in a superior epistemic state if I have good reasons for my true belief than if I lack such reasons. Plato provided the major argument for this claim in *Theaetetus*: If I arrive at a belief by a random process, or because I have been seduced by misleading advertising, the belief may well be true, but it has a certain instability. A different random process, or a different piece of advertising, may dissuade me as easily as the original persuaded me. But if I understand why my belief is true, I will not be so easily dissuaded.

There is a second reason why justification is an important epistemic desideratum: propositions rarely wear their truth-values on their faces, and justification provides our basis for assessing which propositions are true and which are false. This is particularly important if we wish to improve our epistemic situation. Suppose I have some money to invest, and I am considering whether to buy a particular stock. If I arrive at the correct belief by accident or luck I will make exactly as much money as I will make if I arrive at this belief on the basis of appropriate reasons. But my recognition of this fact does not help me *decide* whether to invest. When I must make a decision I would like to have reasons for making my choice – such as indicators that reliably predict the future course of the stock. I am in a better epistemic position if I have such indicators than if I lack them. Overall, then, *justified* true beliefs are epistemically superior to beliefs that are just true. Still, justifications are fallible: the best available evidence at a given point in time may support a false belief. Thus, once again, justified *true* beliefs are preferable to beliefs that are just justified.

Once we associate $E_{sup}$ with belief, truth, and justification we begin building up the implications that are characteristic of a conceptual system: To say that I am in a state of $E_{sup}$ with respect to $p$ implies that I believe $p$, that $p$ is true, and that I am justified in believing $p$.[10] Other implications depend, in part, on the accounts we give of truth and justification, so further insight into $E_{sup}$ requires that we consider these concepts; I will devote a section to each.

A complication appears when we consider instantiation conditions for $E_{sup}$. Since application of this concept to a proposition requires that the proposition be justified and true, the ICs for $E_{sup}$ depend on the ICs for justification and truth. However, our reasons for thinking that a proposition is true are often just our reasons for thinking it is justified (cf. Sec. 5.5). As a result, an account of the ICs for $E_{sup}$ will center on the account for justification. Still (I will argue below) the claim that $p$ is true says more than just that $p$ is justified, so an account of truth will be a key feature of an account of

$E_{sup}$. I postpone further consideration of these issues until our discussions of justification and truth. For now I want to consider an alternative view of our main epistemic goal.

Sartwell has argued that in "our ordinary use of the term" (1992: 167, n. 1) "*knowledge is our epistemic goal in the generation of particular propositional beliefs*" (1992: 167), and that this goal is just true belief. Sartwell agrees that justification is important in the pursuit of this goal, and cites such diverse epistemologists as Bonjour, Goldman, and Moser to support the claim that justification is linked to truth: to say that a claim is justified is to say that we have reasons for believing it is true. But, Sartwell argues, this makes justification a criterion for truth, "a test of whether someone has knowledge, that is, whether her beliefs are true" (1992: 174). Justification, "(a) gives procedures by which true beliefs are obtained, and (b) gives standards for evaluating the products of such procedures with regard to that goal," but justification is not "a logically necessary condition for knowledge" (1992: 174). Now this may be a correct account of ordinary usage, but if so, then knowledge is not $E_{sup}$ since we are epistemically better off if we have a true belief that has passed appropriate standards of evaluation, than if we merely have a true belief – for the reason Plato gave. Goldman (1999: 23–24) is pointing in the same direction when, partly in response to Sartwell, he describes true belief as knowledge in the weak sense, and true belief plus some other condition such as justification as knowledge in the strong sense. In other words, if according to current usage "knowledge" just means "true belief," then knowledge is not so hot. If our current everyday epistemic goal is just to arrive at true beliefs, we should seriously consider adopting the more stringent goal of seeking justified true beliefs. Note that we cannot *stipulate both* that a state is *especially desirable* and *what that state is*. If we stipulate that a term refers to states having one of these characteristics, it is an open question whether those states have the other characteristic. Thus if we use "knowledge" to refer to $E_{sup}$, then knowledge is not just true belief; if "knowledge" means "true belief" then we can do better than just seeking to acquire knowledge.

As we saw in Sec. 5.5, Gettier (1963) presents the other side of this coin when he argues that JTB is necessary for knowledge, but not sufficient. In effect, Gettier assumes that "knowledge" refers to an especially desirable epistemic state, and the force of his examples lies in their meeting conditions for JTB while describing situations in which our epistemic state is not particularly good. Two main lines of response to the Gettier examples underline this point since both seek more stringent requirements, either by agreeing that JTB is not sufficient for $E_{sup}$ and proposing an additional condition; or by holding that JTB is sufficient for $E_{sup}$, but denying that we have justification in the Gettier cases. I am not going to enter into the massive Gettier literature, but I am going to consider justification in greater detail.

## 8.3 Justification

I will approach this subject by examining a currently influential account of justification: Goldman's *reliabilism*. Goldman introduced reliabilism in his 1979 article "What is Justified Belief?" (references are to the 1992 reprint in *Liaisons*). Goldman emphasizes that he aims to explicate "our ordinary standards" of justification: "Unlike some traditional approaches, I do not try to prescribe standards for justification that differ from, or improve upon, our ordinary standards" (105). In other words, Goldman is describing a normative concept (105). But he also specifies several requirements that a successful account must meet. First, the account must explain "in a general way why certain beliefs are counted as justified and others as unjustified" (105). Second, he is seeking "a substantive set of conditions that specify when a belief is justified," and do this in non-epistemic terms (105). Third, Goldman seeks "an explanatory theory, i.e., one that clarifies the underlying source of justificational status. It is not enough for a theory to state 'correct' necessary and sufficient conditions, its conditions must also be appropriately deep or revelatory" (106);[11] it must make clear what it is that makes these *justified* beliefs. "A theory of justified belief of the kind I seek must answer this question, and hence it must be couched at a suitably deep, general, or abstract level" (106). Fourth, Goldman assumes "that a justified belief gets its status of being justified from some process or properties that make it justified" (106), but leaves it open whether someone who has a justified belief must know it is justified, can state or give a justification for it, and whether "there is something 'possessed' by the believer which can be called a 'justification'" (106). Finally, Goldman requires that an analysis of justification be given in a recursive format (107). This format has three elements: a *base clause* which states some initial instances of justified beliefs; a *recursive clause* that describes how to generate further instances out of established instances; and a *closure clause* which states that instances generated by the two previous clauses are the only instances of the concept.[12] Goldman's base clause will state a sufficient condition for justification: it will be of the form "If $p$ is . . . , then $p$ is justified." The requirement that no epistemic terms appear in a definition applies only to the antecedent of the base clause; the closure clause asserts that the conditions given are also necessary. Goldman assumes that the everyday concept of justification already meets these constraints, including the requirement that it is constituted by a set of necessary and sufficient conditions.

Goldman begins his discussion by considering and rejecting some base clauses that have been proposed by others (107–12); he uses two lines of argument to criticize various proposals. In some cases he argues that the proposal fails because it includes an epistemic term. When this desideratum is met, Goldman criticizes proposals by arguing that they include among the justified beliefs some that *we* intuitively recognize to be unjustified. For example, he rejects the proposal that I am justified in believing $p$ if I am

psychologically incapable of doubting *p*: "A religious fanatic may be psychologically incapable of doubting the tenets of his faith, but that doesn't make his belief in them justified" (107). Sometimes Goldman uses imaginary scenarios for this purpose. This is all standard analytic technique. In particular, the appeal to what we intuitively take to be justified provides the basis for Goldman's claim that he is analyzing our ordinary concept of justification.

Goldman moves toward his own account by considering why the attempts he has examined go wrong. He suggests that the failed attempts lack an appropriate *causal* requirement, where causes include both initiators and sustainers of beliefs (112–13). The failed proposals either do not include an account of the cause of the belief, or they include an inappropriate cause. The key feature of inappropriate causes is that they are unreliable: "They tend to produce error a large portion of the time" (113). This leads to Goldman's own proposal:

> The justificational status of a belief is a function of the reliability of the process or processes that cause it, where (as a first approximation) reliability consists in the tendency of a process to produce beliefs that are true rather than false.
>
> (113)

Justified beliefs, then, are those that are caused by reliable processes. The remainder of Goldman's paper is concerned with refining this first approximation, working it into the recursive format, and responding to objections. I will discuss only a few points from this part of the paper.

Refinement proceeds by considering objections and adopting one of two stances towards them. *Sometimes* Goldman adjusts his account to accommodate the criticism. For example, Goldman notes that "The Pope asserts *p*" is not a process that justifies belief in *p* since "we would not regard the belief-outputs of this process as justified" (116). The problem with this process is that it is not "content-neutral" since it refers to a specific individual. Goldman thus requires that a justification-conferring process be content-neutral. Goldman also asks whether we should consider the relevant processes to extend outside the believer's organism, or should limit ourselves to items in the organism, and consider whatever comes from outside as inputs to the process. He opts for the latter "with some hesitation" and offers the following "general grounds" for this decision: justified beliefs result from reliable cognitive operations, and "'*cognitive*' operations are most plausibly construed as operations of the cognitive faculties, i.e., 'information processing' equipment *internal* to the organism" (116). For example, in perceptual processes the photons, pressure waves, and such that impinge on the perceiver are considered inputs, rather than part of the cognitive process; that process begins when these external inputs act on a sensory system. In both of these examples Goldman starts with a point at which his initial

account is unclear and ends up with a more precise version. He is thus maintaining that the ordinary concept being analyzed is also precise on these matters. This point is underlined by Goldman's second strategy. *Sometimes* he leaves the analysis vague on the grounds that the concept being analyzed is itself vague. For example, Goldman invokes vagueness to justify leaving us without a precise account of how reliable a process must be to confer justification, and leaving it open whether reliability should refer to long-run frequency or propensity (114–15). I want to consider several points about Goldman's account.

Assuming that we are analyzing an everyday concept C, let us ask whether it is appropriate to use concepts that ordinary possessors of C do not have. In developing his analysis Goldman makes use of INFORMATION PROCESSING EQUIPMENT, which is a specialized concept of recent vintage, one that is not available to all people, in all cultures, and in all historical periods. Presumably Goldman holds that JUSTIFIED BELIEF is available to all. Goldman could respond that this poses no problem because INFORMATION PROCESSING EQUIPMENT does not appear in the analysis; it appears only in his account of why he constructs the analysis as he does. To be sure, this would make part of the motivation for the analysis unintelligible to many people who presumably have the concept being analyzed, but this may not be a significant objection since ordinary folk use concepts, they do not analyze them. Yet there are further consequences that we should take seriously. Goldman's justification for this aspect of the analysis depends on developments in fields other than conceptual analysis, so the ability to justify the analysis may only appear at some historical stage in our overall epistemic development – a development that is not yet complete. If people at an earlier point in history arrived at the same analysis, they would not be able to give the same reasons as Goldman for at least part of the analysis. People at a later stage of development might prefer a different justification. But this should make us wonder whether the analysis itself (not just its justification) might come out different in a different historical setting.

The question of which concepts are appropriate also arises in the content of Goldman's analysis. Here is the final version of his base clause – although he notes that he has omitted "certain details in the interest of clarity" and that the analysis still faces problems which may require further elaboration.

> If *S*'s belief in *p* at *t* results from a reliable cognitive process, and there is no reliable or conditionally reliable process available to *S* which, had it been used by *S* in addition to the process actually used, would have resulted in *S*'s not believing *p* at *t*, then *S*'s belief in *p* at *t* is justified.
>
> (123)

This analysis includes the concepts COGNITIVE PROCESS and CONDITIONALLY RELIABLE PROCESS, which are not naïve concepts generally available to people across all times and cultures.[13] It seems appropriate to insist that an

*analysis* of the content of concepts that people actually possess should not include concepts they do not possess. Moreover, given Goldman's grasp of problems that are not considered in everyday thought, and his mastery of a rich set of sophisticated and specialized concepts from philosophy and psychology, why should we think that *his* intuitions about which beliefs are justified match those of people who do not share his sophistication? These issues do not arise if we view Goldman as offering a proposal as to what should count as justification given a sophisticated understanding of our epistemic situation.[14]

Consider another feature of Goldman's account: being produced by a reliable process is a sufficient condition for justified belief; a justified believer need not be aware of how the belief was produced, or even that it is justified. "Just as a person can know without knowing that he knows, so he can have a justified belief without knowing that it is justified (or believing justifiably that it is justified)" (118). This feature of the account captures one of Goldman's main motivations. People who are not highly educated in the sciences survive in the world, earn a living, raise a family, and so forth. If the requirements for justified belief are too demanding, we will have to conclude that many of the beliefs at the basis of all this successful activity are not justified; this seems unacceptable. A person who is taking a walk and sees a tree in the path acquires a perceptual belief about this tree and will walk around it, rather than attempting to walk through it or push it aside. This belief is surely justified, even if the believer cannot tell us anything about the physical and physiological processes that caused this belief, or give reasons for believing that perception is generally reliable. A similar point holds for other cognitive processes such as memory and simple calculations. Indeed, to say that such beliefs are not justified is to issue a negative evaluation – to suggest that the person in question has failed to meet some epistemic obligation. Yet if someone is walking down a street with the aim (among others) of not bumping into various objects, there is no more to be done besides looking, listening, and walking. There is no epistemic evaluation that this person has failed to carry out. Goldman's reliabilism is one of a cluster of current views that reject overly sophisticated accounts of justification which imply that many typical everyday beliefs are unjustified.

We can derive some interesting perspective on this motivation by returning to Goldman's extended *argument* for his account of justification. This argument is, after all, an attempt to justify the account, and does not proceed in its own terms. Goldman does not just state the analysis and announce that it is justified if it was arrived at by a reliable process. Rather, as is typical of philosophers defending an analysis, he provides explicit reasons for thinking that his account is correct, defends it against some criticisms, modifies it in response to other criticisms, and challenges competing accounts. In other words, when Goldman endeavors to justify his analysis of JUSTIFICATION, he adopts a much more demanding standard of justification than the one encapsulated in reliabilism. In fact, the standard Goldman uses

is of the same general type as scientists use when proposing an account of some subject matter: evidence is explicitly introduced and evaluated while competing proposals are criticized. As I noted in the previous section, the need for explicit reasons is especially clear in cases that require a decision – for example, when a scientist is trying to decide between competing hypotheses, an investor is considering where to commit money, a physician is evaluating a diagnosis and plan of treatment, or a philosopher is assessing an analysis. It is of no help to be told that we should accept the alternative that was arrived at by the most reliable process. We need an account of what the process is and why it is reliable. In many philosophical, scientific, and technical situations the arguments required to justify a belief are subtle and complex. I agree that arguments of this sort are neither required nor appropriate in all situations. There may well be a considerable range of cases in which different standards of justification are appropriate so that, as Kim observed in the case of causation, accounts of these standards are "so widely divergent that one wonders whether they are all analyses of one and the same concept" (1995: 112). One might reply that Goldman has given an account of a basic level of justification that is appropriate for a wide variety of cases, and that the more sophisticated forms of justification yield beliefs that are better justified because they are arrived at by processes that are more reliable.[15] However, this will not do. On Goldman's account any belief that meets the basic standard is justified, while additional factors concern only degrees of justification. But the basic standard is not sufficient to justify philosophical or scientific beliefs. If reliabilism captures *what it means to say* that a belief is justified, then we will have to conclude that many justified beliefs are not rationally acceptable. At this point we have left behind the systemic role of JUSTIFICATION which provides a common thread through the history of the subject.

TC suggests a different way of describing the situation we have arrived at: the term "justification" does not label a single concept, but a conceptual system that is embedded in our epistemic system. Instead of a single concept of justification, we need a number of evaluative concepts that are appropriate to different situations. Elsewhere Goldman seems sympathetic to this view:

> No unique concept of justifiedness is embraced by everyday thought or language. . . . I present several distinct accounts of justifiedness, each with some hold on intuition. However, these accounts form a close-knit family; so there seems to be a core idea of justifiedness, which my theory will seek to capture.
>
> (1986: 58–59)

TC indicates that this core is found in the shared systemic role of these concepts.[16]

The limitations of reliabilism are particularly clear if we seek to *improve* our current epistemic situation. The development of science includes assessments

of our cognitive processes, discovery of their limitations, and often the discovery of ways to improve our overall reliability. In many cases we have developed technologies that yield greater reliability than we can achieve without external aids. For example, when Galileo first used his telescope for astronomical purposes he encountered a direct conflict between telescopic and naked-eye observations. He responded by arguing that telescopic observations are preferable for astronomical purposes because the unaided eye has a defect that makes it subject to a particular type of illusion when we observe small bright points of light, and that the telescope corrects this defect (Brown 1985). A century or so later, when telescopic observation was well established, Bradley introduced a method of timing celestial events that depends on looking through a telescope while listening to a clock tick seconds. This was considered highly reliable until Bessel discovered individual variations among astronomers using this method, and variations in individuals over time (Boring 1950 Ch. 8; see Brown 1987 Sec. 6.5 for further discussion). We now replace the human observer with more reliable electronic equipment. No doubt these aids are external to the organism, but they combine with our built-in capacities to yield results that are more reliable than we can achieve with our organic processes alone. We have been making improvements of this kind at least since the invention of writing allowed us to improve on our memories, and the process continues. From the perspective of reliability, these aids are not second-class contributors (cf. Brown 2005; Clark 1997). In many situations a failure to use these aids would be an epistemic failing resulting in beliefs that are not justified. Many of these developments rely on a detailed understanding of the nature and limits of evolved cognitive processes, an understanding that requires explicit accumulation and assessment of evidence – knowledge that would not be available if we limited ourselves to reliabilist-justification. As a result, standards of justification develop along with the development of this knowledge, and the concepts we use to describe belief-worthiness undergo change along with other members of our conceptual repertoires.

Returning to Goldman, I suggest that although he conceives of his project as analyzing a familiar concept, we can adopt a different interpretation of his outcome: He offers a *substantive proposal* about what *should count* as justification in some circumstances. Making such proposals is an important and appropriate project. We can underline the importance of this project by recalling that Gettier-type cases typically involve non-deductive justification, and it has long been clear to epistemologists that we do not have an adequate account of this kind of justification. Developing such an account is not a matter of giving a more accurate analysis of what we already know. Such a claim is no more plausible than claiming that we all know how to carry out statistical analyses of data, and that statisticians develop explicit formulations of this common knowledge. Rather, proposals for how to carry out such justifications are in order; when we find a proposal acceptable we can build it into a concept. Given a cluster of justification-concepts that share a

common systemic role, differences among these concepts will be determined by differences in implications and instantiation conditions.

## 8.4 Truth

I turn next to TRUTH which I treat as an epistemic concept only in the sense that it plays a role in epistemic thought; I am *not* adopting what is commonly known as an "epistemic" account of truth. Such accounts define truth in terms of justification – typically in terms of what will be justified under some specified conditions; I will criticize that approach in this discussion. Indeed, I am going to *advocate* the importance of a correspondence account of truth – *in a sense to be explained as we proceed.* The discussion does not aim at an account of an existing concept; rather it should be viewed as an explication in Carnap's sense (Sec. 1.3). Towards this end I will deploy the full machinery of TC in my discussion, beginning with systemic role.

### *8.4.1 Systemic Role*

Consider some typical descriptive propositions from everyday life and science: "The Eiffel tower is in Paris," "Kant was born after Hume," "Top quarks exist," "Energy is conserved in all physical interactions," "In a vacuum, all photons move at the speed of light for all observers." Typical descriptive propositions say something about specific items; what these propositions say may be correct or incorrect. Incorrect descriptions include propositions such as "The Eiffel tower is in London," or "Hume and Kant were born on the same date." We need descriptive concepts that mark the difference between propositions that do and do not correctly describe their subject matter; TRUE and FALSE play this role in epistemic conceptual systems. (I am assuming bivalence for now; I will return to this topic below.) There is nothing mysterious about this. News reporters, physicians, scientists, stock brokers, and others make statements that are correct or incorrect independently of whether anyone believes those statements, and independently of whether anyone has grounds for such belief. We want to determine which propositions are true since these are worthy of belief and should guide our actions.

Other concepts, such as accurate, correct, and like-it-is, play roles similar to truth, but truth has a special function: it applies to propositions and provides an absolute evaluation. Unlike accuracy and correctness, truth is not a matter of degree. Falsity is more complex since some false propositions (e.g., false quantitative propositions) may be more or less accurate than others. This suggests that classifying a proposition as false may raise additional questions, but we may still treat falsity itself as a concept that does not admit degrees. In order to underline the special role of truth and falsify note that we also make epistemic evaluations of items besides propositions,

such as pictures and maps. A picture or map typically provides a great deal more information about an item than a proposition does. Pictures and maps will also be accurate in some respects, but inaccurate in others, and it is often important to be clear about just where these are accurate, and whether one picture or map is more accurate (for a given purpose) than another. Many propositions are quite succinct, making only one claim about an item. And while complex propositions can be more or less accurate, such propositions can usually be broken down into a set of simple propositions. This is an important base case because a simple proposition making a single claim has a clarity and precision that maps and pictures often lack. In this discussion I will use true and false primarily to describe simple propositions. The absoluteness of truth and falsity is an artifact of the limited information carried by these propositions.

Truth and falsity play a further role. While our application of these concepts depends on the available evidence, the status of a claim as true or false typically transcends this evidence. Even when we have strong evidence for or against a claim, we can contemplate the possibility that our assessment is not correct; we need concepts in our epistemic system that allows for such thoughts. In many situations *saying* that a proposition is true has little point. If my broker says that a particular stock will double in the next two months I may ask if that is true, but the broker's assertion that it is true will not give me any new information. New information requires that I seek other sources. But the fact that I may want to carry out such an inquiry underlines the importance of being able to ask if the broker's claim is true.

One issue of substantial interest is whether every proposition is true or false, or whether some propositions are undetermined under some circumstance. Contingent statements about the future provide the classic example of propositions that may not fit this dichotomy, although they are not the only occasion for raising the issue. We have seen that many concepts are open-ended. This is often the case for everyday concepts. For example, many people are unable to decide whether a telephone or a rug is furniture. One plausible explanation for this situation is that everyday concepts are developed only as far as needed for existing purposes; they are not given the kind of rigorous definitions we encounter in mathematics. As a result, we may not know how to categorize an item because the question is not important and has not come up before; thus no answer is built into our current concept. Whether to allow for propositions that are neither true nor false is a question about what we should include in our system of epistemic concepts. We might approach the question by exploring some available conceptual system, but whatever the outcome of this exploration we can still ponder whether there are reasons for introducing an alternative system. In the remainder of this discussion I will focus on truth, and leave it open how many contraries the concept has. Differing views will generate somewhat different conceptual systems, and thus somewhat different truth-concepts. This is also a second respect in which falsity is a more complex notion than is truth.

Truth also has a purely logical role. Suppose I want to express my epistemic approval of all of Pat's beliefs, or all the consequences of an axiom set. These are large sets (the latter is infinite); I cannot express this approval by listing the propositions I wish to affirm. Truth allows me to make the desired assertion. I can affirm that all of the propositions in a set are true, no matter how large that set. In other cases we want to express an evaluation of a proposition whose exact content we do not remember – such as what Pat just said. Again, truth allows us to do this by saying, "What Pat just said is true." There are minimalist accounts of truth that consider this logical role to be the main, perhaps the only, role the concept has (e.g., Horwich 1998). Clearly, I disagree. I will return to minimalism in 8.4.2.

The content of a descriptive proposition is a claim about some subject matter; a true proposition provides information about that subject matter.[17] This idea of "providing information" is the key point of a correspondence account of truth: A proposition corresponds to its subject matter when it carries information about that subject matter. This requires no mysterious "third thing" that relates the proposition to its subject matter, nor is "correspondence" being offered as a definition of "truth." *As I am using the term*, "correspondence" is located in the metalanguage we use for discussing conceptual systems; it occurs in a *metalinguistic commentary* (Sec. 5.2) on the role that truth plays in our system of epistemic concepts. The function of this term in a commentary can be brought out by considering some analogous epistemic situations in which one item may correspond to another.

As a first step, we can describe two distinct items as corresponding in those respects in which they are identical. Consider the Eiffel Tower and a model of the Eiffel Tower; the model and its prototype will share some features but differ in others. As an extreme example, suppose that a detailed copy of the Eiffel Tower has been built in Texas. Every dimension of the tower, the materials, their degree of wear and corrosion on a specific date, and so forth, have been matched. To the extent that the model corresponds to the original, the model can serve as an *epistemic proxy* for the original: we can learn a great deal about the Eiffel Tower by studying the model.[18] But there will always be some properties in which the model does not correspond to the prototype. For example, we cannot learn the latitude and longitude of the original, or the distance of the Eiffel Tower from the Seine, or whether the Tower is currently wet from dew, by studying the model. In a similar way, since identical twins correspond in such characteristics as height and facial features, one twin will serve as well as the other if we are interested in these features. But the twins are distinct individuals, so there will be many features in which they may not correspond, features such as current location, marital status, and profession.

The greater the correspondence between a model and its prototype, the more we can learn about one by examining the other. But there is a pragmatic aspect to our use of models. We typically create a model for a specific purpose, and this purpose plays a role in determining the respects in which

we want model and original to correspond. A tourist's model of the Eiffel Tower that weighed as much as the actual tower would be a very poor model given its purpose. Here we might prefer a scale model. We might construct a plastic model that includes a distinct piece for each distinct structural member of the Tower, with the length of each member reduced in the same proportion, but with no attempt to provide a scale model of the cross-section of each member. There might also be no consistent relation between the weight of a part of the model and the weight of the corresponding part of the original. In this case the number of structural members of model and original would correspond, while there would be no correspondence between the weight of members in the model and the original. The lengths of members in the model and original also correspond in our epistemic sense: if I know the scale, I can determine the length of a piece of the tower from the model. It will take a bit more work to extract this information than in other cases I have been considering, but I can learn about this feature of the original from my model (and conversely).

A true proposition corresponds to its subject matter in our epistemic sense: the proposition provides information about that subject matter. Given that the proposition is true, it is not necessary for me to check that subject matter – the proposition will do. True propositions are especially useful when they provide information about items we cannot check for ourselves, or would not want to check. A trip to India is too expensive; on the date at which this is being written Baghdad is too violent; I do not have the skills to use the Hubble telescope. Propositions are also easier to carry around than models, even scale models. Our epistemic sense of correspondence is the sense in which true propositions *represent* their subject matter. In a similar way, a map is a kind of model that represents some features of its subject, and can serve as an epistemic proxy for those features. Note again that models often carry a good deal of information but also misrepresent in some respects. Thus in using a model it is important to know the respects in which the model represents the original. Since simple propositions present only a limited amount of information about an explicitly stated subject, parallel questions do not arise.[19]

With this correspondence account of truth in hand, we can now consider why we need this concept in our epistemic system. Some philosophers adopt a coherence account of truth: the claim that a proposition is true reduces to a claim about its relations to other propositions. But most propositions make claims about items other than propositions. To say that a proposition is true is to say that it correctly describes its subject, even when this subject is not relations among propositions. Coherence may play a role in *assessing* whether a proposition is true – a coherence account of justification is compatible with a correspondence account of truth – but this is different from claiming that to describe a proposition as true is to make a claim about its relations to other propositions. Even when proposition $p$ is about another proposition $p^*$ (say, about its consistency), $p$ is true just in case what it says

about *p** is correct. This applies to many self-referential propositions. Consider, "This proposition is consistent." The proposition is true if it correctly describes its subject – which happens to be itself. Its truth does not depend on its relations to other propositions.

Next, consider a pragmatic account of truth in the original sense of "pragmatic": to describe a proposition as true is just to say that it is useful. It is not clear that anyone ever actually held this view, and its key defect is clear enough: Propositions that are, in some way, useful, need not be true, and those that are true need not be useful. Moreover, if someone says that a proposition is useful, it is appropriate to ask if this claim about the proposition is true. The correspondence concept of truth embodies the conceptual resources we need to make this point.

In recent years "pragmatic" has taken on a new meaning: It is now used to label accounts that make truth relative to some body of beliefs. These so-called *epistemic* accounts of truth are the most common contemporary alternatives to a correspondence view. They range from the extreme version Plato attributes to Protagoras – what each individual believes is true for that individual – to more moderate views that relativize truth to some social, historical, or professional group. In *Theaetetus* Plato offers two arguments against extreme relativism; I will consider each of these arguments and examine how it extends to less extreme versions. Recall that my aim here is limited to considering what these arguments tell us about the systemic role of TRUTH.

Plato first raises a reflexivity problem: According to Protagoras what each individual believes is true for that individual. But, in effect, Socrates replies: "I believe that Protagoras is wrong, and by his own lights Protagoras must acknowledge the truth of my claim. Since I do not accept a relativized account of truth, I do not have to acknowledge that Protagoras' view is true for anyone." This response is a prototype for a wide variety of reflexivity arguments; reflexivity problems arise for all relativized accounts of truth, even those that are less extreme than Protagoras' version. When someone proposes a truth-concept that is relative to some individual or group, a variation on Moore's open-question argument is always in order: it is legitimate to ask if that claim is true. This is a serious challenge because those who make claims of the sort "Truth is relative to . . . " (TR) are typically not claiming that TR is relative. Consider, for example, the claim that truth is relative to some group (TRG). Those who hold TRG do not usually claim that TRG is a true relative to them, but that others may deploy a non-relative account of truth. The usual claim is that truth is relative – period; and this claim assumes a non-relativized truth-concept. As Putnam notes, "No relativist wants to be a relativist about *everything* whatever" (1981: 158). Even when we are describing beliefs and practices of a particular society, we seek descriptions that are true in the correspondence sense; we do not claim that those descriptions are true for a particular group of researchers, but need not be true for others.

Individuals and groups have a vital interest in truth understood in the correspondence sense. It is not uncommon to find people who harbor beliefs about themselves, their environments, and other people that lead them to disaster – as measured by their own criteria. Including a correspondence concept of truth in our epistemic framework will not eliminate this problem, but it will help us think clearly about why it occurs. A relativized notion of truth can also be introduced, but this is misleading since other concepts such as BELIEF or WARRANTED BELIEF or EXTRAORDINARILY WELL SUBSTANTIATED BELIEF will do the required job while reserving truth for another crucial task: providing the conceptual resources we need to consider whether a belief accurately describes its subject matter. (Huw Price 2003 develops a similar view, although in a different context.)

We need to retain this special conceptual role even if we go to a different extreme and relativize truth to what all will agree to under some idealized condition. A truth-concept that is relativized to an asymptotic future state in which all agree still generates reflexivity problems because it remains legitimate to ask if they have gotten everything right. This pervasive reappearance of reflexivity problems when we attempt to eliminate the correspondence-truth concept is a key indicator of the central role this concept plays in thinking about epistemic matters. One might suggest that in a properly constructed alternative conceptual system reflexivity questions will not arise, but to my knowledge no one has constructed such a system; there is no guarantee that a workable system of this sort can be constructed.

Plato's second reply to Protagoras consists of constructing two mutually contradictory claims about a future event and noting that only one can be true. Consider two possible readings of this response, tied to two different interpretations of Protagoras' thesis. We might interpret Protagoras as proposing a relativized concept of truth *in addition to* the correspondence concept. Plato's prediction gambit amounts to showing why a non-relativized concept of truth is still required. Alternatively, we might read Protagoras as attempting to assimilate truth to belief. In this case Plato's argument shows why we need a distinct concept of truth since not all beliefs are true.

Our discussion suggests a further systemic role for truth that introduces the prescriptive aspect of the concept: truth specifies a basic epistemic end; the reasons for this are not at all arcane. We seek truth because we want to find out how things are. In purely intellectual terms this is an end in itself. It is also an end of considerable practical importance. Whether we are balancing a checkbook, determining the carrying capacity of a beam, deciding if the beer is cold, assessing the best treatment for a disease, considering whether neutrinos have mass, or what have you, we seek the correct answer to our question – that is, we seek the truth. This point holds even when we limit ourselves to purely pragmatic considerations. Consider an engineer who must decide if a particular approximation is sufficient for present purposes. The task is to determine the truth-value of the claim that the approximation is adequate for this purpose. Successful pursuit of a prag-

matic goal depends on the truth of the claim that the proposed means will lead to the desired end. So although we often specify epistemic ends in terms of other values besides truth, truth will reappear.

The pragmatic importance of truth becomes particularly pressing when we recognize that decisions to act on the basis of accepted beliefs typically involve inferences concerning what we might encounter in the future, or in circumstances we have not yet checked. But inferred conclusions are no more reliable than their premises, so true beliefs provide a necessary component of a maximally reliable prediction technique. Whether our inferences are deductive or inductive, we have failed to provide maximally reliable grounds for accepting a conclusion unless the premises are true. We need the concept of correspondence-truth to make this point.

### *8.4.2 Extra-systemic Relations*

Since truth is both descriptive and normative we must consider both instantiation conditions and required actions. We must also keep in mind that TRUTH occurs in a higher-order conceptual system that we use for considering the epistemic status of propositions – propositions that typically do not occur in this system; rather, names for these propositions occur.[20] As is customary, I use quotation marks to form the required name, although I omit quotation marks when the context is clear. Inferences *within* the epistemic system must be distinguished from moves between that system and the system in which *p* lives. Moves of the latter sort take us from propositions of the form:

"*p*" is true (T1)

to

*p*, (T2)

and from T2 to T1. I discuss these moves in the present section. Implications within the epistemic system are between T1 and other propositions of the form:

"*p*" is E,

where E is some epistemic concept; I discuss those in the next section.

Sellars treats the move from T2 to T1 as a kind of entry transition (e.g., MFC 425). TC requires that we extend the notion of an IC to include this case – which requires criteria for assessing if a proposition is true. But, we have seen, the working criteria for such assessment are just those for assessing justification even though a claim of truth goes beyond a claim of justification. Claims of truth require a decision, which introduces extra

content and extra epistemic risk.[21] In one respect this is a common feature of ICs: ICs specify the considerations that are relevant for assessing if a concept has instances. But the application of ICs to a specific case is usually fallible, and the conclusion that a concept has instances is subject to reconsideration. Yet truth differs from other cases because one of the systemic roles of truth *requires* this gap between the ICs and the application of the concept. Nothing in HOUSE, or GENE, or QUARK requires such a gap. The thesis that such gaps are common is embodied in our epistemic conceptual system, not in our conceptual systems for specific subjects. I have acknowledged that there are situations (such as Sosa's headache) in which the gap can be overridden. However, these are special cases, not the base case for epistemic reflection. In addition, the ICs for truth share a feature that we encountered in discussing prescriptive concepts: they are quasi-formal in that the detailed justification criteria vary from case to case. The content of *p* determines the relevant justification conditions. (This is in addition to truth's formal function noted above.)

The role of justification in providing the ICs for truth claims raises another issue. Given a variety of justification concepts, with different versions appropriate in different contexts, TC implies contextual variation in the conceptual content of TRUTH. To my mind this is a surprising result since I began with the view that a single concept of truth applies in all cases. We have here a classic epistemic situation: we can accept this consequence, or take it as a counter-example to TC, or seek some way of wiggling between the two. In my view the arguments on behalf of TC are sufficiently strong that I am prepared to accept the consequence, but underline its limitations. We are not using *exactly* the same concept when we attribute truth to propositions that describe an easily accessible perceptual situation, a general principle such as conservation of mass-energy, and the assertion that top quarks exist. These differences derive from the different ways in which these claims are justified, but ICs are just one element in the content of truth; other elements bring in a much greater degree of similarity. The systemic roles of truth remain constant, and we will find further constant features in the prescriptive side of truth and in intra-systemic implications. This result should be viewed as an example of the way TC allows us to map out the fine structure of identities and differences among concepts.[22]

The prescriptive side of truth requires that we incorporate propositions we consider true into our action and thought. Once we conclude that *p* is true, we should be prepared to assert *p* on appropriate occasions, use *p* as a premise in inferences, teach *p* to our children, and act on the basis of *p*. At the same time, the general fallibility of truth claims indicates that we should be prepared to revise our evaluations of propositions that we currently consider true. Situations that require revision are particularly clear when we are faced with two incompatible claims that we have evaluated as true – such as a consequence of a well-supported theory and an observational result. It

is part of the normative side of truth that situations of this sort cannot be permanently sustained – although there is no guarantee that the appropriate resolution will become clear rapidly.[23]

The move from classifying *p* as true to actually behaving as if *p* is true exemplifies a Sellarsian departure transition. Sellars explains this aspect of normativity in two different ways that are at least verbally at odds with each other. In his earlier writings Sellars uses "inference" to describe only moves that occur within a conceptual system, and holds that ETs and DTs are not inferences since they involve moves into or out of a conceptual system. In his later writings Sellars describes the DT involved in the concept of truth as an inference, but an inference of a special kind. In the usual case there is a sharp distinction between the premises and the principle that justifies an inference. *Modus ponens*, for example, may justify an inference but is not a premise of that inference; Carroll (1895) shows how deduction breaks down if we fail to observe this distinction. But when we infer *p* from T1, Sellars argues, T1 functions simultaneous as a premise and as the justifying principle (SM: 101–2). However, under either description the outcome is a step from evaluating and reflecting on *p* to incorporating *p* into our thought and action. A parallel point holds for adopting truth as our major epistemic end, which requires that we actually pursue truth in our epistemic endeavors.

A potentially confusing feature of Sellars' later discussions of truth is worthy of comment. In SM (e.g., Ch. 4) Sellars *defines* "truth" as "semantical assertability": *p* is true just in case the rules of the language in which *p* occurs allow the assertion of *p*. In one respect this conforms to Sellars' view that everything we firmly believe about a subject is built into a conceptual system. We adopt such a system when we have done all we can to justify the propositions that carry its content. At this point we have accepted a license both to assert any proposition *p* embodied in that system, and to make the epistemic assertion that *p* is true. To be sure, we have accepted *p* as true only relative to our adoption of a specific conceptual system, and we may later change our minds about that system. But our grounds for adopting a conceptual system are also grounds for accepting as true all of the propositions built into that system. This view of truth *seems* to leave us with a sweeping relativism: which propositions we describe as true appears to have been relativized to particular conceptual systems, and we are left without any vocabulary for asking if a proposition that is assertible in some conceptual system is, in fact, true. But Sellars is not a relativist – he is a scientific realist. Unless he is blatantly inconsistent, this relativization of truth cannot be the entire story.

Sellars attempts to reconcile the two positions by maintaining that while truth is relative to conceptual systems, some conceptual systems are more *adequate* than others. For example, in the case of physical science he takes it for granted that there is one maximally adequate conceptual system in each domain, and that we aim to arrive at that system by the long-run pursuit of science. In effect, we have a distinction between *truth* and *ideal truth*: Truth

is whatever is semantically assertable in some conceptual system and different conceptual systems yield different truths; ideal truth is what is semantically assertable in a completely adequate conceptual system. Sellars' account of an adequate conceptual system is an adaptation of Wittgenstein's notion of picturing in *Tractatus* (TC, SM Ch. 5). I will not pursue the details of this account, but I note that Sellars sometimes closely associates picturing with truth. At one point Sellars reminds us that even in the case of "first-level matter of factual discourse" we must distinguish "between the *primary* content of factual truth (truth as correct picture), which makes intelligible all the other modes of factual truth, and the *generic* concept of S-assertibility . . . " (SM 119).

I think that Sellars has adopted a cumbersome and confusing way of saying something that can be said much more clearly. We accept a descriptive conceptual system because we conclude that its built-in propositions are true; if it turns out that this was a mistake, we try again. In doing so, we reject the claim that some or all of the propositions assertable in the abandoned system are true. This is no more mysterious, and need be no more troubling, than cases in which we describe a physical object as red but then, after further examination, conclude that it is not red after all. I see no advantage to defining "true" in such a way that the contents of conceptual systems we no longer accept are true but less adequate than the contents of systems we currently accept. Rather, realists need hold only that one goal of science is to establish conceptual systems in which every proposition is true. If we think that the attempt to find fully adequate conceptual systems is a long-term goal in many domains, then we need concepts to distinguish what is acceptable given a well-supported conceptual system from what is not acceptable in that system. But we already have sufficient means for making this distinction. They are embodied in our concepts of justification, and in such related notions as having sufficient grounds for believing that a proposition is true, and working in a conceptual system that takes the truth of some propositions as established. At the same time, we can still make the intelligible – and often highly desirable – remark that a justified proposition may not be true. The conceptual resources for these thoughts are found in our system of epistemic concepts.

The variation in content with justification clarifies a sense in which there is an epistemic element involved in truth, while TC also holds that this is not the only element. These two features provide some insight into the continuing attractiveness of epistemic accounts of truth, and the continuing resistance to these accounts. A similar point holds for disquotational accounts. The disquotation view holds that the move from T1 to T2 is all there is to truth; rather than describing a property of a proposition, truth is a formal device that allows us to remove quotations marks. For TC this disquotational move is one – but only one – element in the content of TRUTH.

### *8.4.3 Implications*

When TRUTH is combined with the usual propositional connectives we find a set of implications that further underline its differences from justification.[24] Consider, first, the negation of T1:

> it is not case that *"p"* is true.

In bivalent logics this is equivalent to:

> it is the case that "not-*p*" is true.

There is no parallel equivalence in the case of justification. Noting that *p* is not justified does not imply that not-*p* is justified; there may not be enough evidence to establish justification in either case. This result extends directly to logics that admit multiple contraries to *p*: the denial of T1 implies that the disjunction of all admissible contraries is true. But given a set of contraries to justification, denying that *p* is justified does not imply anything about the justificational status of the disjunction of these contraries; it may still be indeterminate. Conjunction also operates differently in the cases of truth and justification, although details may vary with the operative account of justification. The difference is clear for any account that considers a proposition justified if it has some high probability of truth, as long as that probability is less than one. For any set of logically independent propositions, *p*, *q*, . . . , if each proposition in the set is true, then their conjunction is true. But no matter how high we set the probabilistic bar for justification, the multiplication rule for probability guarantees that there will be cases in which each proposition in the set is justified, but their conjunction is not justified.

Still, when we consider implications, we find an aspect of truth that is problematic for TC. According to TC there should be a body of implications connecting truth with other epistemic concepts; these are hard to find. That *p* is true does not imply that *p* is justified, or that there is any evidence that supports *p*, or that anyone believes or ought to believe *p*, or that *p* has any other particular epistemic status. There is, for example, no reason why Plato should have had any particular epistemic attitude to the claim that the sun is a fusion reactor; he was not even in a position to entertain this proposition. Nor do any of the usual epistemic evaluations imply truth. Rather, as I argued above, one systemic role of truth is to provide an ideal that stands beyond such evaluations and provides the conceptual space needed to recognize the intrinsic fallibility of the vast majority of these evaluations. Truth is, it seems, a somewhat special concept that does not fit comfortably into the framework of TC – although it is TC itself that highlights this point. Not all ideals function in this way. For example, there are cases in which consistency can be proved. One aim of Kant's ethics is to allow us to determine our

duty, which is considerably more elusive on teleological accounts. To be sure, knowledge implies truth, but this just indicates that knowledge is an ideal that is at least as elusive as truth. I leave this topic for further research although I want to recall a point from Ch. 1: TC is intended as a contribution to continuing research, not the final word on the subject. From this perspective it is, I urge, a virtue of TC that it allows us to see what is special about truth and indicates a direction for its own further development or eventual replacement.

## 8.5 Non-Propositional Knowledge

Many human cognitive achievements are not included in the realm of propositional knowledge. Although some may object, I will extend the use of "knowledge" to these cases. With a few notable exceptions (some examples are given below) there has been little epistemological attention to this subject. I want to indicate some reasons why this topic deserves greater attention, and why the concept of non-propositional knowledge should play a prominent role in an adequate system of epistemic concepts (cf. Brown 1988, 1994b).

Physical and cognitive skills are clear examples of human achievements. Physical skills are apparent in sports, where people develop varying abilities to serve at tennis, catch fly balls, and shoot baskets. More important examples are provided by carpenters, machinists, musicians, pilots, and surgeons – as well as by anyone who rides a bicycle or drives a car. These abilities play a major role in characteristically human accomplishments, and skill improvement is a major component of human epistemic development. Skills are non-propositional because people often exercise a skill without being able to describe how they do it. There are even cases in which people who have a skill will, when pressed, give a demonstrably incorrect account of what they do when exercising that skill. A classic example is riding a bike (cf. Polanyi 1958: 49–50): many competent riders claim that they keep their balance by shifting their weight in the direction opposite to that in which they are tipping. Yet this account fails to explain why it is so hard to keep one's balance on a stationary bike or on one that is moving very slowly. This example illustrates another point: often a propositional account is not particularly helpful in mastering a skill. One can give a detailed account of the physics of bike riding, but still not be able to ride without practice. This need to learn by practice is one characteristic feature of skills.

I am not claiming that skills are undescribable. We are often able to work out the details of a skill and embody it in a machine that will be faster and more accurate than human practitioners. But our ability to do this is independent of the point that human practioners exercise these skills without such descriptions. Often skills are developed in this non-propositional form before the descriptive eye is turned on them and a propositional account provided. In addition, while machines may embody considerable speed and

precision, they do not (currently) exhibit the flexibility and adaptability we often find in human practitioners.

Computer programming will serve to introduce the notion of a cognitive skill: the outcome of a programmer's work is an explicit and detailed set of steps for carrying out a procedure, yet we do not have comparable detailed accounts of how to write a program. Rather, programmers are trained in a way that is analogous to the way athletes, drivers, and machinists, are trained. We give students examples, allow them to carry out tasks under controlled circumstances, gradually increase the range of circumstances and level of difficulty of their tasks, and count on the ability of most people to improve with practice – and sometimes reach levels of accomplishment well beyond that of their teachers. This is the same process by which we teach students to construct proofs in formal logic and mathematics, play chess, write well (not just grammatically), translate between languages, compose music, and carry out many other human endeavors. When we wish to have a program written or a piece of music composed, we depend on those who have exhibited the ability to carry out these tasks – and this is eminently sensible behavior even though the practitioners we depend on cannot tell us how they do it. To be sure, we also formulate guidelines and maxims, but as Kuhn, Polanyi, and others have pointed out, students do not learn skills by learning these maxims. Indeed, learning the skill is often a prerequisite for understanding the maxim. There are also cases in which there is no sharp distinction between physical and cognitive skills. These include the abilities of painters, sculptors, and musical performers – especially those who improvise – as well as those of laboratory scientists, surgeons, airline pilots, and more. Our overall body of epistemic accomplishments would be significantly diminished without these skills.

While epistemologists mainly focus their attention on propositional knowledge, there are some important exceptions, such as Ryle's (1949) discussion of *knowing-how* and *knowing-that*, Polanyi's (1958) account of *tacit knowledge* and his thesis that knowing is an art, Kuhn's (1962) account of how normal scientists learn the current paradigm, and Putnam's (1978) discussion of linguistic skills. Kuhn's original reasons for adopting the term "paradigm" include the claim that normal science is learned by practice – especially learning how to model solutions of new problems on previously successful problem solutions – and that the skills developed through this process are more fundamental for the pursuit of normal science than any rules that can be formulated. Unfortunately, this theme was largely lost in Kuhn's later writings and in much of the debate generated by his work.

It is currently far from clear how to integrate non-propositional knowledge into our system of epistemic concepts, although it is important that we do so. Major advances in our ability to function in the world – including our ability to improve our stock of propositional knowledge – depend on skills. Still, it is not clear how, if at all, justification and truth apply in this case. It is, for example, common practice to treat truth as a property of

propositions – as I did in the above discussion; but we still need to work out the relation between true propositions and skills. There are also wider issues. Our analytical and critical abilities seem especially apt when we have a propositional formulation to work with – although skills can be criticized and improved, and new skills can be developed, all without a propositional formulation. Some have argued that our propositional knowledge is itself dependent on skills. Putnam, for example, has suggested that our ability to use language depends on a variety of skills that cannot all be expressed linguistically, while for Polanyi this is just one instance of the general point that we know more than we can say.

Given our greater current understanding of propositional knowledge, it is tempting to try to reduce mastery of skills to learning propositions. Since we are not able to formulate the required propositions in many cases, it is tempting to postulate unconscious knowledge of these propositions. Yet this proposal is an explanatory hypothesis and must be evaluated as such. Elsewhere (1988: 172–73) I have examined the related case of postulating unconscious rules to account for skilful behavior that does not follow explicit rules, and argued that it is not a promising explanatory strategy. My main concern here is to stress that these are areas in which research is needed and, as is generally the case, conceptual development is an integral part of such research. It is most unlikely that we will advance our understanding in these cases by peering more carefully into concepts that are already available.

## 8.6 Social Epistemology

There is another central aspect of human knowledge that has received little attention from epistemologists until quite recently: human knowledge is deeply social (cf. Goldman 1999; Hooker 1987; Hull 1988; Kitcher 1993; Longino 1990; Solomon 1994, 2001). The body of human knowledge – both propositional and non-propositional – is distributed across humanity, with each individual mastering only a very small portion. Historically, epistemologists have treated knowledge as a purely individual phenomenon and recent studies of the social side of knowledge have often been built on an individualistic foundation. Goldman has been especially prominent in this regard, with his 1999 book focusing mainly on such problems as how an individual can evaluate testimony and the reliability of experts. There is some discussion of social means by which our overall epistemic state can be improved, but this too is mainly aimed at increasing the reliability and scope of individual beliefs (cf. Brown 2000b). Analysis of the currently dominant concept of knowledge may support this focus on individual knowledge, but then we have another case that calls for conceptual revision. In this section I want only to indicate some of the reasons for this claim and some of the issues that should be addressed in the further development of our epistemic framework.

Let us note the enormous range in which each of us depends on the epistemic accomplishments of other people – accomplishments we are not

able to check for ourselves. Cases in which we accept testimony from those who witnessed an event we were not in a position to witness, and reports of that testimony, are familiar, but barely hint at the range of situations in which we depend on the epistemic accomplishments of others. We rely on others' knowledge whenever we use a textbook or handbook. Scientists consult handbooks to find out properties of chemical compounds and subatomic particles; physicians consult such books to check the appropriate medication for an illness, its side-effects, and its interactions with other drugs; structural engineers use handbooks that list the properties of commonly available steel sections; and the list goes on. In these cases we seek a specific piece of information; in other cases we rely on the overall information and skills of others. We do this whenever we trust our safety to airplane designers, pilots, mechanics, air-traffic controllers, and those who designed and programmed the computers that all of these now depend on. We do the same when we rely on a team of, say, surgeon, anesthetist, laboratory technician, blood supplier, and drug producer. Less dramatically, we do the same when we buy a computer or a piece of software, a measuring instrument, or a refrigerator. Among physicists there is now a fairly sharp division between theoreticians and experimenters; in other fields we find people with high-level manipulative skills, or exceptional mathematical or linguistic abilities. In some scientific fields experimental research requires large teams whose members bring different kinds of expertise to the project. The case of the top quark is an extreme example: one paper announcing its verification (Abachi, *et al*. 1995) had 500 authors, and also depended on the work of myriad technicians and other contributors. Ensuring the quality of such projects is not best pursued by having every (or any) member of the team personally check every computation, every piece of data, every line of computer code, and every electric circuit. Rather, the problem is one of organizing team members to put their particular skills to use in appropriate ways. A being who grasped the principles behind all of the accomplishments mentioned and personally tested every application would be in a superior epistemic state than we are, but this kind of epistemic power has no relevance to human knowers; it is not even an ideal to which we can sanely aspire. The extension and improvement of *human* knowledge is not best pursued by attempting to create individuals who master all abilities. Rather, it is a matter of developing social structures that maintain, organize, improve, and assure the reliability of the myriad contributions on which each of us depends (cf. Hooker 1987; Hull 1988). Hooker refers to these as *epistemic institutions* (1987: 313–15, 1995: *passim*), a concept that we may want to integrate into our epistemic system. It is clear that people who do highly cooperative research have figured out how to make it work a good deal of the time; there is little they can learn by consulting epistemologists, and much that those interested in a theoretical understanding of human knowledge can learn from the work of those engaged in such endeavors.

Another aspect of epistemic social organization concerns the desirability of diverse approaches to a problem (cf., Goldman 1999: 254–60; Kitcher 1993: 68–72, *et passim*). At a given stage some hypotheses or methods may seem more attractive than others, but these do not always include the correct approach. It is thus important to have a structure that encourages some researchers to follow less popular routes. To a degree such options are kept alive by variations among researchers. Some will make different plausibility judgments than others; some will prefer to back the long shot or to buck the crowd; and so on. A social structure in which minority views are encouraged and sustained is desirable just from an epistemic perspective – aside from other values that may be involved. But the problems of knowing how to implement such structures are complex, especially since we are dealing with distribution of limited resources. One cannot support every proposal, or give everyone time on the Hubble telescope or the latest supercomputer. Nor can one publish every paper that someone writes. Those who attempt to keep up with research in even a small number of contemporary specialties depend on the work of editors and referees in providing some filtering of what gets printed. But editorial judgments are unavoidably fallible, and an epistemic community needs mechanisms for reconsidering prior decisions and correcting errors.

The same applies to the educational institutions that free us from the need to rediscover and reinvent everything that our predecessors discovered and invented. These institutions pass on those earlier accomplishment that are deemed worthwhile, but this is another endeavor that involves fallible judgments about what is worth retaining – judgments that sometimes need to be revised. Recall that there have been periods in which Hume's philosophy and Mozart's music were given little attention. In science too theories that are rejected at one stage sometimes make a comeback (see, for example, Cushing 1994; Polanyi 1969). Making and revising fallible judgments – including judgments about what is worth preserving and transmitting – is an unavoidable part of the process by which we pursue and improve our knowledge; epistemic theories should address these matters. Although some work is being done along these lines, epistemologists have barely begun to tickle these issues.

## 8.7 Conclusion: The Status of Conceptual Analysis

I now want to bring together several themes concerning the nature, basis, and purpose of conceptual analysis. I will begin with two questions that overlap: the appropriate "data" for carrying out a conceptual analysis, and the reasons for thinking that these data yield results that have some significant claim to universality. As we have seen, conceptual analysts largely rely on their own intuitions, which are supposed to be generated by concepts they already possess. But it is fair to ask what evidence we have for thinking that I associate the same concept with a word as others do. The question would seem to be empirical. Jackson addresses this question in a defense of concep-

tual analysis. Responding to the suggestion that philosophers should use opinion polls to determine the general understanding of specific concepts he writes:

> My answer is that I do – when it is necessary. Everyone who presents the Gettier cases to a class of students is doing their own bit of fieldwork, and we all know the answer they get in the vast majority of cases. But it is also true that often we know that our own case is typical and so can generalize from it to others. It was surely not a surprise to Gettier that so many people agreed about his cases.
>
> (1998: 37)

Note several points about this reply. First, a professor's assessment of whether students in his course share his reactions is hardly an example of proper polling methodology. Second, in the Gettier case Jackson claims agreement in the "vast majority" of responses. What about the outliers? Are they unimportant? In science it has often been arcane resistant anomalies that provided the basis for major theory change (finches from the Galapagos, black-body radiation, a minute discrepancy in Mercury's orbit). Third, a week before the publication of Gettier's paper, most analytic philosophers agreed on the JTB account of knowledge, and were thus (by their own lights) uniformly wrong about their concept of knowledge. This raises a serious question about the significance of such agreement. Fourth, even granting wide agreement in the Gettier case, incessant repetition of a favorable example is a paradigm of bad methodology. There is major disagreement on such concepts as causation, knowledge, justification, and truth; the list would only expand if we moved on to moral, metaphysical, and aesthetic concepts. Fifth, why in the world would anyone assume that a professional philosopher's intuitions are typical of humanity in general? As Goldman has noted, agreement among philosophers can be explained without invoking wide-ranging conceptual uniformity:

> Philosophers sometimes seem to assume great uniformity in epistemic judgments. The assumption may stem from the fact that it is mostly the judgments of philosophers themselves that have been reported, and they are members of a fairly homogeneous subculture. A wider "pool" of subjects might reveal a much lower degree of uniformity.
>
> (1992: 160, cf. 143–44)

I want to explore this last point somewhat further. Recall Goldman's remark that "we" would not accept a Papal declaration to be sufficient for justification. Given Goldman's claim (in that paper) that he is analyzing an existing concept, he appears to be making a factual claim. It is fair, then, to ask who is included in "we." Code makes this point in discussing Foley's (1987) account of epistemic rationality.

> Richard Foley appeals repeatedly to the epistemic judgments of people who are "like the rest of us" (p. 108). He contrasts their beliefs with beliefs that seem "crazy or bizarre or outlandish . . . beliefs to most of the rest of us" (p. 114), and argues that an account of rational belief is plausible only if it can be presented from "some nonweird perspective" (p. 140). Foley contends that "an individual has to be at least minimally like us in order for charges of irrationality even to make sense" (p. 240). Nowhere does he take up the question of who "we" are.
>
> (1991: 8, n. 7, references and ellipses are all in Code's text; cf. 301–3)

Code has her own conjecture about who tacitly counts as a paradigmatic member of "we": "an adult (but not *old*), white, reasonably affluent (latterly middle-class) educated man of status, property, and publicly acceptable accomplishments" (1991: 7). By way of contrast, many people, both historically and in large parts of the contemporary world, consider those who deny the existence of sorcerers, ghosts, and papal infallibility to be holding bizarre beliefs.

While many analysts make universal claims on the basis of intuitions derived from overly narrow sources, there are also cases in which the range of people whose intuitions count should be quite narrow. The history of science presents us with many concepts that are not available to all people at all times, and that have a good claim for being accurate descriptions of items in their domain. Many of these concepts compete with concepts that are found in various social groups. We encountered several such examples in Ch. 2; we will encounter others in Chs 9 and 10. There is no good reason why we should have any interest in the everyday versions of these concepts – except, of course, for purposes of historical and anthropological study. One could analyze the everyday concepts of space and time, but given the development of relativity it would be a mistake to conclude that these concepts have any special status in a description of the world; the same applies to simultaneity, heredity, force, and many more.

Consider another style of argument for the universality of "our" concepts; it is exemplified by Rescher's response to the challenge that other societies might have a different concept of rationality from ours. I would like to quote and comment on a dozen pages of Rescher's text (1988: 144–56), but a briefer discussion will have to serve. Rescher maintains that,

> it is literally nonsense to say 'The *X*'s have a different conception of rationality from the one we have'. For, if they do not have ours, they do not have any. Whatever analogue or functional equivalent there may be with which they are working, it is just not something that we, in our language, can call 'a conception of rationality'.
>
> (152)

Rescher suggests that this strong claim is actually based on a triviality:

> What is universal about rationality is not something profound about sociology, but something rather trivial about language use; namely, that to accredit another culture as rational at all is to accept it as being rational in *our* sense of the term – which may, to be sure, involve deciding whether their actions live up to *their* standards. The absoluteness of (ideal) rationality is inherent in the very concept at issue.
>
> (150)

There is, I suggest, an odd ambivalence here in at least two respects. One of these concerns whether it is *important* to be rational. To describe someone as having failed to be rational is usually taken as a criticism, but we can still ask why being rational is important. We can bring the issue into clearer focus by looking at a familiar metaphor that Rescher uses: he writes at times of behavior within a culture as a game (150, *et passim*). Treating rule-governed activities as games is illuminating to a degree, but (like the treatment of a theory as a language) leads us astray if pushed too hard or taken too literally. Consider an actual game such as baseball. This game is played in some cultures, but not in others. Those who play baseball get to specify its rules, and if other cultures do not play, so be it. Our understanding of baseball involves a number of concepts that are not found in other cultures, or in other games in our own culture. The failure to possess or behave in accordance with such concepts as home run, or passed ball is not a significant failing. The absence of the infield-fly rule in basketball and Buddhism is not the basis for a critique of those practices. Presumably the failure to be rational is a failing, and the absence of this concept in a culture is another defect. This suggests that more is involved in the concept of rationality than just a set of rules we put together. Before considering what else is involved, consider the second point of ambivalence mentioned above.

Rescher sometimes writes as if we are locked into our current concepts. For example, "The fact that *we* do (and must) apply *our* own idea of the matter is what makes for the universal element of rationality"(150, cf. 145, 147, 149, 153). He also holds that we can change standards, but that once we do this we must take the new standards as absolutely correct (e.g., 145–46). As a result, "You might *force* me to change standards. Or you can, perhaps, brainwash me. But cannot *rationally persuade* me" (149). The ambivalence comes out when we consider Rescher's advocacy of a "chastened relativism."

> We realize (relativistically) that pluralism prevails – that other standards are used by others. But we can (and must) nevertheless accept (absolutistically) our own standards as *appropriate* for ourselves. To recognize a standard as rationally valid is – where rational agents are at issue – already to have adopted it as one's own. We take a cognitive position

> when we adopt a set of standards of truth and validity, but in doing so we assume an evaluative position. But such a position is by its very nature incompatible with the prospect of accepting alternatives, because the holding of a particular evaluative position *consists in* rejecting the alternatives. Even when conceding the prospect of someone's having another position, we cannot see it as available to ourselves.
>
> (148)

Thus Rescher recognizes that we have considerable ability to understand other standards, while insisting that we are, somehow, locked into our own standards. This locking need not be permanent, although at best we can jump from one locked room to another.

Our explorations in the history of conceptual change indicate that Rescher's view of our cognitive abilities is overly limited. Even the history of institutionalized sports is replete with reconsiderations of the rules in the pursuit of a game that better achieves some set of ends (cf. Gould 1996). Moreover, TC provides insight into the dialectic between our own current understanding of rationality and alternative possibilities. We can explore this theme without digressing into a lengthy account of yet another contested concept.

The crux of the matter lies in the systemic role of rationality. Part of this role is to capture a way of pursuing our ends that makes effective use of our cognitive resources. Rescher would surely agree with this remark, but for TC this is only part of the concept. The importance of the multi-dimensionality of our concepts appears once again in the way it allows us to maintain an anchor in existing concepts while exploring ways in which these concepts may be improved. In the present case, means of improvement may include finding better ways of pursuing our own ends. For example, this might occur if we found that people in other cultures have discovered ways of pursuing these ends that are more effective – by our own lights – than those we currently adopt. Were this to occur, we might have good reasons for adopting their procedures and the associated concept – which might turn out to be a variant on our concept of rationality. Moreover, this change might occur in stages. We might, for example, discover implicational patterns or ways of attaching concepts to their domains that differ from ours while maintaining the same goals. Once we incorporate these modifications into our ways of thinking, we might find reasons for modifying our goals, and so on. Note especially that whether we can do this is a question about our cognitive abilities, and that Rescher's own metalinguistic discussions – such as his discussions of pluralism and the possibility of adopting alternative concepts – already recognize all the cognitive resources we would need. In addition, on the approach I am suggesting we can give better reasons for adopting our current concepts than that they are our concepts.

I have been stressing the limits of conceptual analysis, so I will end this chapter by noting some of its uses. Under the guidance of an appropriate theory of concepts, analysis is the key to a better understanding of many

concepts that we currently hold, as well as the limitations of those concepts. Analysis is also required if we are to understand the concepts of other cultures, and of earlier historical periods of our own culture. In particular, historical studies of science (and other fields) require analyses of the concepts we are exploring at their various stages. In the next two chapters I will use TC as the basis for examining conceptual systems and conceptual change in seventeenth century mechanics and twentieth century high-energy physics. These chapters should be viewed as both further applications of TC and as tests of the theory. One more possible application of TC is worth mentioning, although I will not pursue it here. The theory offers an account of ways in which existing conceptual systems may be altered. It is possible that a clearer understanding of these possibilities can contribute to the fruitful pursuit of conceptual change.

# 9 Historical Studies I: Seventeenth-Century Physics

> Scientific investigation, says Popper, *starts* with a problem, and proceeds by *solving* it. This characterization does not consider that problems may be wrongly formulated, that one may inquire about properties of things and processes which later views declare to be non-existent.
>
> (Feyerabend 1975: 274)

In this chapter I use TC to study some major conceptual developments in the physics of Galileo, Descartes, and Newton. Since Galileo and Descartes both sought to replace Aristotelian physics, I begin with a more detailed account of this theory than I gave in Sec. 2.1. I next examine the central concepts of Galilean and Cartesian physics, and compare these to the main Aristotelian concepts. Then I consider Newtonian physics, which was largely developed in opposition to Descartes. I use TC both to illuminate the conceptual structure of each theory and to study the relations between successive systems of physical concepts.

The theories I examine in this chapter have not been discussed in detail during my presentation and initial defense of TC. Thus I hope to extract a double dividend from these studies: to advance our understanding of these theories, and to provide further reasons for taking TC as a basis for such studies. I will be examining central concepts and major conceptual changes that can be documented in the writings of these physicists, but I make no attempt to formulate the complete conceptual framework of any of these individuals, or to study all of the changes that took place in a single scientist's career. For example, I will not discuss Galileo's work on strength of materials, or Descartes' biology and psychology, or Newton's optics and alchemy. In addition, I will largely ignore contributions by Beekman, Huygens, Kepler, and Leibniz, among others. If anyone finds TC to be of value and applies it to additional figures and issues I will be delighted.

## 9.1 Aristotle

We saw in Sec. 2.1 that Aristotelians divide the natural world into two realms: the terrestrial, which encompasses everything from the center of the

universe (where the earth is located as a matter of physical necessity) to the sphere of the moon; and the celestial which includes the moon, sun, planets, and stars.

Terrestrial space is organized into four *natural places*, each associated with one of the four terrestrial *elements*: the center of the universe for earth and the sphere of the moon for fire, with air below fire and water between air and earth. An unconstrained sample of an element is either located at its natural place or moves there spontaneously; this is *natural motion*, and it ends once the natural place is reached. Natural motion is linear, can be either upward or downward, depending on the element involved, and is not eternal; rather, natural motion brings about its own elimination. Any motion that is not natural is *violent*. Each of these terms is associated with a concept; consider their central implications.

On a generic level there is a complete set of mutual implications between ELEMENT, NATURAL MOTION, and NATURAL PLACE, while each specific element-concept (EARTH, WATER, AIR, FIRE) is tied by mutual implications to the concept of a specific natural motion and a specific natural place. In addition, the four elements divide into two pairs: air and fire are light – that is, their natural motions are upward; earth and water heavy – their natural motions are downward. LIGHT and HEAVY are theoretical concepts that help organize our thinking about the composition of the terrestrial world and the motions that occur there. These concepts bring along additional implications. For example, EARTH implies HEAVY, which implies DOWNWARD NATURAL MOTION, and so forth.

Natural and violent motions are defined as contraries; working in the Aristotelian framework, the presence of one of these in an object implies the absence of the other. It is, then, a *conceptual truth* that natural and violent motion – say, horizontal and vertical motion – cannot exist simultaneously in a single object. Violent motion also implies the presence of a sustaining force that is external to the object being moved. Violent motion exists only as long as this force acts, and the object moves only in the direction determined by that force. For example, a typical projectile (e.g., an arrow shot at some target) is an earthy object, its natural motion is downward; motion in any other direction requires a sustaining force. Once the force ceases to act, natural motion takes over and the arrow falls straight down.[1] Consider another earthy object: a stone resting on a table. Since the stone's natural place is at the center of the universe some force must be restraining it. Obviously this force is provided by the table, and when the support is removed the stone immediately moves downward in a straight line. In a similar way, since the natural place of fire is at the sphere of the moon, an unrestrained flame will move upward in a straight line. We see this when we encounter a fire in a closed building: once the roof is breached the fire moves as expected. Note especially that on a generic level the motion of the unrestrained stone and the unrestrained fire are instances of the same kind of motion.

Every non-vertical motion is violent. But while natural motion implies vertical motion, the converse does not hold – a stone thrown upward engages in violent motion. Thus we must know what element we are dealing with to determine whether a vertical motion is natural or violent. Aristotle further recognizes that an object may be moving towards its natural place while an external force is also pushing it toward that place. Still, the dichotomy between natural and violent motion is fundamental; both cannot occur simultaneously. Aristotle holds that in this case the force accelerates the motion, but it is still natural motion: "since movement is always due either to nature or to constraint, movement which is natural, as downward movement is to a stone, will be merely accelerated by an external force, while an unnatural movement will be due to the force alone" (1995c, 301b17-301b30: 494). Thus the presence of a force does not imply violent motion. Presumably a retarding force will slow a natural motion, but as long as the force does not stop the motion or change its direction the motion remains natural. However, violent motion implies the presence of a force, and this implication provides a central GA: As long as one is working within this framework, failure to find such a force is a failure of the researcher. Finding the force that sustains projectile motion was the main Aristotelian research problem in the terrestrial realm. It is, for example, far from clear what force sustains the motion of an arrow after it leaves the bow; proponents of Aristotelian physics attempted to find that force.

Now consider the ICs for natural and violent motion. Since this theory operates at the level of everyday perception, these ICs will give the means by which we detect these motions using our senses. In the terrestrial realm the actual path of a moving object is exactly what it appears to be, so that any non-vertical motion is immediately identifiable as violent. For the reasons given in the previous paragraph, determination of whether a vertical motion is natural or violent requires additional information about the element involved. Identification of instances of an element are also mostly non-problematic. Aristotle relies on general cultural knowledge for such identification – with the qualification that familiar physical objects are mixtures of elements, so that our common examples of the elements are actually cases in which we pick out a dominant element. While cases may arise in which the dominant element is unclear, the theory provides a decision procedure for resolving such cases: remove any restraints or other outside forces, and see how the object moves and how far it goes.

Turning to systemic roles, natural and violent motion have both descriptive and explanatory roles. Their descriptive roles are clear enough, and once we have classified a motion in one of these two ways, we have already indicated what kinds of explanations for the motion are relevant. There is only one kind of explanation for natural motion: the object is moving towards its natural place:

> how can we account for the motion of light things and heavy things to their proper places? The reason for it is that they have a natural tendency

> towards a certain position; and this is what it is to be light or heavy, the former being determined by an upward, the latter by a downward, tendency.
>
> (1995d, 255a24-b31: 426)

Violent motions are also amenable to only one kind of explanation – some sustaining force must account for the motion. However, the exact nature of this force may be far from clear – as the long-standing puzzle of projectile motion among Aristotelians demonstrates. So identification of a motion as violent gives only an outline of an explanation with details to be provided. MOTION and REST constitute another fundamental dichotomy with an explanatory function. Rest at an element's natural place is the natural state in the terrestrial realm. All motion tends towards rest: natural motion ends when a object reaches its natural place or is somehow constrained; violent motion ends when the sustaining force is removed, at which point natural motion ensues. If an object is resting at its natural place no further explanation is required. Rest at any other place requires an explanation that invokes some restraint.

As we saw in Sec. 2.1, Aristotelian chemistry integrates smoothly with terrestrial physics. Aristotelian chemistry introduces additional concepts for the four fundamental qualities. These divide into two sets of contraries: hot/cold, and wet/dry. Each element is characterized by a pair of these qualities, one from each set, so that each element-concept implies two quality-concepts, and a pair of quality-concepts implies an element-concept. However, the qualities are more fundamental than the elements since elements change into each other in a systematic way as one fundamental quality changes at a time. The qualities are not subject to such changes.

In the celestial realm the story is rather different. Here there is only ether, which does not occur on earth. Indeed, ETHER implies CIRCULAR MOTION and this implication embodies the central GA of astronomy: All celestial motion is circular. The motions of the planets seem to be departures from circular motion, but these are only *apparent* departures to be explained by finding an underlying set of uniform circular motions that account for the apparent departures.[2] This is quite different from the terrestrial situation where departures from natural motion (including circular motion) are what they seem to be and are to be explained by finding the force that sustains them. There are no such forces in the celestial realm. Moreover, on earth it would be a serious misconception to think of non-circular motions (whether violent or natural) as departures from circular motion.

Note that I have not invoked natural motion or elements in describing the celestial realm. We can introduce such concepts into astronomy, but they are rather different from the concepts associated with these labels in terrestrial physics. In the heavens there is only ether and only circular motion. We can call ether an "element" and circular motion "natural," but there is no contrast between different elements or kinds of natural motion, no natural

places, and no violent motions. As a result, these concepts lack the substantive implications of their terrestrial counterparts. Describing a celestial motion as natural or a celestial object as elemental does not tell us anything we did not already know. In addition, the only IC we need for instances of elements or natural motion in the heavens is the fact that they occur in the heavens. In the terrestrial case ELEMENT and NATURAL MOTION do real work. Moreover, since the heavens do not contain distinct elements, there is no reason to consider the chemistry of the heavens. This result is consistent with the Aristotelian thesis that there is no generation or corruption in the heavens.

We could, of course, introduce a more general framework that integrates the two systems. This framework would include more abstract versions of the concepts of natural motion and an element: we can think of natural motion as any motion that takes place without an external sustaining force, and we can think of the universe as composed of five elements, one celestial and four terrestrial. There is no harm in doing this as long as we are clear that the abstraction involves a considerable loss of content. For example, the more abstract concept of an element does not imply natural place, and the more abstract concept of natural motion implies only that the motion is either vertical or circular, depending on where it is located. As a result, to understand this framework and apply it for detailed descriptive and explanatory purposes we must still distinguish celestial-natural-motion, which is circular and eternal, from terrestrial-natural-motion, which is linear and self-limiting, and so on. I suggest that we gain more insight into the these concepts by thinking of the celestial and terrestrial frameworks as different conceptual systems that overlap to a small degree. This approach underlines how sharply different the two frameworks are, and how large a conceptual gap had to be overcome by seventeenth-century scientists who sought to break down the distinction between the two realms and develop a single unified physics for both the heavens and the earth.

## 9.2 Galileo

Copernican astronomy challenges Aristotle's two-part universe by, so to speak, putting the earth into the heavens. But the conjunction of Copernican astronomy and Aristotelian physics – the only system of physics available at the time – implies phenomena that are contradicted by observation. One case derives from the daily rotation of the earth. Suppose I drop a rock from the top of a tower. Since the rock falls straight down towards the center of the earth, while the tower rotates from west to east, the rock will land somewhere west of the tower. In fact the rock lands at the foot of the tower, thus the earth cannot be rotating.[3] This is one instance of a large cluster of empirical arguments that Aristotelians used to show that the earth does not move. One of Galileo's aims in his *Dialogue Concerning the Two Chief World Systems* (1967, henceforth *Dialogue* – all Galileo references are to this book unless

otherwise indicated) was to show that these arguments fail because of the way they conjoin Copernican astronomy with Aristotelian physics. Copernican astronomy requires a different physics, which Galileo endeavored to develop. When Copernican astronomy is conjoined with Galileo's physics we arrive at the correct predictions, so these Aristotelian arguments are irrelevant. I will focus here on Galileo's new terrestrial physics.

A key feature of Galileo's approach is his rejection of the thesis that circular motion is natural only in the heavens. Instead Galileo maintains that the earth has two natural circular motions – the daily rotation and annual revolution – and that a stone dropped from the top of a tower falls at the base of the tower because it shares these natural motions. In the case of the daily motion Galileo writes:

> But the diurnal motion is being taken as the terrestrial globe's own and natural motion, and hence that of all its parts, as a thing indelibly impressed on them by nature. Therefore the rock at the top of a tower has as its primary tendency a revolution about the center of the whole in twenty-four hours, and it eternally exercises this natural propensity no matter where it is placed. To be convinced of this, you have only to alter an outmoded impression made upon your mind, saying, "Having thought until now that it is a property of the earth's globe to remain motionless with respect to its center, I have never had any difficulty in or resistance to understanding that each of its particles also rests naturally in the same quiescence. Just so, it ought to be that if the natural tendency of the earth were to go around its center in twenty-four hours, each of its particles would also have an inherent and natural inclination not to stand still but to follow in the same course."
>
> (142, see also 134)

Analogous points apply to the earth's annual motion. Taking these natural motions into account we can explain why the stone falls at the foot of the tower even as the earth moves. Since these are natural motions, no force is required to sustain them; both of these natural motions exist simultaneously in the planet, as well as in the falling stone. Thus on Galileo's account the falling stone is simultaneously engaged in *three distinct motions*: the two natural circular motions, plus the motion of fall; I will postpone Galileo's account of the last of these for a bit.

Galileo also holds that the stone will sustain a motion that is impressed on it; this motion is also circular. Consider a ship moving on the sea. Force is required to maintain its motion against friction, but if we imagine the friction being reduced, less force will be required until, when friction is completely eliminated, the ship will continue moving (Galileo maintains) around the earth at a constant distance from the center forever (145–48). By way of contrast, any motion that involves an increase in an object's distance from the center of the earth is vertical motion. In particular, motion tangent

to the earth involves a continual increase in distance from the center of the earth and could not take place without the application of a force (193–95). The thesis that the stone will sustain an impressed motion is central to Galileo's account of an experiment that could test this part of his new physics against the Aristotelian view. Suppose a sailor drops a stone from the top of the mast of a moving ship, where will it land with respect to that mast? For Aristotelians, once the sailor lets go of the stone it engages in a single motion, straight down to the center of the universe, while the ship continues moving; thus the stone will fall toward the rear of the ship. According to Galileo, as the stone falls it maintains the ship's motion – which was impressed on it before it was dropped. Thus Galileo predicts that the stone will land at the foot of the mast. This case is particularly important because it yields a testable prediction that differs from the Aristotelian conclusion. I want to explore this case in some detail.

In *Dialogue* the experiment is proposed by Simplicio, the Aristotelian spokesman, as a response to Galileo's account of vertical fall. Initially Salviati, Galileo's spokesman, does not challenge the Aristotelian view that the rock will fall towards the rear of the ship (141–42); but he does emphasize that this case is different from the tower since the ship example concerns accidental, rather than natural, motion.

> There is a considerable difference between the matter of the ship and that of the earth under the assumption that the diurnal motion belongs to the terrestrial globe. For it is quite obvious that just as the motion of the ship is not its natural one, so the motion of all things in it is accidental; hence it is no wonder that this stone which was held at the top of the mast falls down when it is set free, without any compulsion to follow the motion of the ship.
>
> (141–42)

I will consider shortly why Galileo accepts the Aristotelian account at this point; for the moment it is more important to note that he soon retracts this concession (144). According to his own account (leaving aside the motion of the earth), the stone is simultaneously engaged in two distinct motions: fall plus the motion of the ship that has been impressed on it. Once dropped, the stone continues in the latter motion as long as there is "no cause for diminution in the property impressed upon it" (149). For Aristotle these two motions are contraries, but Galileo challenges this claim. These motions "are not contraries, nor are they destructive of one another, nor incompatible" (149) because the two motions derive from different causes that do not interfere with each other: "heaviness attends only to the drawing of the movable body toward the center, and impressed force only to its being led around the center, so no occasion remains for any impediment" (149). As de Gandt (1995: 205) notes in discussing the related case of parabolic projectile motion, the claim that different motions can be compounded in a moving

body without destroying each other is a substantive innovation; for Aristotle such joint motion is *conceptually* impossible. One of Galileo's key steps on the route to this innovation is his rejection of the conceptualization of moving objects in terms of the dichotomy between natural and violent motion; this opens up logical space for compound motions. Given Galileo's two natural motions, the falling stone is actually engaged in *four distinct motions*. Galileo also uses this ability to sustain an impressed motion to account for the continuation of projectile motion once the original source of motion has been removed.

Galileo's account implies that the stone lands at the foot of the mast, and we can now see why he did not initially challenge Simplicio's claim that the stone falls towards the rear of the ship. Galileo's immediate goal was to distinguish between fall on a moving ship and fall from a tower before giving his own account of the situation on the ship. When Galileo wrote *Dialogue* the experiment had not yet been done (145). Suppose it is done and the stone falls at the foot of the mast; since this is mere accidental motion, the same conclusion will hold *a fortiori* for the stone's *natural* diurnal motion. This is particularly clear when Galileo later draws an explicit comparison between the ship and the tower (after providing a general discussion of projectile motion). At that stage in the argument Galileo is emphasizing a key consequence of his account of the ship case: observation of the fall of the stone provides no information about whether the ship is moving or stationary. He then poses a challenge to Simplicio:

> Now if in this example no difference whatever appears, what is it that you claim to see in the stone falling from the top of the tower, where the rotational movement is not adventitious and accidental to the stone, but natural and eternal . . . ?
>
> (154)

In other words, if it is impossible to tell whether the ship is moving or stationary by observing the fall of a stone from the mast, it is surely impossible to tell whether the earth is moving by observing the fall of a stone from a tower. Later, when he is discussing an anti-Copernican argument aimed at the earth's annual motion (a cannon ball is shot vertically and returns to its starting point), Galileo again underlines the role of natural motion: "Keeping up with the earth is the primordial and eternal motion ineradicable and inseparably participated in by this ball as a terrestrial object, which it has by its nature and will possess forever" (177–78).

But what if the ship experiment is done and the Aristotelian prediction is confirmed? The way Galileo has set up his argument, this would pose a serious problem for his account of projectile motion, but would not count as an argument against the motion of the earth since the motion of the stone relative to the ship is mere accidental motion. Even if accidental motion does allow us to tell whether the ship is moving, it does not follow that the

same holds for natural motion. I suggest that Galileo did not immediately challenge Simplicio's account of what would happen on the ship in order to set up this feature of his argument.

Now consider Galileo's account of falling objects. It seems to me that Galileo is somewhat unsure on exactly how to think about fall. On his account the actual path of a falling object is not vertical since it includes the two natural motions and perhaps an impressed motion as well. Moreover, there are many passages in which Galileo rejects any role for straight-line motion in dynamics. He denies that straight-line motion ever occurs in an ordered universe (19, 31) and suggests that the only role for straight-line motion is to restore order that has been disrupted (242–43). He maintains that natural motion must be eternal, so that motion in a straight line does not qualify as natural (31–32, 134–36). He even suggests that straight-line motion may not exist at all: when we see such motion, we are actually seeing circular motion from a limited point of view. Thus he considers the possibility that the actual path of a stone falling from the top of a tower is an arc of a semicircle with one end at the top of the tower and the other end at the center of the earth (162–67). He then draws three conclusions from this account (166–67): that only circular motion occurs in this case; that the distance the stone moves in falling to the earth is the same as it would traverse if it stayed at the top of the tower; and that the actual motion is never accelerated. Galileo presents this thesis as only probable, but he does repeat it (264). If this account seems weird to a twentieth-century mind, this can serve as an indicator of how long the journey is from Aristotelian physical concepts to post-Newtonian concepts.

Other passages suggest that Galileo gives up on providing any account of why objects fall. At one point Salviati indicates that he does not know what makes objects fall. To Simplicio's claim that it is gravity Salviati replies:

> You are wrong, Simplicio; what you ought to say is that everyone knows that it is called "gravity." What I am asking you for is not the name of the thing, but its essence, of which essence you know not a bit more than you know about the essence of whatever moves the stars around.[4]
>
> (234)

In a frequently-cited passage from *Two New Sciences* (1974: 158–59; henceforth TNS) Galileo insists that he will investigate some features of accelerated motion, but will not inquire into its causes. The important point for us is that whatever causes fall, and whatever the correct description of the path of a falling body, falling objects are engaged in at least three simultaneous motions.

Now let us consider more carefully what Galileo means by "terrestrial objects." In the midst of his discussion of the natural motion of terrestrial objects Galileo makes a possibly surprising remark about air:

> at least that part of the air which is lower than the highest mountains must be swept along and carried along by the roughness of the earth's surface, or must naturally follow the diurnal motion because of being a mixture of various terrestrial vapors and exhalations.
>
> (142)

The air, like the stone, moves along with the earth in its daily (and annual) motion, but *for a different reason*. Air does not share the earth's natural motion, thus a specific cause must be introduced to explain why air follows the motion of the earth. The context of this passage suggests that we should be wary since it occurs while Galileo is working under the temporary assumption that the Aristotelian account of a stone falling on a ship is correct. But Galileo returns to this topic on the "Fourth Day" of *Dialogue* where it provides the basis of an argument for the motion of the earth. In that later discussion Galileo first tells us that the air does not follow the earth as a result of impressed motion: "The air, being a thing that is in itself very tenuous and extremely light, is most easily movable by the slightest force; but it is also most inept at conserving the motion when the mover ceases acting" (438). He then adds:

> the air, as a tenuous and fluid body which is not solidly attached to the earth, seems to have no need of obeying the earth's motion, except insofar as the roughness of the terrestrial surface catches and carries along with it that part of the air which is contiguous to it, or does not exceed by any great distance the greatest altitude of the mountains. This portion of the air ought to be the least resistant to the earth's rotation, being filled with vapors, fumes, and exhalations, which are materials that participate in the earthy properties and are consequently naturally adapted to these same movements.
>
> (439)

The air close to the earth – the air that we experience – is either carried around by the roughness of the earth, or follows the earth because it is mixed with terrestrial vapors that share the earth's natural motions.

This view of the air close to the earth would have been familiar to many of Galileo's readers; it goes back at least as far as Aristotle. In his *Meteorology* Aristotle writes:

> So at the centre and round it we get earth and water, the heaviest and coldest elements, by themselves; round them and contiguous with them, air and what we commonly call fire . . . but in reality, of what we call air, the part surrounding the earth is moist and warm, because it contains both vapour and a dry exhalation from the earth.
>
> (1995a, 340b: 558)

And, a bit later:

> When the sun warms the earth the exhalation which takes place is necessarily of two kinds, not of one only as some think. One kind is rather of the nature of vapour, the other of the nature of a windy exhalation. That which rises from the moisture contained in the earth and on its surface is vapour, while that rising from the earth itself, which is dry, is like smoke.
>
> (1995a, 341b: 559)

A passage from Descartes' *The World* indicates that this view of the air near the earth was still held, perhaps in a stronger form: "The Philosophers maintain that above the clouds there is a kind of air much subtler than ours, which is not composed of terrestrial vapors, as our air is, but constitutes an element in itself" (1998: 16–17).

Aristotelian arguments against the motion of the earth include arguments from the behavior of the air. Aristotelians argued that a daily rotation of the earth would generate a persistent wind, and the annual motion of the earth should cause us to lose our atmosphere. Galileo must respond to these arguments, but he rejects the two approaches that are available from his account of the motion of stones: air does not share the natural motions of the earth, nor does it sustain impressed motions. Rather, air follows the earth because the air is carried along by the roughness of the earth and by earthy vapors that are mixed with the air near the surface of the earth. Moreover, Galileo attempts to turn the Aristotelian argument against the diurnal motion of the earth into an argument for this motion. First he acknowledges the point of the Aristotelian argument:

> But where the cause for motion is lacking – that is, where the earth's surface has large flat spaces and where there would be less admixture of earthy vapors – the reason for the surrounding air to obey entirely the seizure of the terrestrial rotation would be partly removed. Hence, while the earth is revolving toward the east, a beating wind blowing from east to west ought to be continually felt in such places, and this blowing should be most perceptible where the earth whirls most rapidly; this would be in places most distant from the poles and closest to the great circle of diurnal rotation.
>
> (439)

Then he claims that such winds exist:

> Now the fact is that actual experience strongly confirms the philosophical argument. For within the Torrid Zone (that is, between the tropics), in the open seas, at those parts of them remote from land, just where earthy vapors are absent, a perpetual breeze is felt moving from the east

> with so constant a tenor that, thanks to this, ships prosper in their voyages to the West Indies.
>
> (439)

Additional examples of such winds follow in Galileo's text.

Note how Galileo's view of the nature of air provides a key element in his account of why these winds exist over water but not over land. Galileo treats earth and air as distinct elements. As in Aristotelian physics, these elements are characterized by their dynamical properties, although these properties are quite different than the Aristotelian properties. For Galileo, earth has a pair of eternal natural motions and conserves impressed motions;[5] air does not have these natural motions and has only a minimal ability to sustain impressed motion. Galileo doubts that fire is an element (443), but water fits right between earth and air in this scheme, and the properties of water play a central role in his theory of the tides, which he considered his most important argument for the motion of the earth.[6] Let us follow this argument.

At the beginning of the discussion Galileo points out that water is "not joined and linked with the terrestrial globe as are all its solid parts, but is rather, because of its fluidity, free and separate and a law unto itself . . . " (417). It is because of this that "among all sublunary things it is only in the element of water . . . that we may recognize some trace or indication of the earth's behavior in regard to motion and rest" (416–17).[7] Two features of this element provide the basis for Galileo's claim. One of these, that water does not share the natural motions of the earth, is implicit in the above passage: If water shared these natural motions it would be no more able than a falling stone to reveal the earth's motion. The second feature is that water conserves an impressed motion to a considerable degree, although it takes some time for water to acquire a new motion. Galileo supports this claim by noting that when a water-carrying barge slows down the water moves forward; when the barge speeds up the water moves towards the rear. Thus in a barge with a varying speed we find that:

> the water (being contained within the vessel but not firmly adhering to it as do its solid parts) would because of its fluidity be almost separate and free, and not compelled to follow all of the changes of its container. Thus the vessel being retarded, the water would retain a part of the impetus already received, so that it would run toward the forward end, where it would necessarily rise. On the other hand, when the vessel was speeded up, the water would retain a part of its slowness and would fall somewhat behind while becoming accustomed to the new impetus, remaining toward the back end, where it would rise somewhat.
>
> (424)

The account is repeated in the next paragraph.

Tides, according to Galileo, occur because water trapped in a basin, such as the Mediterranean, is continually subjected to the double (annual and daily) motions of the earth. Each motion provides the water with an impetus to move in a specific circle. If the earth had only one motion the water would be carried along and no tides would occur. But at each moment a second impetus is also impressed on the water by the second motion of the earth. Thus the water moves in an irregular fashion, which generates the tides.[8]

Simplicio raises an objection to this account, and Galileo's response will bring out another important feature of his system of dynamical concepts. The objection is that each of the supposed motions of the earth is circular, and therefore regular, and that it is impossible to construct the irregular motion required by this theory out of regular motions (426). Now Galileo accepts the Aristotelian thesis that circular motion is the only uniform (that is, non-accelerated) motion.[9] Thus he agrees that the two natural motions of the earth are uniform, but argues that the Aristotelian conclusion does not follow: "From the composition of these two motions, each of them in itself uniform, I say that there results an uneven motion in the parts of the earth" (426). Galileo then introduces a diagram that astronomers would recognize as the standard epicycle-deferent diagram (see Figure 9.1), although in this case the larger circle represents the orbit of the earth while the smaller circle represents the rotating earth on that orbit. Galileo uses this diagram in a way that parallels the usage of Ptolemaic astronomers: to show how two uniform circular motions, properly combined, can yield an irregular motion.

A long tradition holds that Galileo's tidal theory is inconsistent with his mechanics, but these arguments ignore the role of the elements in Galileo's mechanics and overly modernize the views attributed to him. While the justification for the interpretation I have proposed stands or falls on the basis of

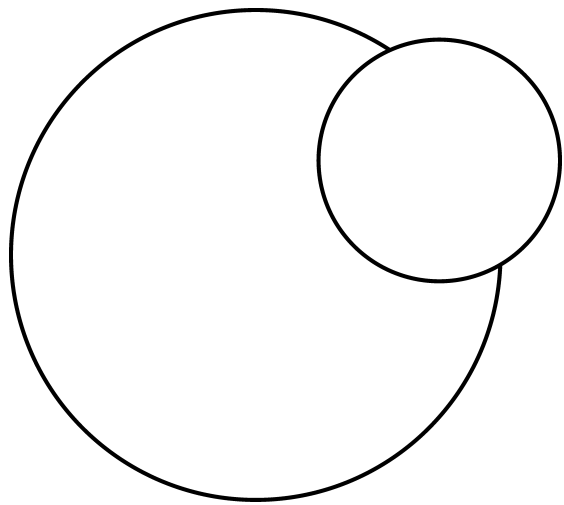

*Figure 9.1* Galileo's Circles

the texts, it is worth noting that my account restores the consistency of Galileo's argument, even though it places him further from the views of later physics than do some readings. Galileo's doctrine of elements is central to his physics; I want to sum up that doctrine. There are three elements. *Earth* is characterized by two natural circular motions and by the conservation of impressed motions. Once a non-natural circular motion has been impressed on an earthy object, the object continues with that motion until something interferes. *Water* does not share the natural motions of earth, but it does conserve impressed motions to a high degree and exhibits a significant resistance to the acquisition of such motion. *Air* does not share the natural motions of the earth and has a much lower ability than the other elements to sustain impressed motion, and much lower resistance to acquiring a motion in a new direction. I have not found any passages in which Galileo discusses resistance of earthy objects to a new impressed motion, although I suspect that Galileo would attribute this property to earth.[10] To the extent that this doctrine of elements plays a central role in Galileo's mechanics, he is building his new mechanics on a modified version of the mechanics that was generally accepted in his day. Galileo's new mechanics does not appear *ex nihilo*, but the inclusion of these traditional features does not prevent him from proposing a genuinely new theory. I want to explore some further features of this new mechanics.

Galileo's account of the stone falling on the ship leads to an account of projectile motion built on the idea that once an earthy object has been pushed it retains its motion unless it is impeded (149–56). This contradicts the Aristotelian view that projectile motion requires an external force to keep the projectile in motion. Recall that finding this force was a central Aristotelian research problem. Some Aristotelians held that air provides the required force but Galileo rejects this view, argues that air impedes this motion (e.g., 135, 153), and criticizes the thesis that air could sustain projectile motion (150–53). Indeed, the difficulties with attributing this role to air were long familiar and in the middle ages an alternative account, known as the *impetus theory*, was proposed.[11] According to this view the force that sets a projectile in motion is imparted to that projectile and this force – the impetus – keeps the projectile moving. Adopting this view raises the question why a projectile stops moving (short of hitting some obstacle). The obvious step is to argue that the impetus is dissipated, and there were two competing views as to why this occurs: Some held that impetus is conserved unless it must work against an opposing force – such as that provided by the air; others held that impetus wears out by itself (Clavelin 1974: 93). The impetus theory must be distinguished from the later inertial view that we will find (in very different versions) both in Descartes and Newton. On an inertial view continued motion of a projectile does not require a sustaining force. Inertia is not a cause of continued motion; rather "inertial motion" is used to label motion that continues without the need of a sustaining force. The impetus theory is a modification of the Aristotelian

framework and is firmly embedded in that framework. Impetus theorists accepted the division of all terrestrial motions into natural and violent, and proposed an account of violent motion in which the causes of the continuation of such motion include the impressed impetus. Introduction of an internal cause of sustained motion is the one key departure from the traditional Aristotelian view.

Galileo adopted a version of the impetus theory in his earliest work (1960: 76–85), explaining the way impetus is imparted to a projectile in terms of two analogies: an object that is heated remains hot even after the source of heat is removed, and a bell struck by a hammer continues to ring after the hammer has been removed. He held that in these analogous cases the "conserved" property diminishes over time on its own, and attributed the same property to impetus. It is more controversial just what view he took in his last books. Many passages in *Dialogue* suggest that he still held an impetus theory (e.g., 22–23, 151–52, 216), but some commentators maintain that he developed at least a first approximation to an inertial account of projectile motion – where inertial motion is circular.[12] Drake, for example, denies that Galileo viewed impetus as a cause, and cites a remark from *Dialogue* in which Galileo is discussing a thrown ball: "What is it that stays with the ball, except that motion received from your arm?" (Drake 1978: 476, n. 14). But Galileo's full sentence reads, "When you throw it with your arm, what is it that stays with the ball when it has left your hand, except the motion received from your arm which is conserved in it and continues to urge it on?" (156). As McMullin notes, "Metaphors of this sort abound in the text of the *Dialogo*, calling into question the frequently-made claim that Galileo's *impeto* is cut off by an ontological abyss from the *impetus* of the Paris theorists" (1967: 17). McMullin concludes: "There is undoubtedly a tension in the *Dialogo* between the metaphor of the *impeto* that causally explains the continuance of projectile motion and an argument-structure that suggests that the *continuance*, as such, of the motion needs no explanation" (28). For current purposes we need not attempt to settle the matter. It is sufficient to note that a crucial shift in thinking about projectile motion was taking place from the view that continued motion requires a sustaining force to the view that it does not. We will see that this shift is clear in Descartes, but that there is still a long way to go from Descartes to Newton.

Occasional passages in *Dialogue* read as if projectiles move in a straight line (e.g., the discussions of centrifugal force at 191–93 and 216). In TNS Galileo analyzes projectile motion as compounded from a combination of a constant-speed linear horizontal motion and an accelerated vertical motion, but it is clear that he considers the horizontal part of the motion to be actually circular. For example, at one point Salviati tells us that by "horizontal" he means a surface that remains equidistant from the center (TNS 172). Later, when he is discussing objections to his account of projectile motion, Galileo has Simplicio remind us of this proper sense of "horizontal." Salviati replies that he uses the linear path as an approximation, and that his practice

is in accord with that of Archimedes who treats the arms of a balance in the same way, and who also treats hanging weighted cords as parallel.[13] The approximation is justified because "the distances we employ are so small in comparison with the great distance to the center of our terrestrial globe . . ." (TNS 223–24). Galileo adds that "if such minutiae had to be taken into account in practical operations, we should have to commence by reprehending architects, who imagine that with plumb-lines they erect the highest towers in parallel lines" (224). Drake notes that Galileo consistently uses a straight-line approximation for small circular arcs from his earliest writings:

> In *De Motu*, where the approach was purely theoretical, the fact that the earth's surface is not literally horizontal demanded notice; in the *Mechanics* it was an accidental circumstance. Thus for the purposes of practical mechanics Galileo regarded inertial motion as horizontal, whereas for the purposes of theoretical analysis any inertial motion was necessarily maintained equidistant at all times from the center toward which the unsupported body would naturally move. This continued to be his practice later on, giving rise to modern debates over the illusory question which treatment Galileo himself regarded as correct for every possible purpose.
>
> (1978: 60)

Still, we must not read Galileo's appeal to practical considerations too narrowly. The discussion in TNS includes the practical considerations involved in actually testing theoretical results. This question brings us to a central methodological theme in Galileo and his successors.

Galileo plays a key role in the new project of constructing a mathematical physics – a project that goes directly against the Aristotelian view of the relation between mathematics and physics. Simplicio raises this issue in *Dialogue* when he argues that mathematics deals in subtleties that "do very well in the abstract, but . . . do not work out when applied to sensible and physical matters" (203). For example, while mathematical spheres and planes touch in only one point, this does not hold for physical spheres and planes. Salviati replies that the correct conclusion to be drawn from this observation is that the physical objects in question are not spheres and planes (203–7). The mathematical physicist analyzes precisely defined objects that only approximate physical objects, and must account for the differences:

> Just as the computer who wants his calculations to deal with sugar, silk, and wool must discount the boxes, bales, and other packings, so the mathematical scientist (*filosofo geometra*), when he wants to recognize in the concrete the effects he has proved in the abstract, must deduct the material hindrances, and if he is able to do so, I assure you that things are in no less agreement than arithmetical computations.
>
> (207)

Yet this approach raises two fundamental questions: What justifies us in believing that mathematical results apply to physical objects, and how can we test such claims? Galileo deals with these issues in greatest detail in TNS. After a demonstration of the law of fall, Simplicio asks for some experiment to show that this reasoning applies to actual cases. Salviati immediately agrees that this is the right question and proceeds to describe an experiment – although this experiment includes a number of special arrangements that must be justified (169–70).[14] Later in TNS, after Galileo has developed his account of the parabolic motion of a projectile, the issue is raised again, and several questions are presented about the relation between abstract mathematical analysis and actual cases of projectile motion. These include objections from the use of a linear approximation for "real" horizontal motion and the effects of air resistance. Galileo agrees that these are all genuine problems and discusses how to construct experiments that minimize their effects (222–29). We have already considered his reply in the case of the linear approximation, and further details need not concern us. What is important for us is Galileo's clear recognition that the introduction of mathematics into physics involves the introduction of approximations so that we should not expect the phenomena we actually observe to conform exactly to the results of mathematical analysis. Additional work is required to confirm mathematical accounts and to determine their limits. This theme will recur throughout this chapter.

We have seen that Galileo, like Aristotle, has a doctrine of elements that is intimately tied to a doctrine of natural motions. I want to use TC to examine Galileo's versions of these concepts. Doing so will help clarify both the internal structure of Galileo's own physics and its conceptual relations with Aristotelian terrestrial physics. It will also provide the basis for comparisons of Galilean concepts with those of later physicists.

The *systemic roles* of these concepts are very similar in the two frameworks. In both cases the doctrine of elements serves to distinguish fundamental types of physical bodies, where each type is characterized by its dynamical behavior. A concept of natural motion also occurs in both frameworks where it plays a role in characterizing the elements and also distinguishes motions that require explanation in terms of a specific force from those that do not require such explanation. These functional similarities provide an important bridge from the Aristotelian framework to the Galilean.

Differences in detail become dominant when we look at the *implications* among these concepts, although we also find some overlaps. Recall that for Aristotle each element has only one natural motion so there is a mutual implication between an element and its natural motion. All natural motions on earth are vertical, linear, self-limiting, and distinguished by their endpoints; circular terrestrial motion is violent motion. Since natural and violent motions are contraries, the presence of one of these types of motions implies the absence of the other. To my knowledge Aristotle does not discuss

whether two violent motions can exist simultaneously in an object. If not, then any moving object must have only one kind of motion at a time.

For Galileo these implications are completely disrupted. He makes no use of VIOLENT MOTION. He holds that NATURAL MOTION implies ETERNAL MOTION, and thus must be circular, although not all circular motions are natural – as is illustrated by the motion of rock in a sling, or a wheel. The element earth has two natural motions – so distinct motions can exist simultaneously in an object. A falling object has a third motion, whose nature is not exactly clear. A projectile will have a fourth motion as well: an impressed motion that the projectile retains. Galileo recognizes only three elements – earth, water, and air. Water and air do not have the natural circular motions of earth. Indeed, WATER and AIR imply an absence of natural motions. The ability to conserve an impressed motion is now included in the defining characteristics of the elements: the concepts of earth and water both imply this ability, and water implies a considerable resistance to acquiring a new impressed motion. The concept of air implies a greatly reduced ability to sustain impressed motion, along with only minimal resistance to a change of motion.

The ICs for instances of EARTH, WATER, and AIR are straightforward in both frameworks since they rely on cultural background knowledge and yield the same instances for both Galileo and Aristotle. Identification of falling objects involves both similarities and differences. Galileo and Aristotle agree that any object that appears to be moving vertically downward is falling, but Galileo identifies projectiles as falling objects while Aristotle denies this. There is also wide agreement on the identification of projectiles, although for Galileo a rock dropped from the mast of a moving ship is a projectile (150) – as is any object that is dropped while on a vehicle moving relative to the earth. None of these are projectiles in Aristotle's view. Galileo and Aristotle also disagree on how we determine the actual path of a moving object. For Aristotle this is simple since the actual path is exactly what it appears to be. Moreover, objects that appear to be at rest are actually at rest. For Galileo the issue is much more complex. Given that we do not see the natural motions of earthy objects, none of these actually have the path they appear to have; in all such cases we must add in these natural motions. Thus there is a theoretical element in the ICs for paths of earthy objects, whether they appear to be moving or stationary. We have seen that Galileo gives only a probable account of the actual path of a stone falling from a tower. The situation is even more complex for a stone falling on a moving ship. Whenever we share a non-natural motion with an object, there is an additional motion that we do not see; this must be included in an account of the actual path. Galileo does not offer an account of these paths.

The water that constitutes the seas is continually engaged in a complex motion generated by two impressed motions; this complex motion is its actual path since there are no additional natural motions. The motions of air over land (ignoring the weather) are just what they appear to be.

Differences between the motions of air and water result from the difference in their ability to conserve impressed motions. Air over long stretches of water is also moving just as it appears to be relative to the earth – although in fact it is the earth, not the air, that is moving. Neither Aristotle nor Galileo considers all cases we might introduce. I noted above that Aristotle does not discuss the possibility of two violent motions existing simultaneously in an object; Galileo does not consider the case of falling water.

Galileo's conceptual innovations are far-reaching and highly original, but he still works under the general limitations of human thought. He cannot begin from nowhere and create new concepts out of whole cloth – and if he could, no one would understand him. Instead, he works from concepts he inherited and introduces new ideas by way of modifications of those concepts. But this limitation does not prevent radical innovations. Still, the full development of a viable new physics required more time and more cognitive resources than any one individual could command; I will examine the contributions of two more scientists who pursued this project.

## 9.3 Descartes

Descartes continues the project of unifying the universe and attempting to capture its nature in mathematical terms. Descartes developed his physics over the course of his life: in *The World*, which was not published in his lifetime, the *Discourse on Method* and the three treatises (*Optics*, *Geometry*, and *Meteorology*) that provide its substance, and the *Principles of Philosophy* (henceforth PP). I will base my discussion primarily on PP although I consider earlier works when they help clarify issues. There are significant differences between the original Latin text of PP and the later French translation. Since it is generally agreed that Descartes collaborated on and approved the French version, I will draw on the French unless otherwise noted.[15] Descartes wrote PP in opposition to Aristotelian philosophy and science, paying little attention to Galileo.

In PP Descartes hedges on the crucial issue of the earth's motion. According to his own account of motion (discussed below) the earth does not move even though it is carried around the sun and rotates daily (III 28, 29; IV 22). Still, Descartes advocates several views that are characteristic of the new astronomy. He holds that sun is a star, and thus that there are multiple suns (III 13, cf., 9). He rejects Ptolemaic astronomy on empirical grounds, in particular the phases of Venus (III 16). He holds that the Copernican and Tychonic systems are equivalent when treated solely as hypotheses, but that according to the correct account of motion Tycho attributes more motion to the earth than does Copernicus (III 17–19).[16] He lists many ways in which the earth is like the planets (III 8, 11, 27) and even explicitly includes the earth among the planets (III 13). He argues that the fixed stars are not all located on a single sphere (III 23) and are much farther away than the planets (III 7, 20, 40, 41). Descartes' account of the cosmos

suggests that there are planets around other stars (III 54, 115, 199). He holds that comets move through the heavens, including the heavens beyond Saturn (III 41) and the heavens surrounding other stars (III 119, 126, 127). Most importantly for our purposes, Descartes holds that there is, on a fundamental level, only one kind of matter and thus one physical theory that applies throughout the universe (II 22–23). I will examine his account of matter first, and then consider his account of motion.

The unity of matter and project of mathematization go together in Descartes' doctrine that matter has a single essential property – extension. Note three features of this doctrine. First, it paves the way for holding that geometry is the key to the study of the material world. Second, it follows that the universe is a plenum; a vacuum cannot exist since there is no space without matter (II 16). Third, since all matter has the same essence, whether in the heavens or on the earth, all matter is subject to the same fundamental laws. Yet this does not eliminate *all* differences among forms of matter. Descartes maintains that there are three kinds of matter – three elements (III 52) – although they differ only in the sizes and shapes of their particles, with no sharp boundaries between the different types of matter. The three elements contribute to the filling of space in different ways.

Descartes develops his account of matter by constructing a hypothetical story of how the world could have reached its present state. He assumes that initially the world consisted of a huge number of particles of roughly equal size and varying shapes (III 46). Each particle moves around its own center, and groups of particles move around other centers.[17] These initial particles cannot all be spherical since no set of spheres can completely fill a volume of space (III 48), but they become spherical by continually banging against each other. The edges and corners that are knocked off through these impacts (Descartes calls them "scrapings") constitute the first element. These are the smallest particles in the universe, move rapidly, easily change shape, and fill any gaps (III 49–51). The middle-sized spherical particles that result from these interactions constitute Descartes' second element. In addition, under certain conditions particles of the second element clump together to form larger particles, which constitute the third element. The entire visible universe is made up of these three elements: "the Sun and the fixed Stars of the first, the Heavens of the second, and the Earth, the Planets, and the Comets of the third" (III 52: 110). The apparently empty space between astronomical bodies is actually filled with particles of the second element interlaced with particles of the first.

One role of the three elements becomes particularly clear when we look at Descartes' theory of light. The full title of his first attempt at a comprehensive account of the universe is *The World or A Treatise on Light*; in the *Discourse on Method* Descartes explains why he took light as his central theme. Given the richness of the universe, some unifying perspective must be found if we are to proceed in a coherent way; the study of light will lead us

to consider all major features of the universe (2001: 34–35). There are three main concerns for a theory of light: its production, its transmission through space, and its reflection and refraction by physical objects; each of these is based on one of the elements. There are just two original sources of light, stars and fire, both consisting of the first element. Light itself, Descartes holds, is a kind of pressure – a tendency to movement – that results in the perception of light when it strikes an eye (III 62–64, 2001: 66–70). This pressure is transmitted by the second element. Objects made of the third element intercept light and either reflect it or transmit and refract it (III 52).

In *The World* the three elements become fire, air, and earth – the only elements Descartes admits in this book. Descartes denies any fundamental status for Aristotle's four basic qualities (hot, cold, moist, and dry), "which are themselves in need of explanation" (1998: 18), although he is open to the possibility that pure forms of the elements are associated with special places in the universe (1998: 19). He also holds that the elements change into each other. (See, for example, the 1631 letter to Villebresieu quoted in Gaukroger 1995: 226.) However, this doctrine of elements plays a very limited role in Descartes' account of the physical world. In *Meteorology* Descartes offers a rich, complex account of entities and phenomena in the terrestrial realm, and attempts to reduce all the phenomena he discusses to the behavior of particles of different sizes, shapes, and states of motion; the elements play no special role in this account. Descartes does introduce water, earth, and air towards the beginning of this book, but only because these are familiar, not because they have some special status. He tells us that "water, earth, air, and all other such bodies that surround us are composed of many small particles of various sizes and shapes . . . " (2001: 264), but that:

> I do not conceive the small particles of terrestrial bodies as atoms or indivisible particles; rather, judging them all to be made of the same material, I believe that each one could be redivided in an infinity of ways, and that they differ among themselves only as pebbles of many different shapes would differ, had they been cut from the same rock.
>
> (2001: 268)

There are two themes in this passage. One of these – that there are no indivisible atoms (cf., PP II 20) – does not concern us at the moment. The second theme does: Although we encounter bodies of various types, they are all ultimately composed of a single kind of matter.

In PP (III 52) Descartes introduces three elements that parallel those of *The World*, but does not associate them with any of the four traditional elements; he refers to them only as "the first element," "the second element," and "the third element." While these elements play a role in Descartes' cosmology, they all have the same essence (extension). The logic of Descartes' view suggests eliminating the notion of an element altogether.

While Descartes does not do this, elements do not play the foundational role in his mechanics that they play in the mechanics of Aristotle and Galileo. Instead, Descartes introduces a small number of laws of nature that apply equally to all matter. These laws provide the core of his mechanics. Thus I will leave the elements aside, and turn to Descartes' account of motion. In discussing this account I will introduce the main concepts and implications among them first; I will consider instantiation conditions and systemic role towards the end of the section. In all cases I will make comparisons with predecessors as we proceed.

Descartes holds that motion is a permanent feature of the physical world. At creation God placed matter in motion, and the *quantity of motion* in the world remains unchanged. Given the original set of particles, a fixed quantity of motion, and the laws of nature, the world as we know it develops necessarily (III 46–47). Conservation of the total quantity of motion in the world follows, Descartes holds, from the immutability of God, but we need not consider his argument for this claim. His view of the proper measure of motion is of greater importance for physics. If we can determine this measure in the case of particular objects, we need only add it up over all objects to arrive at the universal constant. Descartes also holds that quantity of motion is conserved in specific interactions between particles, and this claim will provide one main focus of our discussion. His concept of motion, along with his doctrines of the constancy and proper measure of motion, are best discussed in the context of his *laws of nature*. In PP Descartes states three such laws plus a set of seven subsidiary *rules of impact* that fill out the significance of the third law.[18] These provide ten GAs that carry much of the content of Descartes' physical concepts.

Descartes' first law states that, "each thing, provided that it is simple and undivided, always remains in the same state as far as is in its power, and never changes except by external causes" (II 37: 59). The crucial concept in this law is STATE. States are properties (Descartes calls them "modes") of physical objects that persist unless changed by the action of an external cause. It has long been recognized, Descartes notes, that objects do not change their shape or begin to move without an external cause; shape and rest serve as paradigm examples of states. Descartes' key claim is that *a particular kind of motion is also a state*. This claim is not completely new: in ancient and medieval astronomy the circular motion postulated in the heavens is eternal – indeed, in this case there are no forces that could change the motion. However, in his comments on the second law Descartes emphasizes that circular motion is not a state; I will explore the reasons for this shortly. In Aristotelian terrestrial physics there are no motions that would count as Cartesian states. In particular, Aristotle's natural motions would not count as states since they spontaneously cease when the moving object reaches its natural place.

Before specifying which kind of motion counts as a state, Descartes notes that the first law solves the traditional problem of projectile motion. Given a

state of motion, the fact that an object is moving provides all the explanation we need for its continued motion. This account generates another problem – why moving objects ever stop – but in Descartes' plenum universe this question receives an easy answer:

> For there is no other reason why things which have been thrown should continue to move for some time after they have left the hand which threw them except that, in accordance with the laws of nature, having once begun to move, they continue to do so until they are slowed down by encounter with other bodies. It is obvious, moreover, that they are always gradually slowed down, either by the air itself or by some other fluid bodies through which they are moving, and that, as a result, their movement cannot last for long.
>
> (II 38: 60)

A previously intractable problem has been solved while a new problem that this solution generates is solved at once.

These remarks on why projectiles stop suggest that the state in question is, at least, motion at a constant speed. The direction of motion is the subject of the *second law*: "each part of matter, considered individually, tends to continue its movement only along straight lines, and never along curved ones . . . " (II 39: 60). In the absence of external agents, objects continue to move in whatever straight line they are currently following; there is no privileged direction for this motion, and this law holds for all matter everywhere in the universe. Descartes explicitly includes both constant speed and direction in a later remark (after introducing the third law): a body

> which is at rest has some force to remain at rest, and consequently to resist everything which can change it; while a moving body has some force to continue its motion, i.e., to continue to move at the same speed and in the same direction.
>
> (II 43: 63)

I will refer to *straight-line motion at a constant speed*, as *uniform motion*. Any change of direction or speed constitutes a change of state. Note especially the parallel between a body remaining at rest and one that is moving uniformly; I will return shortly to the forces mentioned in this passage.

Things are, however, not as simple as they may seem, because bodies never actually move in straight lines in the Cartesian plenum. Since there are no voids, other bodies are always acting on any given body; a moving body must push other bodies out of its way, these bodies must push others, and so forth. Such motion can be sustained only if there are groups of bodies moving in closed paths (II 33). Descartes usually describes these paths as circular, but notes that they need not be exact circles and may even be "extremely irregular" (II 33: 56); Garber notes that only closed paths are required (1992a:

220, n. 31). Still, bodies do not spontaneously move in closed paths. The closed paths that constitute the actual motions of bodies require impacts from other bodies. Taken on its own, "every moving body, at any given moment in the course of its movement, is *inclined* [my italics] to continue that movement in some direction in a straight line, and never in a curved one" (II 39: 60–61). Thus the actual motion of physical objects is (roughly) circular, but the fundamental physical phenomenon is an inclination to move in a straight line, even if this inclination is never actualized:

> the observation of actual motions does not suffice for a rational description of the universe; it is necessary also to include virtual motions – tendencies to motion, "efforts." To give an account of the state of a natural body, it is necessary to say what it would do in the next instant if nothing impeded it. The contrary-to-fact conditional . . . is even more inevitable in the Cartesian context because motion is always impeded from the moment it commences. . . .
>
> (de Gandt 1995: 123)

In various texts Descartes writes of an object's inclination, tendency, and determination to move in a particular direction. Garber argues that there is an important distinction between tendency and determination:

> Determination is an aspect of the motion a body has. . . . But Descartes does not use the word 'tendency' in this way. For him a tendency is not a motion or an aspect of motion, but a property a body has by virtue of which it would move if it were unimpeded.
>
> (1992a: 219–20)

Garber thus considers it significant that the second law is stated in terms of tendencies. However, I am not convinced that Descartes adheres to this distinction. In explaining the third law Descartes shifts between "tendency" and "inclination" (he does not use "determination"), and holds that a stone whirled in a sling has a tendency to move radially away from the center – although it would not move in this direction if it were unimpeded.[19] We will see that a moving body may have multiple determinations at a given time, but only one motion. Thus I will follow the common practice of treating "tendency," "inclination," and "determination" as synonyms. DETERMINATION is a central concept in Descartes' theory of motion; I want to explore this concept before introducing the third law.

Immediately after stating the second law Descartes tells us that the determination to move in a straight line "is confirmed by experience . . . " (II 39: 61): If a stone that is whirled in a sling leaves the sling it moves off in a straight line tangent to the circle. But this appeal to empirical support is misleading: although the stone may appear to move in a straight line, Descartes has already argued that the actual path is circular – a point that he

repeats in his discussion of the second law: "as stated before, in any movement, a circle of matter which moves together is always in some way formed" (II 39: 60). In the same article Descartes offers a second bit of empirical evidence supporting the claim that objects whirled in a sling tend to move in a straight line: "our hand can even feel this while we are turning the stone in the sling, for it pulls and stretches the rope in an attempt to move away from our hand in a straight line" (II 39: 61, cf. III 58).[20] This is a different determination than the one we have been considering, but the two determinations are both tendencies of the stone "to move [directly] away from the center of the circle which it is describing" (II 39: 61, "directly" was added by MM).[21] It is a key feature of Cartesian physics that determinations in different directions can exist in a body simultaneously (III 57). This is particularly important in Descartes' theory of light, where he maintains that light is a determination to move, without any actual motion; the simultaneous existence of multiple determinations explains why a source can emit light in all directions at once (III 64, 2001: 69–70). Still, inclusion of this second determination raises a question: Why does the rock move in one direction rather than another when it leaves the sling? As far as I can tell, Descartes never answers this question.

The importance of determination is particularly clear in Descartes' accounts of reflection and refraction in *Optics*; it will be worthwhile to consider these in some detail. Descartes uses an idealized model to develop his account of *reflection*: Consider a ball hit by a racket, moving downward at an angle to the vertical. The ball hits a horizontal surface at B and rebounds to the right (see Figure 9.2).[22] The problem is to find the relation

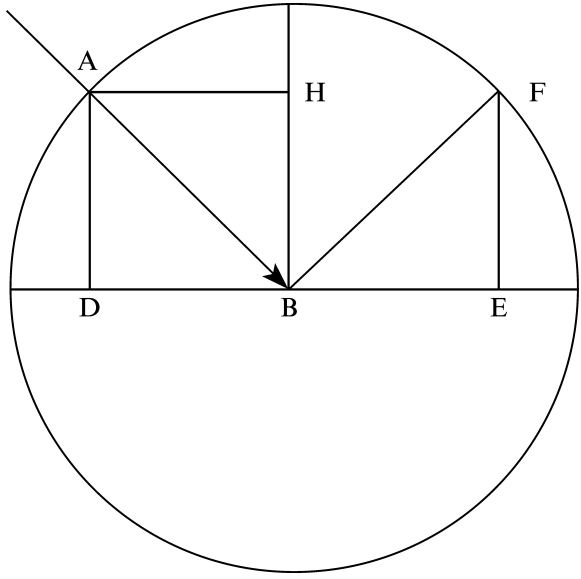


*Figure 9.2* Descartes' Reflection Diagram

between the angle of incidence ABH and the angle of reflection HBF. Descartes writes:

> But in order not to involve ourselves in new difficulties, let us assume that the ground is perfectly flat and hard, and that the ball always travels at a constant speed, as much in descending as in reascending, without asking ourselves in any way about the power which continues to move it after it is no longer touched by the racket, and without considering any effect of its weight, or of its bulk, or of its shape.
>
> (2001: 75)

Descartes now sharply distinguishes the cause of the ball's movement from the cause of its moving in a particular direction. The two causes are distinct because they can be varied independently of each other:

> the power, whatever it be, which causes the movement of this ball to continue is different from that which determines it to move in one direction rather than in another, as is quite easy to know from the fact that it is the force with which the racket has impelled it upon which its movement depends, and that this same force could have been able to make it move in any other direction as easily as towards *B*; whereas it is the position of this racket which determines it to tend toward *B*, and which could have determined it to tend there in the same way even though another force had moved it. Which already shows that it is not impossible that this ball be diverted by the encounter with the ground, and hence that the determination which it had to tend toward *B* be changed, without anything being changed by this in the force of its movement since these are two different things.
>
> (2001: 75–76)

In considering the cause of movement Descartes shifts between discussing the force applied by the racket, and whatever causes the ball's continued motion after it has been struck. I will consider what Descartes means by "force" (or "power") in more detail below. For the moment we can note that Descartes is concerned with the ball's continued motion, and that speed is the relevant measure of this motion (cf. Sabra 1981: 84). My current concern is the role that the distinction between the ball's speed and determination to move in a particular direction plays in the arguments that follow.

Descartes argues that although the speed acts only along the line AB, the determination can be decomposed in many different ways. In particular, it can be decomposed into a horizontal component and a vertical component.[23] When the ball hits the ground the vertical component of the determination is destroyed, but the horizontal component *and the speed* are not affected. Descartes emphasizes that the ball does not stop, even briefly, when it reaches B because "if its movement was once interrupted by this

stop, there could be found no cause which would make it start up again afterward" (2001: 76). Motion and rest are contraries, and a change from a state to its contrary always requires a cause. Thus if the moving ball once stops, an additional cause will be required to start it moving again (cf. Rohault 1969: 81).

The next step in the argument requires a premise that is not stated here, but which follows from Descartes' fourth rule of impact: the ball will rebound in the direction opposite to its original motion. In the present case this means that the downward vertical determination is replaced by an upward vertical determination. Since the speed has not changed, the ball will move a distance equal to AB in the same time.[24] Descartes implements this claim by drawing a circle with center B and radius AB:

> let us say that in as much time as the ball will take to move from *A* to *B*, it must infallibly return from *B* to a certain point on the circumference of this circle, inasmuch as all the points which are the same distance away from *B* as *A* is, are to be found on this circumference, and inasmuch as we assume the movement of this ball to be always of a constant speed.
>
> (2001: 76–77)

As the ball fell it moved to the right the distance DB; so it will move the same distance to the right from B in the same time. Thus Descartes constructs BE equal to DB, and a line at E perpendicular to the ground. The point F, at which EF meets the circle, is the unique point at which all required conditions are met. Geometry now yields the result that ABH equals HBF.

In the case of *refraction* Descartes uses three models, although the main argument is developed in terms of the first: A ball is hit by a racket at an angle, but the ground is replaced by a loosely woven cloth; the ball passes through the cloth and loses speed.[25] Descartes considers the case in which the ball loses half its speed:

> to know what path it must follow let us consider once more that its movement differs entirely from its determination to move in one direction rather than another, from which it follows that the quantity of these [two factors] must be examined separately.
>
> (2001: 77–78, "two factors" was added by the translator)

The determination before impact is again resolved into a horizontal and a vertical component, and the horizontal component remains unchanged as the ball passes through the cloth (Figure 9.3). Descartes constructs a circle centered at B with radius AB; after passing through the cloth the ball moves along the path BI. The problem is to locate I. As a result of the reduction in speed, the ball takes twice as long to traverse radius BI as it took to traverse

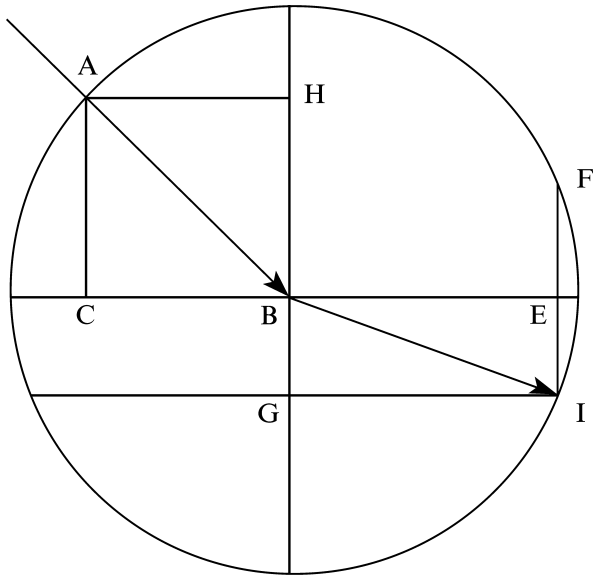


*Figure 9.3* Descartes Refraction Diagram

AB. Since the horizontal determination is not affected, BE is twice the length of CB. We construct the line FEI perpendicular to the cloth; the point I, at which FEI meets the circle, determines the direction BI. The angle of incidence is ABH; the angle of refraction is GBI. Descartes emphasizes that the effect of refraction must *not* be measured by comparing these angles, but rather by comparing the lengths CB and GI (which is equal to BE). Given this specification it follows that the problem was solved as soon as it was maintained that BE is twice CB, but there is a point of the rest of the construction. If we take the radius AB as our unit, we have the familiar law of refraction: the ratio of the sines of the angle of incidence and refraction is a constant determined by the materials through which the ball is moving.[26]

There are many problems with this argument, but I will consider only three that are relevant to our further discussion. First, the angles of reflection and refraction depend on the horizontal components of the determinations, and these are determined by ratios of the speeds (cf. Gabbey 1980: 252–54). Thus the determination to move in a particular direction is treated as a quantitative notion. But if determination depends on speed, it is no longer clear why Descartes sharply distinguishes the determination from the speed. Fermat, and other critics of Descartes' proofs, took determination to refer only to direction (Sabra 1981: 119–20), but this is not correct. Elsewhere Descartes makes it clear that determination involves both direction and speed (Sabra 1981: 120–21; Garber 1992a: 188–93). Rohault (1969, Ch. 13) regularly writes of "quantity of determination" as distinct from

"quantity of motion." In retrospect it seems that Descartes is on his way to the concept of velocity as a vector, but has not thought the issue through clearly. Discussing determination in the context of the rules of impact Garber concludes that "Descartes recognized the importance of both magnitude and direction, without knowing exactly how to combine them" (1992a: 246). Descartes does not consider the possibility that a concept combining speed and direction may play a more fundamental role in understanding motion than either speed or direction alone. It may well be that he was hampered by his clear understanding that the speed and direction of a motion can be independently varied. In any case, for Descartes, motion involves a determination to move in a straight line, so MOTION implies DETERMINATION which implies DIRECTION. MOTION also implies SPEED and perhaps DETERMINATION implies SPEED as well, although it is unclear whether we are dealing with the same speed in both cases, and how direction and speed relate.

Descartes' use of moving balls as a model for light raises a second problem since he holds that light is a determination to move, not an actual motion. Descartes is aware of the problem and has a response: "For it is very easy to believe that the action or the inclination to move which I have said must be taken for light, must follow in this [i.e., reflection and refraction] the same laws as does movement" (2001: 70). Yet Descartes provides no justification for this claim, and does not adopt it in other cases. For example, while Descartes holds that an object can have multiple determinations, he is clear that a body can have only one movement at a time (PP II 28). Thus while MOTION implies DETERMINATION, the converse does not hold.

The third problem arises from the role that change of speed plays in the account of refraction. In *The World* (Ch. 14) Descartes holds that the transmission of light is instantaneous, and in *Optics* he talks of transmission from one point to another in an instant (2001: 67, 69). Some commentators take this to be his consistent view (e.g., Gaukroger 2002: 148; Shea 1991: 235), but in PP Descartes seems a bit more cautious, saying only that light "is transmitted in the shortest space of time to the greatest distance" (III 64: 117). But even granted this more cautious view, it is difficult to see how, for Descartes, light can halve its speed by passing through the cloth.

Before considering the third law and the associated rules of impact, I want to note two issues that we should keep in mind. The first concerns the role of approximations in Descartes' physics. Unlike Galileo, Descartes does not discuss the use of approximations, but his theory of motion makes use of extreme approximations – as measured against his own view of the universe. For example, his first law of nature applies only to things that are simple and undivided, but Descartes holds that no such entities exist (e.g., II 20, 2001: 268). Similarly, the rules of impact are developed for cases in which two isolated bodies collide even though isolated bodies cannot exist in a plenum. We will consider as we proceed whether these idealizations generate significant problems.

Second, the rules of impact apply to bodies that are at rest or in motion with a specific speed and direction; these are all treated as absolute properties of bodies, not as relative to a frame of reference. Thus we must include MOTION and REST among the central concepts of Cartesian physics. Descartes provides an account of MOTION which may seem in accord with this approach: whether a body is moving (along with its speed and direction of motion) are to be assessed in comparison with contiguous bodies (II 24–30). A body that is stationary with respect to the contiguous bodies is at rest; a body that is moving with respect to those contiguous bodies is in motion with its speed and direction determined with respect to the contiguous bodies. These assessments of motion and rest hold even if the contiguous bodies are themselves in motion when judged (by the same criterion) in a wider context.[27] But Descartes also emphasizes that when we consider a body and its contiguous neighborhood, motion is reciprocal: each body moves with respect to the other (II 29). He explains why we typically consider one body to be at rest and the other in motion – for example, why we consider a person who is walking to be moving and the earth at rest – but adds:

> we must remember that all the real and positive properties which are in moving bodies, and by virtue of which we say that they move, are also found in those contiguous to them, even though we consider the second group to be at rest.
>
> (II 30: 54)

There seems to be a real conflict in PP between relative and absolute views of motion. Garber tries to rescue Descartes on the basis of a text that is not included in PP (1992b: 308–10). In this text Descartes says that the only genuine distinction between motion and rest is given by the mutual separation of two bodies, or the lack of such separation. Garber notes that if we apply this test to a given body and those immediately contiguous to it, whether they are separating is a determinate fact – and thus it is determinate whether the body of interest is in motion or at rest. This approach may save Descartes' treatment of motion and rest as distinct states, but we will have to consider whether it is sufficient to deal with all relevant cases once we have the rules of impact before us. We are now ready to consider the third law.

Descartes' third law is rather more complex than the first two:

> when a moving body meets another, if it has less force to continue to move in a straight line than the other has to resist it, it [loses its determination], retaining its quantity of motion. If, however, it has more force; it moves the other body with it, and loses as much of its motion as it gives to that other.[28]
>
> (II 40: 61)

This law has two distinct parts, for which Descartes offers distinct proofs (II 41, 42); I will not analyze these proofs. More importantly, for our purposes, the law makes use of three concepts that we will have to consider: QUANTITY OF MOTION, the FORCE OF MOTION TO CONTINUE IN A STRAIGH LINE, and the FORCE A BODY HAS TO RESIST MOTION.

Let us begin with QUANTITY OF MOTION. As noted above, Descartes holds that God introduced a permanently conserved quantity of motion into the universe at creation (II 36). In the third law Descartes takes it for granted that each collision between bodies also involves a conserved quantity of motion, and that this quantity is the sum of the quantities of motions of the various bodies involved in the collision (cf. Garber 1992a: 207). What determines the quantity of motion of a moving body? Several passages suggest that the quantity of motion is speed multiplied by size. For example,

> when one part of matter moves twice as fast as another twice as large, there is as much motion in the smaller as in the larger; and that whenever the movement of one part decreases, that of another increases exactly in proportion.
>
> (II 36: 58)

In discussing impact, we will see, Descartes consistently treats this product as a conserved quantity. Henceforth I will take quantity of motion to be the product of size and speed, and abbreviate it as QM; I will use CQM for the claim that this quantity is conserved. However, in interpreting QM it is important to avoid attributing later concepts, such as momentum, to Descartes since he does not have either the concept of mass or velocity (cf. Garber 1992b: 313–14).

Consider velocity first. In later mechanics velocity is a vector which combines speed and direction, but Descartes treats speed and direction differently. This distinction is central to his proof of the first part of the third law (which echoes an argument we encountered in *Optics*):

> The first part of this law is proved by the fact that there is a difference between motion considered in itself, and its determination in some direction; this difference makes it possible for the determination to be changed while the quantity of motion remains intact.
>
> (II 41: 62)

Although determination seems to include both direction and speed, in his account of impact Descartes regularly uses "determination" as a synonym for "direction."

Now consider mass, a concept introduced by Newton. Mass is a measure of the amount of matter in a body, and is an intrinsic property of that body; two bodies of the same volume can have quite different masses, depending

on how densely the matter in those bodies is packed. But this view is incompatible with Descartes' claim that extension is the only intrinsic property of body. Descartes discusses the relation between quantity of matter and volume in his account of condensation and rarefaction (II 5–7). On this account, when a body is rarified its shape changes generating gaps that are immediately filled with other matter. When a body becomes denser, some of this additional matter is moved away; when no additional matter is left, the body achieves its maximum density. A given volume, however, always contains exactly the same quantity of matter. This suggests that the proper measure of the size of a physical object is just its volume (II 5, 19), and thus that *QM is volume multiplied by speed*.

The conceptual gap between Descartes' physical concepts and later developments is underlined by his account of weight. For Newton weight is a relational property of a body that depends on its location, but is proportional to its mass (for details see Sec. 9.4). But on Descartes' account, weight is neither an intrinsic property of a body nor related to an intrinsic property. Descartes' account of weight is directly related to his theory of *vortices*. As we have seen, Descartes holds that physical objects have a determination to move in a straight line, but that actual motion in the world is roughly circular. Given an initial chaotic state of matter, a universe without voids, and the laws of motion, Descartes holds that the universe would necessarily develop a large number of centers of rotation, best thought of as vortices in a fluid (III 23–34). The sun and each of the stars is at the center of a vortex and the heavens surrounding each of these move in a continual circular flow. In addition, Descartes allows for smaller vortices that occur in the large vortex around a star. For example, there is a vortex around the earth that carries our moon. Descartes is also clear that the vortex around an object turns the object on its axis (IV 22). Now recall Descartes' account of an object that is constrained to move in a circle, such as a stone being whirled in a sling, and assume that the sling provides the only constraint on the object. As long as the sling restrains the stone, the stone presses on the sling in the direction away from the center of motion, and thus exhibits a determination to move in that direction (III 55–59); this same tendency is found in all matter trapped in a vortex (III 60–62). But Descartes also holds that smaller particles generally move faster than larger particles (III 51); the smallest particles are more agitated – have more internal motion – than medium sized particles; these are more agitated than the largest particles. The effect of this difference is that more agitated particles have a greater determination to recede from the center than those which are less agitated (e.g., III 52, 57; IV 15). Given Descartes' view that an object's determinations can change while its QM remains the same, his point seems to be that an agitated particle has a high QM plus a variety of determinations to move in various directions; if these determinations line up, a rapidly moving particle will result.

Now consider objects near the earth and focus on the earth's daily rotation which is brought about "by the heavenly matter which surrounds it and

which permeates all its pores . . . " (IV 22: 190). Because of this circular motion all particles tend to move away from the earth, and this tendency is greater for the smaller particles. But for a particle to move away from the earth, other particles must be moving towards the earth creating a circular flow (IV 26).[29] In effect, the particles moving away from the earth – which constitute the primary phenomenon – create a pressure on slower particles to move towards the earth. This inward pressure will overwhelm the centrifugal pressure of the larger particles which will then be under a net pressure towards the center of the earth. The net downward pressure constitutes weight (IV 20–27). For Descartes weight is solely the result of this external pressure by other bodies, and he explicitly concludes that "weight does not correspond to the quantity of matter in each body" (IV 25: 192). If we can imagine a world consisting only of the planet earth and some object at a distance from the earth, then for Descartes this object would have no weight and no tendency to move towards the earth. Aristotle and Newton disagree, although for different reasons.

Our next task is to consider the various forces that Descartes invokes and their relation to QM. These are best discussed in the context of the rules of impact – which provide Descartes' most developed fragment of a mathematical physics. Descartes considers a highly idealized situation: collision of two hard bodies in isolation from any other bodies. All motions are along a single straight line and rebounds occur without either body being compressed. An isolated moving body has a single determination that coincides with its direction of motion, while an isolated stationary object has no determination in any direction. Descartes provides seven rules that fall naturally into three sets.

I begin with Rules 4–6 which deal with cases in which a moving body B hits a stationary body C. Rule four (henceforth R4) tells us that if C is even slightly larger than B, C will not move "no matter how great the speed at which B might approach C. Rather, B would be driven back in the opposite direction . . . " (II 49: 66). Descartes adds that the faster B moves, the greater the force with which C resists: "a body which is at rest puts up more resistance to high speed than to low speed; and this resistance increases in proportion to the difference in speeds" (II 49: 66). It follows from the third law that B's quantity of motion remains unchanged after rebound, but its direction (and thus its determination) is reversed.

R5 addresses the case in which the stationary body is slightly smaller than the moving body. After impact the two bodies will move with the same speed in the same direction as B was originally moving. Descartes is clear that this result holds "no matter how slowly B may advance toward C" (II 50: 67). In the French edition he adds: "it is impossible for B to have so little force that it would ever be insufficient to move C. . . " Descartes gives two examples to show how the resulting speeds can be calculated; Table 9.1 summarizes one of these examples. The requirement that the speeds are equal after impact is an additional assumption that – along with CQM – determines the final

speed and implies that after collision QM is distributed in proportion to the relative sizes of the two objects. A parallel calculation might seem appropriate for R4. (I will use the same label for a rule and for cases that fall under that rule, where no ambiguity results.) For example, if the stationary body C were three times the size of B, and B were moving with four units of speed, after impact each body would be moving in B's direction with one unit of speed. But Descartes does not permit this calculation. Instead, R4 tells us that if C is smaller than B, C's resistance increases as B's speed increases, and C will never move. If B is larger, C never has sufficient force to resist being moved.

R6 deals with cases in which B and C are the same size. Here, Descartes tells us, C would be driven forward and B would recoil. He also gives an example of how the final speeds are distributed: If B initially has four units of speed, after impact C will have one unit and B will have three. The French edition provides an explanation of how Descartes arrives at this result: he treats this case as intermediate between the two previous cases, and averages the resulting *speeds* from those cases. If we assimilate this case to R4, B rebounds and transfers no speed to C. If we do an R5 calculation both objects move in B's direction, each with two units of speed. Averaging these cases leaves B with three units of speed and C with one unit. In fact, Descartes' method of calculation yields a general conclusion that he does not state: C acquires one-quarter of B's original speed while B recoils with three-quarters of its original speed. However, since QM is a scalar quantity, it is also conserved by other outcomes – such as B and C moving off in opposite directions – each with half of B's original speed. Again, CQM is not sufficient to determine the outcome; nor is it sufficient in the cases

*Table 9.1* Descartes' Speed Calculation

| | *Units of size* | *Units of speed* | *QM* |
|---|---|---|---|
| *a) Before impact* | | | |
| B (moving) | 3 | 4 | 12 |
| C (stationary) | 1 | 0 | 0 |
| Total QM before impact | | | 12 |
| *b) After impact* | | | |
| B (moving) | 3 | 3 | 9 |
| C (moving) | 1 | 3 | 3 |
| Total QM after impact | | | 12 |

discussed below. Descartes must make additional assumptions to determine the outcomes in each case. In accordance with TC, these assumptions constitute part of the content of the concepts that occur in them.[30] Descartes does not tell us how he decides whether B continues or rebounds, but his averaging technique leaves B moving faster than C, so rebound is the only coherent possibility. Unfortunately, R6 contradicts Descartes' third law. The second part of the law says that when B has sufficient force to move a previously stationary body, B "moves the other body with it." This suggests that both bodies move in the same direction, but R6 requires that if B and C are the same size they move in opposite directions after impact.

We can now note that a stationary body's FORCE TO RESIST BEING MOVED is a *relational concept* since the force depends on the relative sizes of the interacting bodies. A stationary body has unlimited capacity to resist being moved by a smaller body (although the force is finite in any given case) and no force to resist being moved by a body of equal or larger size. It follows that a moving body's FORCE TO MOVE A STATIONARY BODY is also relational – as is its FORCE TO CONTINUE IN THE SAME DIRECTION and its FORCE TO CONTINUE AT THE SAME SPEED, when impacting a stationary body.[31] The moving body's QM does not provide a measure of any of these forces. As MM point out (66, n. 53): "resistance to motion depends entirely on relative size. Quantity of motion plays no role whatever, except that it must be conserved." It is worth pressing this point with some examples. Consider a stationary body C with four units of size, and three different impacts by a body B that has a QM of twelve units. First (R4), B has three units of size and four units of speed: C does not move while B changes direction and maintain its speed. Second (R5), B has six units of size and two units of speed: C moves while B loses speed but retains its direction. Third (R6), B has four units of size and three units of speed: C moves while B loses speed and changes direction.[32] We can also see that Garber's attempt to rescue the Cartesian account of motion will not work in this case. Descartes requires an absolute distinction between motion and rest that cannot be reduced to consideration of the separation between two bodies.

The remaining rules deal with cases in which neither body is initially at rest. R1-R3 concern collisions between two bodies that are moving towards each other. R1 is the only rule of impact that gives the same result as later physics (II 46): If two bodies of equal size move towards each other with equal speeds, then after impact they will move away from each other at equal speeds. Still, we should keep in mind the vast differences between Descartes' conceptual framework and those of his successors. It would, for Descartes, be a serious conceptual error to shift to a frame of reference in which one of the objects is stationary – this is a fundamentally different situation. R2 and R3, which deal with cases in which either the sizes of the bodies or their speeds are not equal, underline the point that Descartes does not allow assimilation of one case to another by shifting frames of reference.

R2 concerns two objects, B and C, moving towards each other with the same speeds, but B is slightly larger. At impact "only C would spring back" (II 47: 65); both bodies retain their speeds, but the smaller body changes direction. Given Descartes' remark that C would "spring back" they presumably still move as distinct objects. This is another example of the tendency of bodies to maintain their state: "a body which is joined to another has some force to resist being separated from it, while a body which is separate has some force to remain separate" (II 4: 63). Comparing R1 and R2 we find that the force an object has to continue moving in the same direction again depends on its relative size in an impact.

R3 deals with cases in which B and C are the same size but B's speed is greater than C's: "one half of B's additional speed would be transferred from it to C" (II 48: 65) and the two bodies would continue to move in B's direction with the same speed. Suppose we compare R3 with R6 (which also concerns bodies of the same size), thinking of rest as a speed of zero. In R3 the faster body maintains its direction while in R6 it changes direction. However, there is no inconsistency here since, for Descartes, the comparison is not appropriate. The fact that C is moving in one case and stationary in the other entails that these are different kinds of cases. Rest should not be thought of as a speed of zero, but as a case that falls under the concept REST, which is incompatible with SPEED.

Descartes does not consider situations in which both size and speed differ, but one case seems clear. Since the outcome of impact is the same if just B's size is larger, and if just B's speed is larger (both move in B's direction at the same speed), if B is both larger than C and moving faster, after impact B and C will presumably move in B's original direction at the same speed. It is, however, unclear what will happen if one object is larger but the other is faster. Note especially that Descartes seems to consider the *magnitude of the differences* in size and speed as irrelevant.

Our third set, which contains just R7, deals with cases in which B and C are moving in the same direction but B is faster so that it overtakes C (II 52). In the French version of PP Descartes considers three cases; in all cases the slower object C is larger.[33] In discussing these cases Descartes introduces a new consideration: the ratio of two ratios. Some notation will be helpful:

r(sp) = B's speed/C's speed, r(sp) > 1 for B to overtake C;
r(sz) = C's size/B's size, r(sz) > 1 for all cases considered;
R = r(sp)/r(sz).

Descartes considers the three possible values of R (II 52: 68).

> R > 1: "B would transfer to C as much of its speed as would be required to permit them both to travel subsequently at the same speed and in the same direction."
>
> (R7a)

> R < 1: "B would be driven back in the opposite direction, and would retain all its movement."
>
> (R7b)

> R = 1: "B must transfer some of its motion to C and spring back with the rest."
>
> (R7c)

Descartes gives numerical examples to illustrate the first two cases, but not the third. Nor is there any discussion of cases in which a body overtakes either a smaller body or one of equal size, although Descartes adds "and so on" to his second example. He also tells us, "These things require no proof, because they are obvious in themselves" (II 52: 69); a bit of algebra gives some sense to this claim since R is the ratio of B's initial QM to C's initial QM. His three cases give coherent results from this perspective. Further, where C is not larger than B, R > 1. Thus, after impact B and C continue to move in the same direction but now at the same speed.

Suppose we attempt to calculate R for cases covered by R1–R6. When one object is stationary we immediately run into a problem: viewing the stationary object C as having a speed of zero puts a zero in the denominator of r(sp) so we cannot proceed. Moreover, if we treat R as the ratio of quantities of motion, the denominator of R is zero. The appropriate conclusion, I submit, is that QM does not apply to stationary objects; REST does not imply QM. This squares with Descartes' view that rest is fundamentally different from motion.

Next, consider two bodies moving towards each other. R1 deals with cases in which speeds and sizes are equal, so R = 1. Descartes' claim that the bodies move away from each other after impact accords with R7c, but there is no transfer of motion in cases covered by R1. In cases covered by R2 and R3 we have two objects with different quantities of motion, but we have no criterion for deciding which quantity to put into R's numerator, and which in the denominator; thus R is indeterminate. This problem does not arise for R7 because the numerator is QM for the faster object. We see again that cases in which objects move towards each other are different from cases in which one object overtakes another – in accord with Descartes' treatment of directions of motion as absolute. There is no internal inconsistency here because in Descartes' conceptual framework the three types of cases are genuinely different and thus involve different concepts. CQM does apply to all cases but is never sufficient to determine outcomes, and the unstated additional rules that Descartes seems to use vary among the different cases.

Several scholars point out that, for many seventeenth century thinkers, "Interactions between bodies were seen as contests between opposing forces, the larger forces being the winners, the smaller forces being the losers . . . " (Gabbey 1980: 243, cf. Garber 1992a: 233–34; Gaukroger 2002: 122). In the case of Descartes' rules of impact it is unclear how to measure the forces,

what counts as a contest, and what counts as a winner. Table 9.2 summarizes the qualitative outcomes. Garber, defending the contest view, points out that R2, R3, and R7a all deal with collisions in which the body with the larger initial QM maintains its direction, and equates this with winning the contest (1992a: 239). But in R7a the body with the smaller initial QM also maintains its direction and gains QM; the other body loses QM. Why not consider a body that gains QM the winner? The body with the smaller initial QM also gains QM in R3, but not in R7b. We have two cases in which bodies have the same initial QM: in R1 there is no gain or loss of QM, but there is in R7c. In R7c only one of the bodies (the one that is smaller and moving faster) changes direction, but in R1 we have two bodies with equal size, speed, and QM, and both reverse their directions. R4–R6 involve a different kind of contest – one between motion and rest (see Gabbey 1980:

*Table 9.2* Results of Impacts According to Descartes' Rules

a) Both bodies are initially in motion

| *Rule* | *QM* | *Effect on B* | *Effect on C* |
|---|---|---|---|
| R1 | QM(B) = QM(C) | changes direction<br>QM unchanged | changes direction<br>QM unchanged |
| R2 | QM(B) > QM(C) | maintains direction<br>QM unchanged | changes direction<br>QM unchanged |
| R3 | QM(B) > QM(C) | maintains direction<br>loses QM | changes direction<br>gains QM |
| R7a | QM(B) > QM(C) | maintains direction<br>loses QM | maintains direction<br>gains QM |
| R7b | QM(B) < QM(C) | changes direction<br>QM unchanged | unchanged |
| R7c | QM(B) = QM(C) | changes direction<br>loses QM | maintains direction<br>gains QM |

b) C is initially stationary

| *Rule* | *Size* | *Effect on B* | *Effect on C* |
|---|---|---|---|
| R4 | SZ(B) < SZ(C) | changes direction<br>QM unchanged | unchanged |
| R5 | SZ(B) > SZ(C) | maintains direction<br>loses QM | acquires direction<br>acquires QM |
| R6 | SZ(B) = SZ(C) | changes direction<br>loses QM | acquires direction<br>acquires QM |

260–72 for this point and an account of the various contests). We could say that in R5 and R6 B wins since C's state of motion is changed, but B's speed is also changed, which is a change of state. In R4 neither object changes its state of motion or rest, so there is no winner of this contest. But B's direction is changed. Perhaps there is a second contest here that C wins. I suggest that if there is a coherent account in these rules, the view of impacts as contests will not help us locate it.

Descartes' discussion of impact invokes many forces: A stationary body has a force to resist being moved; a moving body has forces to move a stationary body, to maintain its direction of motion, and to maintain its speed. In an impact between two moving bodies, each has forces to change the speed and direction of the other, as well as to maintain its own speed and direction. Many, perhaps all, of these forces are distinct. We may require further distinctions depending on whether two objects are moving in the same or different directions. It would be tedious to work through all possible cases and attempt to decide when we are dealing with different forces, and when with the same force in different circumstances. Nor is it clear that Descartes has provided enough detail to work out all these possibilities – although he seems to think that he has provided a unified account of force. In an article titled "In what the force of each body to drive or resist consists" he writes:

> We must however notice carefully at this time in what the force of each body to act against another or to resist the action of that other body consists: namely, in the single fact that each thing strives, as far as is in its power, to remain in the same state, in accordance with the first law stated above. . . . One which is at rest has some force to remain at rest, and consequently to resist everything which can change it; while a moving body has some force to continue its motion, i.e., to continue to move at the same speed and in the same direction. Furthermore, this force must be measured not only by the size of the body in which it is, and by the [area of the] surface which separates this body from those around it; but also by the speed and nature of its movement, and by the different ways in which bodies come into contact with one another.[34]
>
> (II 43: 63)

The passage includes factors that play no role in the rules of impact, presumably because Descartes is concerned here with motion in the actual world, not with the idealized cases covered by the rules. But this raises a question about the relevance of the rules to actual cases. This question is also raised by another article that deals with real objects (II 26); here Descartes asserts that the force required to initiate motion is the same as that required to stop it. Yet the rules of impact do not mention any cases in which a body's motion is stopped. According to these rules, when a moving object's speed is reduced it loses either 1/4 or 1/2 of its previous speed. This may be an

artifact of the idealization. In his remarks on projectiles (II 38) Descartes notes that in the real world moving objects are quickly brought to a stop as a result of multiple impacts. But the rules of impact imply that stopping a moving object is a more complex process than initiating motion in a stationary object, which requires only the slightest nudge from a moving object of equal or greater size. In his comments on the third law Descartes also says that when hard moving bodies, "strike a yielding body to which they can easily transfer all their motions, they immediately come to rest" (II 40: 61–62). No such cases are included in the rules of impact, and Descartes does not tell us how CQM works in this new case. In fact, Descartes is well aware that his laws and rules do not accord with experience; I will consider his response to this situation when we look at the ICs for his dynamical concepts.

Our discussion underlines the deep conceptual difference between MOTION and REST in Descartes' framework. Descartes holds that motion and rest are contraries, but this description is a bit misleading. In its common usage two properties are contraries if they cannot both exist in the same object at the same time and same respect, but it is possible that neither of a pair of contrary properties exist in a given object. Aristotelian natural and violent motion are contraries, and neither occurs in an object at rest. For Descartes, motion and rest are *contradictories* – every object must be in one of these states.[35] The distinction between contraries and contradictories is worth noting because it helps clarify some points. Descartes emphasizes that "movement is not contrary to movement, but to rest . . . " (II 44: 63, repeated at II 56). Yet a moving body must have a specific speed and direction at an instant; different speeds form a set of contraries, as do different directions. Descartes' point seems to be that differences between various speeds and between various directions are less fundamental than the difference between motion and rest. He also tells us that "determination in one direction is the opposite of determination in another" (II 44: 63), and that "the determination of movement in one direction is contrary to its determination in the opposite direction . . . " (II 56: 71). But we have seen that an object can have multiple determinations simultaneously – his account of light requires this. Thus different determinations are not contraries.[36] Unfortunately, this raises another question about the relevance of the rules of impact for actual cases since direction of motion and determination are treated as equivalent in the idealized situation.

This discussion of implications among the key concepts of Descartes' dynamics has shown that his conceptual system for physics is neither consistent nor complete. Still, Descartes' work provides an important historical stage in the development of physics. I want to examine some of its main overlaps and contrasts with the implicational structure of Aristotle's physics.

Consider STATE, understood as any property of an object that remains unchanged absent interactions with other objects. If we seek an instance of this concept in Aristotle's terrestrial physics the only candidate is rest at an

object's natural place. An unconstrained object at any other place in the terrestrial realm moves spontaneously to its natural place. Aristotle would agree that no object resting at its natural place "will ever begin to move unless driven to do so by some external causes" (II 37: 59), but reject Descartes' next sentence: "Nor, if it is moving, is there any significant reason to ever think that it will cease to move of its own accord and without some other thing which impedes it." Still, a form of motion that continues without an external force has a long-standing precedent in the traditional account of the heavens. While the motion that constitutes a Cartesian state is linear, not circular, and can occur anywhere in the universe, the view that motion can be self-sustaining does not appear *ex nihilo*.

Aristotle and Descartes agree that motion and rest are fundamentally different, but Descartes departs from Aristotle in several respects. For Aristotle's REST implies NATURAL PLACE since unconstrained rest can occur only at an object's natural place. Descartes' thesis that all material objects share a single essence removes any reason for including NATURAL PLACE in his physics; unconstrained rest can occur anywhere in the universe (ignoring impacts). The move to a single essence also eliminates any reason for attributing an intrinsic structure to terrestrial space and for treating the vertical and horizontal directions differently. Descartes' extension of this single essence to the heavens eliminates a central reason for providing different accounts of celestial and terrestrial motion. Cartesian UNIFORM MOTION shares one feature with Aristotle's NATURAL MOTION: neither requires an external sustaining force. But Aristotelian natural motion is not uniform (motion to a natural place implies acceleration), and leads to its own termination. Cartesian uniform motion can occur in any direction, anywhere in the universe, and is eternal.

Aristotle and Descartes both consider weight a phenomenon that physics must explain. But the differences between Aristotelian and Cartesian physics require different accounts of weight. For Aristotle weight is a manifestation of motion to a natural place. A single earthy object alone in the universe would move spontaneously toward the center of the universe and thus exhibit weight. For Descartes weight depends on impacts by other objects so that an isolated object would not exhibit weight. It is illuminating to consider how Aristotle and Descartes would deal with the motion of a helium-filled balloon were they to encounter one. Each would treat this as an instance of the same general type of phenomenon as weight, although their detailed accounts would be quite different. For Aristotle we have a mixed object dominated by the element air moving toward its natural place; for Descartes we have a result of impacts and the tendency of smaller objects to move away from a center of rotation more rapidly than larger objects. In general, we find the differences between the Aristotelian and Cartesian accounts of matter, motion, and weight set against a shared background. Many of the differences in the two frameworks can be generated by changes in the implications that link fundamental concepts. Some of these changes consist of dropping links to concepts that Descartes eliminates.

Now consider the ICs for the central concepts of Cartesian mechanics. Recall that ICs are criteria for deciding if a concept is instantiated, and thus provide the basis for empirical evaluation of a conceptual system. To a large extent both Aristotle and Descartes attempt to give accounts of phenomena that can be picked out on the basis of unaided observation and general knowledge (fall, weight, projectile motion), quite independently of either theoretical structure. Still, theoretical considerations play a large role in Descartes' ICs. For Aristotle rest and motion are just the familiar everyday phenomena, but Descartes' ICs for MOTION and REST depend on a technical account: whether a body is in motion or at rest is determined by comparing it with adjacent bodies. As a result, Aristotle and Descartes disagree on whether the moon, for example, moves. Still, given Descartes' account of motion, application of these concepts requires just ordinary perception. The case is not so straightforward for the various forces that Descartes invokes because of the extreme nature of the idealizations at the heart of his account of these forces. Aristotle does not permit idealizations as a means for understanding real phenomena; Galileo does. We encountered Galilean idealizations in his treatment of a small portion of the circumference of a large circle as a straight line, and two strings pointing to the distant center of the earth as parallel. These are limited idealizations that give close approximations to the actual situation. In Descartes' account of the rules of impact the idealizations are much more extreme – by his own lights. In order to limit consideration to one impact at a time Descartes considers isolated bodies, ignoring the plenum nature of the universe. Since his account of motion depends on adjacent bodies, in the idealization it is indeterminate which bodies are moving and which are at rest.

The rules of impact give results that are empirically false, and Descartes is aware of this. In the French edition of PP, immediately after his discussion of the rules, he writes: "Indeed, experience often seems to contradict the rules I have just explained" (II 53: 69). He appeals to the use of idealizations to account for this contradiction, and gives some indication of how to move to a less idealized case by discussing a body immersed in a fluid (II 54–61). Still, these accounts are not developed in a way that would allow any empirical outcome to provide a challenge to a law of nature or rule of impact. Indeed, the discussion of R7 in the French edition ends: "And the demonstrations of this are so certain that, even if experience were to appear to show us the opposite, we would nevertheless be obliged to place more trust in our reason than in our senses" (II 52: 69, n. 62). We encountered a somewhat similar situation in Galileo where determining the state of motion of a stone requires that theoretical consideration override naïve observation. But in Galileo's hands theoretical considerations are quickly connected to observable situations – such as fall on a moving ship – which could contradict the predictions of his theoretical account. An often-cited remark of Galileo's may seem to close the gap between his view and Descartes'. Galileo praises earlier astronomers who accepted the motion of the earth even

though it contradicted experience: "they have through sheer force of intellect done such violence to their own senses as to prefer what reason told them over that which sensible experience plainly showed them to the contrary" (1967: 328). Yet Galileo's practice shows that he considers this situation unsatisfactory. In *Dialogue* he endeavors to overcome it by refuting arguments which seem to show that the motion of the earth is contrary to experience, and by seeking evidence for this motion. The latter requires tying the concepts of his physics to experience in a way that Descartes never does. In particular, we have no way of evaluating the instantiation of the various force concepts that Descartes introduces.[37]

The IC for DETERMINATION is also problematic. While a given object may have multiple determinations in different directions at the same time, it is not clear from Descartes' account how these determinations are to be identified. A light-emitting object has determinations in every direction, but a fire may also be in motion (as judged by Descartes' criterion). This motion will bring along determinations of its own, and it is unclear how these relate to the determinations generated by the emission of light. In his discussion of a rock whirled in a sling Descartes mentions two determinations – one tangential and one radial. He suggests that we can identify the radial determination by sensation, but his attempt to establish that there is also a tangential determination is inadequate. In this case Descartes' primary aim is to argue that there is no circular determination, which he does by means of the observation that when the stone is released it moves off in a straight line. But according to his own physics, this is not true – all actual motions follow closed paths. Thus motions that appear curved do not reveal the physically fundamental form of motion – which is linear; but motions that appears to be linear should be treated as misleading since this kind of motion does not occur in the physical world. Moreover, a circular determination seems to appear in a later discussion of the stone in the sling (III 57), and several commentators argue that circular determination appears at various places in Descartes' writings and letters.[38] DETERMINATION is one of Descartes' central physical concepts, but he never provides the IC required for an adequately developed descriptive concept.

Since particles in the plenum continually receive impacts from many directions, their direction and speed frequently change. But to determine direction we must track a particle over some distance, and to determine speed we must track it over both distance and time. Yet neither direction nor speed is stable enough to permit such measurements. When they appear to be stable, this must be written off as an illusion to be overridden by theoretical considerations. Since a body's QUANTITY OF MOTION is size multiplied by speed, the problems with determining speed infect this concept too.

Now consider *systemic roles*. In the case of ELEMENT I noted above that although Descartes continues to use this language, the older concepts are dropped. In earlier systems of mechanics the elements distinguish types of objects that follow different mechanical laws. In Descartes' system there is one set of laws for all physical objects; thus the systemic role of ELEMENT has

been eliminated. For Descartes STATE covers properties of an object that remain unchanged unless some external force acts on that object. As he notes, the idea that some properties remain constant in this way is familiar, and he extends the scope of this concept by applying it to uniform motion, but the role of STATE remains unchanged. Forces also continue to play their familiar role as specifying what is needed to alter a state. Cartesian forces are also invoked to maintain a state of rest or uniform motion. But the forces Descartes considers are all relational: their magnitudes depend on properties of two interacting bodies. This suggests that these forces do not exist in a body, whether at rest or in motion, in the absence of impacts. On the other hand, there are also places where Descartes suggests that continued uniform motion requires a sustaining force (e.g., his discussion of reflection of light). It seems that the systemic roles of the various forces he invokes have not been clearly developed.

The roles of NATURAL MOTION and VIOLENT MOTION have been absorbed by STATE. Although both rest and uniform motion count as states, the distinction between motions that require an external sustaining force, and those that do not, is retained. The role of QUANTITY OF MOTION is to specify a property that is conserved both in individual interactions and in the universe as a whole. This may be a new systemic role – one that takes on central importance as physics develops. DETERMINATION may also be a new systemic role. Although Descartes never develops the concept adequately, it can (with some generosity) be read as a precursor of the idea that direction and speed are equally fundamental in understanding motion.

On the whole, Descartes has failed to provide an adequate system of descriptive concepts for dealing with motion. Problems arise particularly on two dimensions. In the case of implications we have found both incompleteness and inconsistency; in the case of ICs we have found that Descartes' concepts are not sufficiently tied to their subject matter to allow us to determine if they are instantiated. As a result of the latter failing, the theoretical claim that this conceptual system provides a model for the physical world is untestable. Nevertheless, for part of the seventeenth century Cartesian physics played a dominant role in the thought of many physicists and was the view Newton had to overcome.

## 9.4 Newton[39]

Westfall notes that by the time Newton came on the scene Aristotelian physics was no longer in play: "As far as men active in the study of nature were concerned, the word 'overthrown' is not too strong. For them, Aristotelian philosophy was dead beyond resurrection" (1983: 14). Cartesian physics was at center stage. Newton had learned a good deal of mathematics from Descartes' *Geometry*, and there is no doubt that Newton was thoroughly versed in at least Books I, II, and III of PP since he wrote a detailed critique of this material.[40] Although *Principia* contains few explicit references to

Descartes, Newton systematically argues that the vortex theory is incompatible with each of Kepler's laws of planetary motion and with the motion of comets. Much of this argument is in Section 9 of Book II and its concluding "Scholium." Newton opens the "General Scholium" that appears at the end of the second and third editions of *Principia* with a summary of the case against the vortex hypothesis; Cotes "Preface" to the second edition sums up this case at greater length. Newton returns to the critique of Cartesian physics in the "Queries" he included in the final edition of his *Opticks* (1952: 362–65, 368–69, 397–400).

Some of Newton's definitions at the beginning of *Principia* could almost have been written by Descartes.[41] For example, Newton's second definition reads (404):

> *Quantity of motion is a measure of motion that arises from the velocity and the quantity of matter jointly*.
>
> (D2)

This is followed by the comment:

> The motion of a whole is the sum of the motions of the individual parts, and thus if a body is twice as large as another and has equal velocity there is twice as much motion, and if it has twice the velocity there is four times as much motion.
>
> (404)

However, Newton's first definition (403) has put us on notice that his D2 does not mean quite the same thing as it would from Descartes.

> *Quantity of matter is a measure of matter that arises from its density and volume jointly*.
>
> (D1)

Descartes would make just one key change, substituting "size" for "density." But for Descartes size is volume, and the shift from volume to density has, we will see, major ramifications.

The next two definitions (404–5) could have also come from Descartes' pen.

> *Inherent force of matter is the power of resisting by which every body, so far as it is able, perseveres in its state of either of resting or of moving uniformly straight forward.*
>
> (D3)

> *Impressed force is the action exerted on a body to change its state either of resting or of moving uniformly straight forward.*
>
> (D4)

In addition, Newton's first law of motion (416) also looks Cartesian:

> *Every body perseveres in its state of being at rest or moving uniformly straight forward, except insofar as it is compelled to change its state by forces impressed.*
>
> (L1)

We will see that the import of these claims is quite different in the two systems of physics, but there is ample justification for approaching Newton's physics in terms of a contrast with Cartesian physics. Throughout this discussion I will assume general familiarity with Newtonian physics and focus discussion on Newton's key conceptual innovations. I state Newton's remaining laws of motion (416–17) here for ease of reference.

> *A change in motion is proportional to the motive force impressed and takes place along the straight line in which that force is impressed.*
>
> (L2)

> *To any action there is always an opposite and equal reaction; in other words, the actions of two bodies upon each other are always equal and always opposite in direction.*
>
> (L3)

Before moving to a more detailed discussion it will be useful to have a sketch of the structure of *Principia* before us. Following the prefatory material we find a series of definitions, then a scholium in which Newton discusses space and time, then the three laws of motion and six corollaries. The bulk of *Principia* begins after these and is divided into three "Books." The first two books are both titled *The Motion of Bodies* and are supposed to be purely mathematical; Book III, *The System of the World*, applies the mathematics to astronomy. This application requires of a body of empirical evidence and a set of methodological rules; I will postpone consideration of these until we need them. Book I deals with motion in non-resistive media; Book II adds consideration of motion in resistive media and is central to Newton's critique of the Cartesian plenum. In both of these mathematical books Newton explores a wide range of situations irrespective of whether they have any direct application to physics. In addition, each book is divided into sections. In the first section of Book I Newton proves several mathematical lemmas that he will use throughout the text; other lemmas are proved as needed. The numbered "Propositions" begin in Sec. 2.

Much of *Principia* is concerned with what Newton calls "centripetal force." He introduces this concept in D5 as any force by which bodies "*are impelled, or in any way tend, toward some point as to a center*" (405). The next three definitions introduce three different *measures* of the quantity of centripetal force. To explain these measures Newton first considers an object

such as a magnet that serves as a force center. The total magnetic force, which depends on the "bulk or potency" (406) of the magnet, is its *absolute quantity of centripetal force*; this is the force measured at the point from which it acts. The *accelerative quantity of centripetal force* is the acceleration that a force causes on a body at a given distance from the force center. It is a central feature of Newtonian physics that in the case of gravitation this measure is independent of the mass and kind of body on which the force acts. Finally, the *motive quantity of centripetal force* is measured by the quantity of motion (D1) generated in a body in a given time. This measure depends on the body's distance from the force center (like the accelerative measure), but is also directly proportional to the body's mass. For reasons that will become clear shortly, Newton gives weight as an example. *Note that these are different measures of a single force, not different forces.* Newton emphasizes that he is now discussing forces purely mathematically; he is not proposing any physical causes (407–8).

We are now ready to discuss Newton's conceptual system, beginning with STATE. While there is considerable overlap with Descartes' notion, there is also a major difference because Newton fully integrates motion and rest; rest is motion with zero speed, not a distinct state governed by its own laws. For a specific body, the quantity of force required to bring about a particular change of speed in a given time is independent of the initial and final speeds. This holds, for example, whether the change is from zero units to ten, from ten to twenty, from twenty to ten, from ten to zero, or from five units towards the right to five units towards the left. The fact that the last of these involves a brief passage through zero speed has no special significance.

Newton also breaks with Descartes in treating change of speed and change of direction as the same dynamical phenomenon, although Newton does not use the later vector concept of velocity as a single item that includes both features. The first step towards this integration occurs in L2. The statement of this law does not make it immediately clear that Newton is integrating change of direction with change of speed because he does not specify what counts as "a change of motion." The earliest numbered propositions of *Principia* deal only with changes of direction: Newton explores the forces required for deviations from motion in a straight line, as required by L1. As de Gandt (1995: 256) points out, the full integration of the two kinds of change becomes explicit in Newton's proof of I.40 where he treats a single force as generating change in both speed and direction. The subject of the proposition is "*To find the orbits in which bodies revolve when acted upon by any centripetal force*" (528). In this proof Newton uses "velocity" as a synonym for "speed" and "acceleration" as a synonym for "change of speed." But this language notwithstanding, he examines a body moving in a curved path, resolves the force acting on this body into components along the path and perpendicular to it, and tells us that the perpendicular component "will in no way change the velocity of the body in that path but will

only draw the body back from a rectilinear path and make it turn aside continually from the tangent of the orbit . . . " (528). The other component, "acting along the body's path, will accelerate the body and in a given minimally small time will generate an acceleration proportional to itself" (528). At this point Newton has made the key conceptual steps of including both change of speed and change of direction in "change in motion" and recognizing that a single force can produce both effects in a body simultaneously. As Westfall puts it, Newton has recognized "the dynamical identity of uniform circular motion and uniformly accelerated motion in a straight line. Heretofore in the history of mechanics, these two motions had been treated as irreducible opposites" (1983: 416). Conceptually, change of direction and change of speed are now recognized as aspects of a single phenomenon; it is a small linguistic step to extend the word "acceleration" to include both. I will henceforth use "acceleration" in this extended sense, even though Newton does not do this.

Using this later terminology, L2 embodies two fundamental conceptual innovations. First, according to this law FORCE implies ACCELERATION: The familiar forces typified in everyday experience by pushes and pulls yield a constant acceleration, not a constant speed. Second, the implication between ACCELERATION and FORCE is mutual. As a result, ACCELERATION acquires a new systemic role: Any deviation from motion in a straight line requires a force. Every such deviation implies the presence of acceleration, irrespective of whether there is any change of speed. Given this mutual implication, FORCE also acquires a new role: A single force may be responsible for a change of speed, a change of direction, or both together.

FORCE acquires another new role in Newton's physics: Weight is now conceptualized as a force. A body's weight does not measure the amount of matter in the body, but instead a specific force acting on the body. I want to summarize how this change comes about, and explore its significance. The first step to note is Newton's introduction of the *new concept* MASS, and his distinction between mass and weight. Mass is an *intrinsic property* of a given body; it provides a measure of the quantity of matter in that body which Newton defines as the product of volume and density (D1). Weight is a *relational property*; it is measured with respect to some other body, and depends on the masses of *both* bodies and on the distance between them. The most familiar case is the weight of a body near the earth. On Newton's account this weight is the force of gravitation on the body. Given that a body's mass and the mass of the earth are constants, this force varies (inversely with the square of) the body's distance from the center of the earth. Thus bodies do not have a single characteristic weight. Moreover, once we recognize that a body's weight depends on the earth's mass, it is a small step to conclude that the same body would have a different local weight if it were moved to the moon or one of the other planets (III.8 Cor.1). In addition, L3 requires equal and opposite forces between the two bodies we are considering. Gravitational attraction is mutual so that it is equally correct to talk about a

stone's weight with respect to the earth, and the earth's weight with respect to the stone – where the two weights are numerically identical. Although the mass of the earth is (essentially) constant, the earth has myriad weights with respect to myriad objects, and these weights vary as the distances of the objects from the earth vary; similar points hold for Jupiter, the moon, and so forth. Indeed, since gravitational force occurs between all bodies, any two bodies can be described as having weight with respect to each other, and this weight will also vary as the distance between the bodies varies (III.8). One conclusion to be drawn from this variability is that weight is not a fundamental physical property of bodies; mass is a fundamental property.

Note how different this concept of weight is from the concept we associate with "weight" in everyday thought. The common concept is confined to bodies near the earth where variations in distance have little practical effect (the relevant distance is measured from the center of the earth). This is the weight we measure with scales and balances, and for many practical purposes it is constant for a particular body. As a result, we do not usually distinguish mass from weight. Even physicists use the everyday concept in their mundane lives. We have, then, another example of an everyday concept with a practical function that exists alongside the scientific concept in the minds of the scientifically educated. There are both historical relations and systematic similarities between the two concepts. The everyday concept predates the Newtonian concept, and the former can be viewed as a special case of the latter under a set of limiting conditions. But this practical concept does not play a role in Newtonian and post-Newtonian physics; in physics a new concept replaced the older concept of weight. I want to use the tools of TC to examine further the content of the Newtonian concepts MASS and WEIGHT.

Note, first, some *implications* that are characteristic of WEIGHT in a Newtonian framework. Since all of an object's various weights are proportional to its mass, WEIGHT implies MASS, but there is no converse implication from MASS to WEIGHT. A single body alone in the universe would have mass but no weight. Astronauts experience a close approximation to this situation since they must regularly deal with the mass of bodies (exhibited by their inertia – to be discussed below) even though these bodies do not exhibit weight. WEIGHT also implies a second body, the DISTANCE between these bodies, and an equal FORCE on each. Since WEIGHT implies FORCE, it also implies ACCELERATION. None of these implications hold for MASS.

Turning to ICs, consider how we measure weight. The heft of a stone in my hand remains a rough measure of its local weight, and the standard instruments for measuring weight continue to function in the familiar ways for many purposes. However, these instruments are not satisfactory for precision measurements. One problem arises because the earth is not a perfect sphere – it bulges slightly at the equator and is flattened slightly at the poles. This small variation in the distance between the surface of the earth and the center shows up in precise measurements of weight. Although the effect is

small, it was measurable in Newton's day. Newton considers this case at length (826–32) in discussing III.20, "To find and compare with one another the weights of bodies in different regions of our earth" (826). He calculates that "the gravity at the pole is to the gravity at the equator as 230 to 229" (827), and reports results of careful measurements made by several people at various latitudes. Newton also notes that the latitude effect "is so small that in geographical matters the shape of the earth can be considered to be spherical, especially if the earth is a little denser toward the plane of the equator than toward the poles" (829).[42] Newton also notes the effect of buoyancy due to the air on measurements of weight (828). The case of a helium filled balloon will serve again to underline the point. Two relevant forces act on the balloon: gravitational force and a buoyant force resulting from the air. The balloon rises when the buoyant force is larger. In an airless environment, such as on the moon, the same balloon will fall. Thus familiar means of measuring weight give only an approximate value, and the significance of the approximation depends on the precision we are seeking. I will return to the role of approximations in Newtonian physics toward the end of this section.

The distinction between mass and weight separates *systemic roles* that had previously been combined in a single concept. In effect, two new roles are introduced, each taking on some of the features previously associated with weight (by everyday thought, but not by Descartes) and introducing new features. WEIGHT denotes one of the forces that may be acting on a body in a given situation. MASS denotes an intrinsic property of bodies and has two further functions in Newtonian dynamics. First, a body's mass gives that body's contribution to the gravitational attraction between it and any other body. Given Newton's definition of mass as the product of volume and density, the thesis that mass plays the specific role it does in gravitational attraction is a substantive additional claim. Second, mass occurs in L2 where it provides the measure of a body's resistance to acceleration. This resistance is the same for any force acting on the body, whatever its source. In Newtonian physics there is no reason why these two roles should be played by the same property. Nor is there any reason why these different kinds of mass need have the same numerical value, or the same values for different materials.[43] Newton was well aware of these issues and cited experiments that tested and confirmed these numerical identities. I will discuss these shortly, but first I want to note one profound consequence of this identification. Let the mass of the earth be M, the mass of some test body be m, the distance between the centers of these two bodies be R, and the gravitational constant (needed to transform a proportion into an equation) be G. According to Newton's law of gravitation, the gravitational force acting on the test body is $F = GMm/R^2$. According to L2 this force generates an acceleration that is inversely proportional to the body's mass ($a = F/m$). Substituting ma for F in the gravitational equation, the body's mass cancels out and it follows that all body's dropped from the same height fall to the earth with the same acceleration, whatever their masses.

We can now see the basis for Newton's distinction between the accelerative and motive measures of centripetal force. Two bodies of different masses at the same distance from a reference body will acquire the same acceleration with respect to that body; thus they will have the same accelerative measure of their quantity of force. But the motive measure of force will differ for these two bodies because different forces will be acting on them (they have different weights). In each case the force moving the body is proportional to its mass, so the larger body has a greater motive measure of force. This larger force is just sufficient to overcome that body's larger resistance to acceleration – which is why they have the same accelerative measure of force. In a given period of time the two bodies will acquire the same speed, but their quantities of motion – speed times mass – will be different.

Proposition III.6 states that the weight of a body toward any planet is proportional to that body's mass. In his justification Newton first notes that many observations have shown that all bodies fall to the earth from a given height in the same time (making the required adjustment for variations due to air resistance).[44] He then reports an experiment that addresses the question of whether bodies made of different materials behave differently.

> I have tested this with gold, silver, lead, glass, sand, common salt, wood, water, and wheat. I got two wooden boxes, round and equal. I filled one of them with wood, and I suspended the same weight of gold (as exactly as I could) in the center of oscillation of the other. The boxes, hanging by equal eleven-foot cords, made pendulums exactly like each other with respect to their weight, shape, and air resistance. Then, when placed close to each other [and set into vibration] they kept swinging back and forth together with equal oscillations for a very long time. Accordingly, the amount of matter in the gold (by book 2, prop. 24, cors. 1 and 6) was to the amount of matter in the wood as the action of the motive force upon all the gold to the action of the motive force upon all the [added] wood – that is, as the weight of one to the weight of the other. And it was so for the rest of the materials. In these experiments, in bodies of the same weight, a difference of matter that would be even less than a thousandth part of the whole could have been clearly noticed.
>
> (807, passages in square brackets are due to the translators.)

Such a difference would have raised a serious problem for Newton's dynamics.

Since Descartes holds that weight is a result of external forces acting on a body, he should agree that bodies do not have a fixed weight. But for Descartes a body isolated in a local vacuum would have no weight irrespective of any nearby bodies. In Newtonian physics a body in a local vacuum (as in Boyle's experiment) would still have weight with respect to other bodies. Moreover, Descartes does not admit any intrinsic property of a body that is

related to its weight; thus many implications of Newton's weight-concept are not found in Cartesian physics.

I now want to explore some further issues concerning Newton's concept of force. There is some disagreement among commentators over the status of the older force-concept in Newton's physics. The problem arises because Newton explicitly distinguishes the "inherent" or "innate" force of a body in motion from a force that is impressed on the body from the outside. D3 introduces the concept of an inherent force of matter; Newton then comments:

> This force is always proportional to the body and does not differ in any way from the inertia of the mass except in the manner in which it is conceived. Because of the inertia of matter, every body is only with difficulty put out of its state of either resting or of moving. Consequently, inherent force may also be called by the very significant name of force of inertia. Moreover, *a body exerts this force only during a change of its state, caused by another force impressed upon it* [my italics], and this exercise of force is, depending on the viewpoint, both resistance and impetus: resistance insofar as the body, in order to maintain its state, strives against the impressed force, and impetus insofar as the same body, yielding only with difficulty to the force of a resisting obstacle, endeavors to change the state of that obstacle.
>
> (404)

D4 explains impressed force as an action that changes a body's state:

> This force consists solely in the action *and does not remain in a body after the action has ceased* [my italics]. For a body perseveres in any new state solely by the force of inertia. Moreover, there are various sources of impressed force, such as percussion, pressure, or centripetal forces.
>
> (405)

Commentators reading these passages differ on whether Newton has achieved a fully inertial physics, or retained elements of the older view that a force is required to maintain uniform motion. The issue turns on how we understand the role of the *inherent force of matter* in Newton's physics. Jammer (1999: 120), Shapere (1967), and Westfall (1971: Ch 8, 1983: 416 *et passim*) are among those who hold that Newton does have a fully inertial physics; Gabbey (1971, 1980) and Cohen (1999) are among those who reject this claim. Cohen recently put his view this way:

> The primary distinction made in the definitions in the *Principia* is between those "forces" that preserve a body's state of motion or of rest and those that change the body's state. Today's reader will be puzzled by def. 3, in which Newton introduces a "force" of inertia, using "force" in

> a sense very different from later usage. No doubt this was a legacy from the traditional (pre-inertial) natural philosophy which held that there is no motion without a mover.
>
> (1999: 56; see also 98)

But we should be careful before accepting this interpretation since it leads to rather strange results by a fairly direct route.

To see the problem let us assume an inherent force that maintains body *A*'s uniform motion, and assume that *A* is hit from behind by a brief force that increases *A*'s speed, the direction remaining unchanged. According to the remark I italicized in Newton's comment on D4, the impact leaves no *additional* force in *A*. If a force maintains *A*'s new speed, it is exactly the same force that was maintaining *A*'s previous speed. This also holds if *A* was at rest before the impact: The same inherent force that was maintaining *A* at rest now maintains *A*'s motion. But in Newtonian physics a body can achieve any speed whatsoever, so this constant inherent force can sustain any speed from zero on up. This leaves no specific relation between this force and the body's speed. As Shapere notes (1967: 204–5), it then makes no sense to talk of the inherent force as a *cause* of *A*'s continued uniform motion.

Consider another approach. Mass shares this independence of speed with the supposed force, and D3 explicitly states that this force "does not differ in any way from the inertia of the mass except in the manner in which it is conceived." While mass is not a force, perhaps the inertia of mass is a force? But in D3 the key expression "perseveres in its state" is ambiguous. It could be referring either to the ability of a body to continue in its state of motion *tout court*, or to a body's tendency to maintain its state *when acted on by an impressed force*. On the latter reading no force would be involved in maintaining motion between external interventions. The *force* of inertia would appear only when an impressed force acts on the body. This reading is supported by the passage I italicized in Newton's comments on this definition: "a body exerts this force only during a change of its state, caused by another force impressed upon it." It is in this case that the force of inertia appears in its two guises. If we focus on *A*, its force of inertia appears as "resistance insofar as the body, in order to maintain its state, strives against the impressed force." But *A* is also acting to change the state of *B* (the object that is acting to change *A*'s state), as required by the third law. From *B*'s perspective, *A*'s force of inertia is impetus. In a similar way, *B*'s force of inertia is identical with its impetus to change *A*'s motion. None of these forces occur in the absence of impacts.[45]

The notion of a force that appears only in an interaction would have been familiar to Newton and his readers at least from Descartes' R4 which also postulates such a force. To be sure, beyond this minimal overlap, Newton's view is quite the antithesis of Descartes' since Descartes' R4-force is always sufficient to prevent a stationary body from being moved by a smaller body, while L2 implies that any force, no matter how small, generates some

acceleration in any body, no matter how large. If a force acts on *A* (ignoring oblique impacts), *A* accelerates as long as this force acts, and *A* reaches a new velocity. The magnitude of the velocity *increment* is inversely proportional to *A*'s mass. From the perspective of resistance, the inherent force resists change of motion in that the greater the mass, the smaller the speed increment for a given impressed force. Now suppose this force acts for an extended period of time – a case that Descartes does not discuss. L2 implies that any force can accelerate any body to any speed whatsoever *given enough time*. Once the force is specified, the body's mass determines the acceleration, and thus determines how much time will be required to reach a particular speed. The body's inertia plays its role by determining the acceleration generated by an impressed force. Maxwell summed up the point nicely, albeit at a much later date: "a body requires a certain force to produce in it a certain change of motion, which fact we express by saying that the body has a certain measurable mass" (quoted in Jammer 1999: 182).

While these considerations strongly support the view that Newton's physics does not include a force that maintains a constant velocity, there is (as Gabbey emphasizes 1971: 40, 1980: 278) a troubling sentence in D4: "For a body perseveres in any new state solely by the force of inertia." Given its context, it is hard to read this sentence as referring to cases in which a body resists an impressed force. Thus while the preponderance of evidence suggests that Newton arrived at a fully inertial physics, a shadow of a doubt remains. We are, however, dealing with a period in which physical concepts (and language) were in flux. Even if Newton did not himself achieve a fully inertial physics, his followers arrived at this result quite rapidly. I have been quoting from the third edition of *Principia*, published in 1726. D'Alembert, in a book published in 1743, wrote:

> When we speak of the "force of a body in motion" either we form no clear idea of what this expression means or we understand by it only the property which moving bodies have of overcoming obstacles encountered in their paths or of resisting them.
>
> (quoted in Jammer 1999: 11)

A remark by Maclaurin, in a book published in 1748, is of special interest since Maclaurin knew Newton personally (Westfall 1983: 830–31) and claims to be presenting Newton's views. Commenting on Newton's first law of motion Maclaurin writes:

> As body, therefore, is passive, in receiving its motion and the direction of its motion, so it retains them or perseveres in them, without any change, till it be acted upon by something external. . . . *From this law it appears, why we enquire not, in philosophy, concerning the cause of the continuation of the rest of bodies, or of their uniform motion in a right line* [italics mine]. But if a motion begin, or if a motion already

> produced is either accelerated or retarded, or if the direction of the motion is altered, an enquiry into the power or cause that produces this change is a proper subject of philosophy. . . .
>
> (Maclaurin 1968: 114)

Gabbey cites another of Maclaurin's remarks as evidence that the older notion of force is retained even here (1971: 42–43, 1980: 279–80):

> Body not only never changes its state of itself, in consequence of its passive nature or *inertia*, but it also resists when any such change is produced: when at rest, it is not put in motion without difficulty; and when in motion, it requires a certain force to stop it. This force with which it endeavors to persevere in its state, and resists any change, is called its *vis inertiae*. . . .
>
> (Maclaurin 1968: 99)

However, the term "persevere in its state" is subject to the same ambiguity noted above, and the context makes it clear that Maclaurin is discussing cases in which a body's state of motion is being changed. Here, as in the case of Newton's texts, older language is preserved while the underlying concepts change.

For my purposes in this book it is not necessary to come to a definitive conclusion on Newton's understanding of force. It is sufficient to note that during Newton's life, and in Newton's own mind, the concept was undergoing change. A fully inertial physics was certainly achieved by the middle of the eighteenth century. In this physics force is proportional acceleration – which includes both change of speed and change of direction – and an object moving at constant velocity requires no sustaining force.

The most important force in *Principia* is gravitational attraction, although Newton explicitly avoids commitment to any physical account of its nature:

> I use the word "attraction" here in a general sense for any endeavor whatever of bodies to approach one another, whether that endeavor occurs as a result of the action of the bodies either drawn toward one another or acting on one another by means of spirits emitted or whether it arises from the action of aether or of air or any medium whatsoever – whether corporeal or incorporeal – in any way impelling toward one another the bodies floating therein.
>
> (588)

The same approach is underlined in Query 31 of *Opticks* (1952: 376), where Newton cites electricity and magnetism as other familiar attractions in nature, and suggests that there may be many more, but refuses to consider what causes these attractions. In *Principia* Newton mentions these other forces in his "Author's Preface to the Reader":

> For many things lead me to have a suspicion that all phenomena may depend on certain forces by which the particles of bodies, by causes not yet known, either are impelled toward one another and cohere in regular figures, or are repelled from one another and recede. Since these forces are unknown, philosophers have hitherto made trial of nature in vain. But I hope that the principles set down here will shed some light on either this mode of philosophizing or some truer one.
>
> (382–83)

Jammer (1999: 202–3) notes that Samuel Clarke adopts the same approach in his debate with Leibniz. In *Principia*, then, gravitational attraction is a tendency of material objects to move towards each other. Newton argues that this is a universal tendency: every bit of matter attracts, and is attracted by, every other bit of matter in the universe. Justification of this claim is a central theme in *Principia*, and exploration of this justification will bring out the content of the concept.

Newton begins his detailed study of motion under the control of a centripetal force in Book I, Sec. 2. At this stage Newton is examining motion under control of a force directed to a point; it is not assumed that there is a body at that point. I.1 establishes that when the center of force is stationary, radii from a moving body to this center pass through equal areas in equal times; I.2 establishes the controverse. I.3 extends I.2 to cases in which the center of force is moving (e.g., looking ahead, the motion of a moon around a planet). I.4 concerns bodies moving in uniform circular motion; Newton proves that the centripetal force responsible for this motion is directed toward the center of the circle, and proportional to (in modern terminology) the square of the angular velocity divided by the radius. This is followed by a series of corollaries in which Newton considers the forces required for different ratios of the orbital period to the radius – all dealing with circular motion. Among these, Cor. 6 deals with a period that increases as the 3/2 power of the radius; Newton proves that the centripetal force varies inversely as the square of the radius. In a scholium following these corollaries Newton notes that this is the relation "for the heavenly bodies" (452) – a rare departure from the strictly mathematical character of Book I. Beginning with I.6 Newton addresses a variety of problems in which we are given a type of motion controlled by a centripetal force, and seek a mathematical description of the force. The exploration is quite general. Newton considers, among other cases, motion on a circle with the force directed to any point within the circle, and with the force directed to a point on the circumference; motion on an ellipse with the force directed to the center of the ellipse and to a focus; and motion on a particular kind of spiral (logarithmic). This is known as the *direct problem*. The *inverse problem* – finding the orbit given the force – is more difficult. Newton gives its solution for the case of an inverse square force in I.13 Cor. 1.[46] There is much more in Book I; it contains 98 propositions and many scholia, but we have what we need for present purposes.

Book III begins with four "Rules for the Study of Natural Philosophy" (794–95); it will be useful to have these before us.

> *No more causes of natural things should be admitted than are both true and sufficient to explain their phenomena.*
>
> (NP1)

> *Therefore, the causes assigned to natural effects of the same kind must be, so far as possible, the same.*
>
> (NP2)

> *Those qualities of bodies that cannot be intended and remitted [i.e. qualities that cannot be increased and diminished] and that belong to all bodies on which experiments can be made should be taken as qualities of all bodies universally* (material in brackets added by the translators).
>
> (NP3)

> *In experimental philosophy, propositions gathered from phenomena by induction should be considered either exactly or very nearly true notwithstanding any contrary hypotheses, until yet other phenomena make such propositions either more exact or liable to exceptions.*
>
> (NP4)

These are followed by six "phenomena" – empirical *generalizations* about major constituents of the solar system that will provide the basis for Newton's account of the world. The first two phenomena report results by named astronomers verifying two key generalizations for the motions of the known moons of Jupiter (Ph1) and Saturn (Ph2): Radii from these moons to the respective planets pass through equal areas in equal times; and there is a 3/2-power ratio between orbital period and radius for these moons. Ph3 states that the five primary planets (Mercury, Venus, Mars, Jupiter, and Saturn) encircle the sun.[47] Ph5 tells us that radii from these planets to the sun pass through equal areas in equal times, while this does not hold for radii from the planets to the earth. Ph4 reads: "The periodic times of the five primary planets and of either the sun about the earth or earth about the sun – the fixed stars being at rest – are as the 3/2-powers of their mean distances from the sun" (800). Given Ph3 and Ph5, Newton's (temporary) agnosticism about whether the earth or sun moves is between the Copernican and Tychonic systems; Aristotelian-Ptolemaic astronomy, like Aristotelian physics, is not in play.[48] Ph6 states that a radius from the moon to the center of the earth passes through equal areas in equal times. These phenomena, along with mathematical results from previous books and the rules for natural philosophy, provide the basis of Newton's argument for universal gravitation. I want to sketch Newton's main steps to this conclusion. (For detailed analyses see Densmore 1996: 285–395; Harper 2002; Stein 1991.)

According to L1, objects with no forces acting on them move in straight lines, thus some forces act on the moons and planets. Proposition III.1 establishes two results: that Jupiter's moons are deflected from straight-line paths by a force directed to the center of Jupiter; and that this force varies inversely as the square of a moon's distance. The first result follows from the first part of Ph1 (equal areas in equal times) plus either I.2 or I.3. The second result follows from the relation between period and time for Jupiter's moons plus I.4 Cor. 6. Newton adds that the same results follow for Saturn's moons, but with the force directed to the center of Saturn. III.2 establishes analogous results for the primary planets with forces directed to the center of the sun.

III.3 deals with our moon. Ph6 tells us that a radius from the moon to the earth describes equal areas in equal times, thus there is a force directed to the center of the earth. But there is only one moon, so Newton cannot compare orbital periods to establish that this is an inverse-square force. Instead, he argues, this result follows from "the very slow motion of the moon's apogee" (802).[49] This is justified by I.45 Cor. 1 which shows that if the apogee of a planet's orbit does not move, an inverse-square law follows. But there is a complication because there is in fact a small motion of the moon's apogee. As the moon moves from east to west, the apogee at each orbit is three degrees, three minutes further east than on the previous orbit. Newton argues that this is small enough to be ignored in the present context (less than 1 percent); I will consider the role of approximations in *Principia* shortly. Newton also notes that he will show later in the text that this movement of the apogee results from the action of the sun.

According to III.4, "*The moon gravitates toward the earth and by the force of gravity is always drawn back from rectilinear motion and kept in its orbit*" (803). By "gravity" Newton means whatever force is responsible for the familiar fall of everyday objects near the earth; Newton's claim is that the same force (whatever its ultimate source) is also responsible for the moon's deviation from a straight line. His argument for this proposition is complex and I will give only its bare structure. (For detailed discussion see Densmore 1996: 294–309.) Newton begins by introducing three empirical results that he will need for the argument. First, he surveys values for the mean distance from the earth to the moon given by several astronomers; these range from 59 to 60 2/5 earth radii. Tycho's value, he notes, is considerably lower (56 1/2 earth radii) and he explains why he thinks Tycho is mistaken. He assumes a value of 60 earth radii (which simplifies the calculations below). Then he adopts values for the time it takes for the moon to complete one orbit and for the circumference of the earth:

> a revolution of the moon with respect to the fixed stars is completed in 27 days, 7 hours, 43 minutes, as has been established by astronomers; and that the circumference of the earth is 123,249,600 Paris feet, according to measurements made by the French.
>
> (804)

In Book I (either I.4 Cor. 9 or I.36 will suffice for the present calculation) Newton established a relation between the distance a body controlled by any centripetal force moves along the arc of a circle in a given time, and the distance it falls as a result of that force in the same time. If we imagine that the moon has no tangential velocity, but is just falling under control of this force, it will fall 15 1/12 Paris feet in one minute. Note especially that an inverse-square law has *not* been assumed at this point. Newton now calculates that under the control of an inverse-square force, at the surface of the earth (i.e., at a distance of one earth radius from the center) the moon would fall approximately this distance in one second. More precisely, in one second the moon would fall "15 feet, 1 inch, and 1 4/12 lines" (804, a line is 1/12 of an inch). Newton then invokes Huygens' pendulum measurements to conclude that an ordinary body falls "15 Paris feet, 1 inch, 1 7/9 lines" in one second;

> therefore that force by which the moon is kept in its orbit, in descending from the moon's orbit to the surface of the earth, comes out equal to the force of gravity here on earth, and so (by rules 1 and 2) is that very force which we generally call gravity.
>
> (804)

Note how this proof integrates mathematical results, empirical data, and rules for natural philosophy. Without NP1 and NP2 one could still maintain that two different causes are responsible for the motion of the moon and for terrestrial fall. It is these rules that take us from a shared property to a common cause.[50]

III.5 states that the deviations from straight-line motion of the moons of Jupiter and Saturn, and of the primary planets, are also caused by gravity directed towards their respective centers of motion. These are all

> phenomena of the same kind as the revolution of the moon about the earth, and therefore (by rule 2) depend on causes of the same kind, especially since it has been proved that the forces on which those revolutions depend are directed towards the centers of Jupiter, Saturn, and the sun, and decrease according to the same ratio and law (in receding from Jupiter, Saturn, and the sun) as the force of gravity (in receding from the earth).
>
> (806)

Newton adds three corollaries to this proposition. First, "there is gravity toward all the planets universally" (802) since all the bodies in question are of the same kind. He also invokes L3 to conclude that these attractions are mutual. Second, the gravity directed toward each of these bodies conforms to the inverse-square law. Third (on the basis of the previous two corollaries), all these bodies

> are heavy toward one another. . . . And hence Jupiter and Saturn near conjunction, by attracting each other, sensibly perturb each other's motions, the sun perturbs the lunar motions, and the sun and moon perturb our sea, as will be explained in what follows.
>
> (806)

Newton adds a brief scholium in which he announces that the above proofs justify calling the centripetal forces that keep all these bodies in their orbits "gravity": "For the cause of the centripetal force by which the moon is kept in its orbit ought to be extended to all planets, by rules 1, 2, and 4" (806). At this point we have mutual attraction among celestial bodies, but we do not have universal mutual attraction of all matter. Three more propositions are required to establish this result.

III.6 establishes that all bodies gravitate to every planet (Newton includes satellites as secondary planets), and that at a given distance from a planet the gravitational force is proportional to the mass of the body in question. Newton's pendulum experiment comparing the gravitational behavior of different materials is reported in the proof of III.6. In a series of corollaries Newton argues that weight does not depend on a body's shape or texture; that all bodies near the earth are heavy in this sense; that bodies are not equally full of matter, that a vacuum exists, and that gravity is a different force than magnetism since the latter does not depend on mass. Several of these corollaries are aimed directly at Descartes, who is mentioned in the discussion of Cor. 3 – which also contains the only explicit reference to Aristotle in *Principia* (Cohen 1999: 203).

III.7 concludes that "*Gravity exists in all bodies universally and is proportional to the quantity of matter in each*" (810). This is followed by Cor. 1: "Therefore the gravity toward the whole planet arises from and is compounded of the gravity toward the individual parts" (811). Newton addresses an important objection in the justification for this corollary:

> If anyone objects that by this law all bodies on our earth would have to gravitate toward one another, even though gravity of this kind is by no means detected by our senses, my answer is that gravity towards these bodies is far smaller than what our senses could detect, since such gravity is to the gravity toward the whole earth as [the quantity of matter in each of] these bodies to the [quantity of matter in the] whole earth.
>
> (811, material in brackets added by the translators.)

Cor. 2 adds that these attractions conform to the inverse-square law.

Finally, III.8 states that the attractions of two homogeneous globes can be treated as if located at their centers – even though it arises from the joint actions of all of the particles of each body. This proposition is central to the application of Newton's results since it allows us to treat astronomical

bodies (among others) as masses concentrated at a point – at least to a close approximation. Proving this result was a major step on the way to writing *Principia*:

> After I had found that gravity toward a whole planet arises of and is compounded of the gravities toward the parts and that toward each of the individual parts it is inversely proportional to the squares of the distances from the parts, I was still not certain whether that proportion of the inverse square obtained exactly in a total force compounded of a number of forces, or only nearly so. For it could happen that a proportion which holds exactly enough at very great distances might be markedly in error near the surface of the planet, because there the distances of the particles may be unequal and their situations dissimilar. But at length, by means of book 1, props. 75 and 76 and their corollaries, I discerned the truth of the propositions dealt with here.
>
> (811)

The remainder of Book III includes application of these results to other problems, including the motions of comets and the tides. Stein notes (1991: 219–20) that these provide further support for the thesis of universal gravitation. We are now ready to explore Newton's concept of gravitation from the perspective of TC.

The systemic roles of GRAVITATION are clear enough. Newton presents gravitation as a descriptive concept but it is also a central explanatory concept. Newton's main concern in *Principia* is with astronomy where his problem situation is set by the evidence in conjunction with L1. Planets and satellites move in curved paths while L1 requires a force that is responsible for every deviation from a straight line. Gravitation explains these deviations. But L1 does not dictate that the same kind of force occurs in all cases, or that the forces between celestial objects have anything in common with forces we experience on earth. So a major part of Newton's explanatory accomplishment consists of bringing these diverse phenomena under a single concept. The desirability of doing so is enshrined in NP1 and NP2.

GRAVITATION implies a mutual attraction between any two pieces of matter that is proportional to their masses and inversely proportional to the square of the distance between them. These are the only properties of bodies that are relevant. Sellars notes that inferences we do not make are as significant as those we do make (Sec. 4.4); *a fortiori* implications that are explicitly blocked tell us as much about conceptual content as implications that are licensed by a concept. This is especially striking when we consider an historical context in which a new conceptual system blocks inferences that were previously licensed. Thus we have seen Newton emphasize that shape and texture are *not relevant* to an object's weight, and that weight is independent of the particular material constituting a body. There is also a mutual implication between GRAVITATION and MASS, along with the distinction

between mass and weight, and the further implications that were discussed above.

Now consider the ICs for GRAVITATION. These are complex because Newtonian theory requires that this concept is instantiated everywhere, and this requires different criteria in different places. Let us begin with the situation described in Newton's first three phenomena: multiple celestial bodies moving in closed paths around a single object. In this case we can compare the distances of the orbiting objects from the central object, and the periods of the various orbits. If the periods vary as the 3/2 power of the distance, we have sufficient grounds for concluding that these bodies are moving under the influence of gravitation. This conclusion depends on two of Newton's results since the 3/2-power ratio establishes only an inverse-square relation. But in III.6 Newton shows that all such motion is proportional to the body's mass. Establishing the 3/2-power ratio requires collecting data on the body's period – data that must be collected over time; this is not the kind of parameter that can be assessed at a glance or by a single measurement. The same holds for determination of distance since it is actually the mean distance that is required. These measurements could yield different results from those required by Newtonian gravitation, so they are indeed criteria for determining if GRAVITATIONAL FORCE is instantiated in these cases.

Our moon presents a different situation since it is a unique orbiting object so comparisons of periods and distances are not available. As we saw in our sketch of Newton's argument for universal gravitation, the key criterion is the motion of a point on the moon's orbit, such as the apogee. A stationary apogee is sufficient to establish an inverse-square force between the orbiting body and the central body. Again, III.6 establishes the dependence on mass. This condition will suffice for any sufficiently isolated orbiting body.

Continuing with astronomical cases, we still need ICs for interactions between bodies that are not orbiting each other, and for points in space that are far from any bodies. If an isolated object is stationary or moving in a straight line, we may conclude that no (net) gravitational force is present. Accelerated motion implies a net gravitational force and generates the problem of finding the specific forces acting. This will typically be done by calculations of the effects of other bodies in the neighborhood. Failure to establish such forces can challenge the theory. For a point in space where no bodies are present, the accelerative measure of force provides an in-principle IC: we can determine whether a gravitational force is present by determining whether test bodies of different masses exhibit the same acceleration. In the absence of actual bodies there is no way to test if a gravitational force is present at a particular location, although astronomical events such as the passage of a comet can provide a test. In addition, recent technology allows us to introduce bodies at some points in space where they do not otherwise occur.

For places near the earth, acceleration measurements of bodies with different masses again provide an IC for the presence of a mass-dependent

force. Measurements at different altitudes provide an IC for the inverse-square variation. While this requires measurements of fairly high precision, they were already possible in Newton's day using the pendulum. Parallel points hold near other astronomical bodies. In our day spacecraft have been sent to several bodies in the solar system; comparisons of their actual and predicted behavior provide such tests. A similar point applies to the behavior of astronauts and other objects on the moon.

It is an important feature of these ICs that the tests all involve approximations. This is a direct consequence of the fact that gravitation is a quantitative concept so that determining if it is instantiated in a particular circumstance requires measurements that check quantitative values determined from theory. But measurements always involve a range of possible error and thus automatically yield approximate results. Newtonian theory introduces an additional reason why measurements of gravitational effects must be approximations. Given that gravitation is universal, every celestial object and every point in space is gravitationally affected by many objects, so that it is impossible to measure the gravitational effect of a single body. We can improve approximations by seeking relatively isolated bodies so that (according to theory) their behavior will be dominated by a single gravitational force, but this will not eliminate approximations completely. Indeed, there is a trade-off between the precision of a measurement and the ability to ignore small effects: a body that can be treated as isolated for purposes of relatively crude measurements, cannot be treated as isolated in more refined measurements.

Additional complications arise for terrestrial objects. Motions of projectiles are affected by the air which introduces an additional force that depends on velocity, and also yields effects due to an object's shape. I have already noted some factors involved in measuring an object's weight at different places on the earth. These are examples of the kinds of factors that must be taken into account when assessing whether a quantitative concept is instantiated.

Theoretical predictions also involve approximations. In general, quantitative predictions do not come from theory alone; they require empirical inputs. For example, to determine the orbit of Mars we must know at least its mass, the mass of the sun, the distance between them, and the planet's orbital speed; a more precise calculation will require data on other planets that affect the orbit. To calculate the planet's location at a particular time we must determine its location at some other time. All these empirical inputs result from measurements that are only approximately correct; they thus yield only approximately correct predictions. Back on earth, when Newton calculated the difference between the gravitational force at the equator and the poles (at the surface of the earth) he assumed a uniform density for the earth; more realistic data would be difficult to acquire, would depend on the state of the relevant technology, and involve multiple approximations.

Another source of approximation in theoretical calculations comes from mathematical difficulties. For example, testing the theory's conclusions about

a planet's orbit requires that the orbit be calculated. Newton had to introduce a series of approximations in order to carry out these calculations. (Newton's assumption that the earth has uniform density also serves to simplify calculations.) Moreover, for reasons that became fully clear only after substantial further developments in mathematics over the next two centuries, it is in general not possible to calculate orbits precisely when three or more bodies interact in accordance with Newton's gravitation law (exact solutions are possible for some special cases). Successive approximations can yield highly accurate results, but it is a consequence of the mathematics at the foundation of Newtonian dynamics that, for the most part, only approximate results can be derived from the theory.

I hasten to add that this reliance on approximations is *not* offered as a criticism. Quantitative theories generally carry more information and are subjected to more demanding tests than qualitative theories (cf. Popper 1992). Reliance on approximations in prediction and measurement is a fact of life of quantitative science which has generated an interplay of progressively more precise predictions and measuring techniques; these place greater demands on theory. Recall that a major piece of evidence leading to the replacement of Newtonian gravitational theory by general relativity was a shift in the perihelion of Mercury of 43 seconds of arc per century that could not be accounted for by Newtonian theory.

Newton addresses issues raised by approximations repeatedly in *Principia*. His discussion of weight measurement at different latitudes is one example; I will mention a few more cases by way of illustration. In a scholium that follows his statement of the laws of motion Newton considers using pendulum experiments to test consequences of these laws and their corollaries for colliding bodies. "However," he writes, "if this experiment is to agree precisely with these theories, account must be taken of both the resistance of the air and the elastic force of the colliding bodies" (425). Newton then describes experiments that will yield estimates of these factors. In I.31 Newton considers how to find the position at a given time of a point moving on an ellipse. After solving the problem using a complex geometrical argument Newton adds a scholium which begins, "But the description of this curve is difficult; hence it is preferable to use a solution that is approximately true" (514), and develops an alternative approach. In Book II, after proving some lemmas concerning motion of convex bodies through a fluid, Newton lists his simplifying assumptions:

> we are supposing that the bodies are very smooth, that the tenacity and friction of the medium are nil, and that the parts of the fluid which by their oblique and superfluous motions can perturb, impede, and retard the flow of the water through the channel are at rest with respect to one another as if icebound and adhere to the front and back of the bodies. . . .
>
> (747)

Turning to Book III, recall how Newton's proof that the earth exerts an inverse-square force on the moon depends on the moon's apogee being stationary – which is only approximately correct. Newton notes that we might account for the small motion of the apogee by adjusting the gravitational force, but (by I.45 Cor. 1) this requires a force that varies as the 2 4/243 power of the distance. Instead he argues that the motion is caused by the action of the sun in accordance with the inverse-square law. There are many other examples throughout *Principia*. Newton is also careful about the use of approximate inputs to theoretical arguments:

> in every case in which he deduces some feature of celestial gravitational forces, he has taken the trouble in Book I to prove that the consequent of the "if-then" proposition licensing the deduction still holds *quam proxime* so long as the antecedent holds *quam proxime*. For instance, two corollaries of Proposition 3 show that the force on the orbital body is at least very nearly centripetal so long as the areas swept out in equal times remain very nearly equal.
>
> (Smith 2002: 156)

At several points Newton introduces techniques that allow one to deal with certain small gravitational interactions – such as the effects of planets on the orbits of other planets – as modifications to larger interactions. (See Nauenberg 2001: 189–93 for a summary.) This is the beginning of *perturbation theory*, a technique that remains central to physics all the way down to the quantum level.

More generally, Newton regularly follows a method of successive approximations that takes him from consideration of relatively simple situations to those of greater complexity.[51] For example, in developing his mathematical analyses Newton moves from considering the motion of a mass point in a force field directed towards a point in space, to an actual body (approximated by a perfect sphere) moving in such a field, to a two-body problem in which the force is generated by a second body, to multiple-body problems. He also moves from motion in a non-resisting medium to consideration of the effects of resistance. In Book III, where he is dealing with the actual planets and integrating his mathematical results with astronomical data, Newton pays special attention to hard cases where approximations he has introduced break down – such as the orbit of the moon, and the interaction between Jupiter and Saturn when they are close to each other. In a similar way, Newton first treats each planet as a body moving in the sun's gravitational field with the sun at a focus of the ellipse, and then moves to treating a planet and the sun as interacting bodies, with the result that neither body strictly moves around the other. As we have seen, he postpones a decision between the Copernican and Tychonic planetary systems until he can show that neither is strictly correct, but that the Copernican account provides a better approximation.

Returning to TC, I want to consider whether this reliance on approximations in the ICs for Newton's key concepts implies that these are in some sense approximate concepts. It may be useful to recall here why I have included ICs in the conceptual content of descriptive concepts. The guiding idea is that a descriptive concept is incomplete without criteria for determining if it has instances. The question now is whether a limitation to approximate tests implies an approximate element in the content of these concepts. Could we hold instead that the ICs are exact, even though our means of applying them are approximate? I do not want to rule out this possibility in general, but whether it is appropriate depends on the specific case. Recall the top quark. In this case the ICs are unavoidably statistical because we have no conception of a non-statistical procedure that would serve. To be sure, later developments might lead to such a procedure, but that would involve some degree of conceptual change. In the Newtonian case the reasons why we can have only approximate instantiation tests lie deep in the conceptual system; they derive from the general nature of the measurements required to test quantitative theories and from the Newtonian mathematical structure. As a result, I urge that we recognize an approximate *element* in the content of these concepts. Concepts are complex items and it is illuminating to note such an aspect; there is no additional gain in trying to decide if we should apply this label to the concepts *tout court*.

I want to consider one more set of concepts that Newton considered of central importance: ABSOLUTE SPACE, ABSOLUTE TIME, and ABSOLUTE MOTION. Consider time first: "Absolute, true, and mathematical time," Newton tells us, "in and of itself and of its own nature, without reference to anything external, flows uniformly and by another name is called duration" (408). This is contrasted with time as we normally measure it, which yields only "relative, apparent, and common time. . . ." Time is measured by uniform motion, and Newton is skeptical about our ability to measure absolute time: "It is possible that there is no uniform motion by which time may have an exact measure. All motions can be accelerated and retarded, but the flow of absolute time cannot be changed" (410). Newton is more optimistic about our ability to recognize absolute space because it has a special relation to absolute motion, which in turn is intimately related to the forces acting on a body. "Absolute space, of its own nature without reference to anything external, always remains homogeneous and immovable. Relative space is any movable measure of this absolute space . . ." (408–9). A body may have numerous and various relative motions, but its true motion is its motion with respect to absolute space. Only an impressed force can alter true motion:

> The causes which distinguish true motions from relative motions are the forces impressed upon bodies to generate motion. True motion is neither generated nor changed except by forces impressed upon the moving

> body itself, but relative motion can be generated or changed without the impression of forces upon the body.
>
> (412)

If a force acting on body A changes A's motion, the motions of all other bodies relative to A are automatically changed, but only A's true motion is changed. As a result of this connection with forces, absolute space is more central for Newton than absolute time.[52] Given the mutual implication between FORCE and ACCELERATION, Newton believed that acceleration could reveal the presence of absolute motion.

Newton provides two experiments – one actual, one a thought experiment – to demonstrate the existence of absolute motion. In both cases absolute motion is detected by the presence of a force, and both involve circular motion, so Newton's assimilation of circular motion to accelerated motion is vital. The actual experiment concerns a bucket of water hanging from a cord. The bucket is turned (say, clockwise) and the cord is twisted as a result. If the bucket is then given a counterclockwise spin, the twisted cord will keep it in motion for a time. At the beginning of the bucket's spin the motion has not yet been communicated to the water, which is thus in motion relative to the bucket. As the water begins to spin the shape of its surface changes. The water rises up the sides of the bucket and the water's surface becomes concave. Soon there is no relative motion between the water and the bucket. But the concave shape of the surface indicates the existence of a force:

> The rise of the water reveals its endeavor to recede from the axis of motion, and from such an endeavor one can find out and measure the true and absolute circular motion of the water, which here is the direct opposite of its relative motion.
>
> (413)

The thought experiment concerns two identical balls of matter connected by a rope and far from any other material objects which could indicate relative motions. The arrangement is symmetric around an axis through the center of gravity of the balls. If the object spins around this axis there will be a tension in the cord – which indicates that motion is occurring – and this tension will change as the rotational speed varies. There is, again, a direct tie between rotational motion and force.

The relation between force and absolute motion plays a key role in Newton's critique of Descartes' physics in *De Gravitatione*. Descartes, we have seen, also attempts to define a sense in which there is a matter-of-fact as to whether a body is moving. Newton notes this Cartesian claim (1962: 123) but argues that (among other defects) Descartes' account of motion implies

a relativist account. In this discussion Newton takes it as fundamental that there must be a determinate fact as to whether a body is moving.

ABSOLUTE MOTION is, then, a central concept in Newton's framework, where its function is to mark the distinction between true and apparent motions. In principle this distinction applies to every motion, but Newton is able to provide an IC only in cases where acceleration occurs. Since ACCELERATION implies FORCE, and FORCE implies ABSOLUTE MOTION, we would seem to have established a case of absolute motion whenever we detect acceleration. Exactly how we detect acceleration can vary. In the bucket experiment the concavity of the water's surface indicates the presence of a force; in the case of the globes the tension in the cord is our indicator. But other cases can be tricky. Suppose I push with a steady force in a constant direction on a box that is initially stationary relative to the immediate environment. I can feel the force that I am applying, and the speed of the box changes; together these clearly indicate that the box, rather than the environment, is moving. But frictional forces oppose the motion, and these forces vary with speed. After a while friction balances the force I am applying and – since only the net force on the box is relevant – the box moves at a constant speed. So the fact that I feel myself pushing only guarantees acceleration in the absence of other forces. Yet once a steady state is reached it does not follow that the box is moving absolutely since my push may have stopped a previous absolute motion. The important point is that detecting a specific force is not sufficient for identifying absolute motion since other forces may be at work and only the net force counts. In addition, as long as the box is changing speed relative to its environment, that environment is also changing speed relative to the box. Thus detecting a change of speed is not sufficient for concluding that absolute motion is occurring. It seems that in this case *both* evidence of the presence of a force and evidence of acceleration are required to conclude that absolute motion is occurring.

Once Newton establishes the universal role of gravitational attraction both of these criteria are met by the motions of celestial bodies: they are all responding to gravitational forces, and their motions in curved paths provide evidence that they are accelerated. In this case we have an adequate IC for ABSOLUTE MOTION.[53] In general, however, Newton does not provide adequate means of recognizing whether absolute motion is occurring. According to TC it follows that he has not provided an adequately developed concept of absolute motion. Many natural philosophers of his day and afterward reject any role for this concept in physics; many arguments against the inclusion of this concept amount to denying that it plays any empirical role in physics – which is implied by the lack of an adequate IC. At best Newton provides ICs for curved motions around a central point, and for a limited range of linear accelerated motions. He provides no IC for non-accelerated absolute motions or for many cases of linear accelerated motions. We have also noted that he despairs of any IC for absolute time.

## 9.5 Conclusion

In this chapter I examined three cases in which introduction of a new conceptual system for physics involves systematic changes in an available system: the Galilean and Cartesian systems which were developed by modification of Aristotle's framework, and Newton's system, which began from a modification of the Cartesian framework. Within the limits of the history I have explored, no case occurs in which there is a transition from Galileo's framework to a new framework. Descartes is largely dismissive of Galileo's work (see, for example, de Gandt 1995: 118–20), while Newton absorbs and uses what he considers to be Galileo's main results. I am *not* claiming that these conceptual modifications are the only thing involved in the introduction of a new framework. It is clear that new experiments and observations played a central role; in Newton's case new mathematics is also crucial. Perhaps other factors are also involved. But the examples do illustrate the aspect of scientific development that I am concerned with in this book: that introduction of fundamentally new concepts does not require that one ignore existing concepts and begin anew. According to TC a conceptual system is a rather complex item. Each of the three dimensions is itself complex, permitting changes to some aspects of a dimension while leaving others intact. For example, one can add or drop some implications while leaving others largely unaffected; or one can introduce a new systemic role while leaving some of the earlier roles intact. In addition, changes may be more drastic on one dimension than on the others. As a result, there is no incompatibility between continuity and innovation, and little point in attempting to stick one of these labels on a given transformation. Throughout the chapter I have illustrated how TC can guide the systematic exploration of ways in which a new system introduces innovations while maintaining continuity with its predecessors. I continue to postpone general discussion of the significance of this kind of change for the development of science until the final chapter.

I want to end the chapter by highlighting one outcome of the developments we have been considering: a central role for mathematics in physics was advocated throughout the seventeenth century; mathematics actually came to play that role by the end of the century. As a result, quantitative considerations – which once were central only in astronomy – pervade all physical science. These developments were possible because both mathematics and the collection of empirical data through observation and experiment achieved levels of power and sophistication well beyond anything previously available. This move to quantitative physics brings along a recognition of the role of approximations, and reflection on their use. For Galileo approximations arise because physical objects are never exact instances of geometrical concepts, although he considers other issues as well. But in Newton's hands the dependence on approximations, and the understanding of how they are to be handled, reach new heights. More powerful

mathematics, improved evidence collection, and growing sophistication in the understanding and use of approximations all play a central role in the development of physics over the following centuries. All are taken for granted in the fragment of late twentieth-century physics that I discuss in the next chapter.

# 10 Historical Studies II: Interactions

> Experimental science is continually revealing to us new features of natural processes and we are thus compelled to search for new forms of thought appropriate to these features.
>
> (J. C. Maxwell, quoted in Pais 1986: 454)

Before the twentieth century physicists recognized two fundamental forces, gravitation and electromagnetism, but the discovery of radioactivity followed by the development of atomic and subatomic physics led them to recognize two further interactions: the *strong force* that holds the nucleus of atoms together, and the *weak force* that is responsible for a variety of particle decays and other phenomena. Quantum theory led to a major rethinking of the nature of physical interactions, and by the end of the twentieth century there were well-developed quantum theories of the strong, weak, and electromagnetic interactions, along with an ongoing project of integrating gravitation into this framework. Leaving quantum gravity aside, I will examine the unified account of the other three interactions – known as the *standard model* (SM) – as it existed at the end of the twentieth century. We will see that this model takes us far from the everyday notion of a force. Because of this I will generally talk about *interactions* rather than forces. Many physicists adopt this language although the language of forces has not vanished and I will not avoid it altogether.

Sec. 10.1 provides a qualitative – almost visualizable – picture of these interactions as they are understood in SM. This part of the discussion should be accessible to non-mathematical readers, but is potentially misleading on its own. Since Newton the dominant mode of physical theorizing has been mathematical; Sec. 10.2 gives an outline of the mathematical structure of SM. This will be far from an account that a physicist would consider adequate, but I attempt to bring out enough of the characteristic mathematical features of SM to indicate the enormous amount of conceptual development that was required to get us to this theory. Some of the mathematical concepts used in this account are explained in the Appendix; references of the form *An* refer to sections of this appendix. Throughout the

discussion in these two sections I focus on the construction of new theories by analogy with established theories. TC provides the background for this discussion, although there will be few explicit references to it. I then return to the explicit use of TC to discuss three final examples of conceptual change. Sec. 10.3 examines the introduction of a new concept, isospin, which is central to SM; Sec. 10.4 examines changes in the concept of a force; Sec.10.5 considers changes in what counts as a unification.

In developing this chapter I have relied heavily on two textbooks: Cottingham and Greenwood (1998) and Rolnick (1994); I cite the former as CG, the latter as R. These books are already compressed treatments that leave out many mathematical details.

### *10.1 Qualitative Picture*

We can approach SM by thinking of nature, at the most fundamental level, as consisting of three types of elementary particles (i.e., particles not composed of other particles). Two of these, *leptons* and *quarks* are constituents of ordinary matter; particles of the third type carry the fields by which particles interact.[1] Leaving fields aside for now, the crucial difference between leptons and quarks is that leptons do not respond to the strong interaction (SI); all quarks and leptons respond to the weak interaction (WI). By analogy with ELECTRIC CHARGE, physicists introduce the concepts STRONG CHARGE and WEAK CHARGE. All quarks and leptons have weak charge; quarks have strong charge which leptons lack. In addition, all quarks and half the leptons have electric charge.

SM recognizes six leptons: the electron, muon, and tau, plus a characteristic neutrino associated with each of these. The first three each carry the same negative electric charge so they respond to both WI and the electromagnetic interaction (EI); neutrinos are electrically neutral and respond only to WI. In addition, the charged leptons all have rest mass; in SM neutrinos do not have rest mass and thus move at the speed of light.[2] Leptons are grouped into three sets of two, each set consisting of a charged particle and its associated neutrino; these sets are commonly referred to as *families*.[3] In WI an electron, muon, or tau is always accompanied by its characteristic neutrino.

There are also six known quarks which have been given whimsical names; these are also divided into three families: up, down; charmed, strange; and top, bottom. While all quarks have electric charge, these are fractions of the charge on the electron, $e$, which was long considered to be the minimal charge that occurs in nature. The up, charmed, and top quark each has a charge of $2/3e$; the down; strange; and bottom each has a charge of $-1/3e$. Thus each quark family consists of a quark of charge $2/3e$ and one of charge $-1/3e$. Quarks respond to EI, WI, and SI.

Every fundamental particle has a corresponding anti-particle. A particle and its anti-particle have the same mass, lifetime, and spin (I will discuss

spin shortly), but have opposite values for other properties. Leptons, for example, have a property known as *lepton number*; a particle and its anti-particle have opposite lepton numbers (i.e., their sum is zero). An electrically charged particle and its anti-particle have opposite charge. Neutrinos are electrically neutral; it is unknown whether each neutrino has a distinct anti-particle (Dirac neutrinos) or is identical with its anti-particle (Majorana neutrinos). Each quark has an anti-particle with opposite charge and opposite values of other characteristic quantum numbers (see Perkins 2000: 377–78 for a summary).

Particles that respond to SI are known as *hadrons*. In addition to the quarks there are myriad composite hadrons that are systems of quarks; this class includes neutrons and protons, along with many less familiar particles. The class of leptons includes only fundamental particles. The electric charge of a composite hadron is the arithmetic sum of the charges of its constituent quarks; composite hadrons may be electrically positive, negative, or neutral. Their anti-particles consist of the corresponding anti-quarks. The electrically neutral neutron, for example, consists of one up quark and two down quarks (udd), while an anti-neutron consists of their anti-quarks. An electrically charged composite hadron has the opposite charge from its anti-particle. Composite hadrons further divide into two classes: *baryons* – such as protons and neutrons – are each made up of three quarks; *mesons* – such as the electrically positive, negative, and neutral pions – are each made up of one quark and one anti-quark; the quarks in a meson may, but need not, be each other's anti-particles.

We must consider one more grand division between particles – *fermions* and *bosons* – a difference related to *spin*. It is somewhat helpful to think of each particle as if it were spinning on an axis, although this is an analogy and not exactly correct (see Sec. 10.3 for details). Spin is quantized, and physicists introduce a unit of spin, ħ, equal to Planck's constant (h) divided by $2\pi$. Thus we can express the spin of any particle as a number times ħ. The spin of a fermion is always ħ multiplied by *half an integer* (1/2, 3/2, etc.); the spin of a boson is always ħ multiplied by an *integer*. Particles with zero spin behave as bosons, so we can view these as cases in which ħ is multiplied by the integer zero (recall the discussion of zero in Sec. 2.2). Spin can be either positive or negative. In the spinning-particle analogy we can think of clockwise spin as positive and counterclockwise spin as negative. It is customary to specify just the multiplier when giving a particle's spin, so physicists speak of particles having spin 1/2, 1, -3/2, and so on. All quarks and leptons are fermions. Spins of the quark constituents of composite hadrons add. Thus baryons (three quarks) are fermions; mesons (two quarks) are bosons.

Bosons and fermions exhibit very different behavior. Fermions obey the Pauli exclusion principle: two particles in a system cannot have all of the same quantum properties (degrees of freedom). For example, electrons in an atom can occur with various energies, and more than one electron can have the same energy – provided they differ in other properties such as angular

momentum or spin. However, a limited number of degrees of freedom are associated with each energy level. The combination of a limited number of degrees of freedom plus the exclusion principle generates the organization of the electrons in an atom, which determines the atom's chemical properties. The exclusion principle does not apply to bosons, which tend to move into the same quantum state. This difference with respect to the exclusion principle is tied to another fundamental difference. Quantum theory deals with probabilities. Suppose we are interested in a set of states and wish to determine the probability that a number of particles will be distributed among those states in a particular way. This requires calculating the number of allowable ways in which the particles can be distributed. Given the exclusion principle, there can be either one fermion or none in each state; there is no limitation on the number of bosons that can occur in the same state. Thus the probabilities must be calculated by different rules for fermions and bosons. The two types of particles follow different statistics: Fermi-Dirac statistics for fermions and Bose-Einstein statistics for bosons. This difference in statistics is related to the different spins associated with these particles.

Now consider the fields involved in interactions among quarks and leptons. In SM particles interact by exchanging other particles that serve as mediators of the field. Each type of field is meditated by a characteristic boson; for the three fields I am considering they are all spin-one bosons. These mediators constitute a third class of fundamental particles in addition to quarks and leptons.

EI was the first interaction to be understood in these terms; the theory of this interaction is known as *quantum electrodynamics* (QED). SM accounts of the other interactions have been modeled on QED – with the kinds of variations we have learned to expect when such modeling takes place. EI is mediated by the *photon*, represented by $\gamma$; photons are themselves electrically neutral. The thesis that charged particles interact by exchanging photons was introduced by Bethe and Fermi in 1932 (Schweber 1994: 78). Consider two charged particles, say two electrons: QED tells us that they interact when a photon emitted by one is absorbed by the other. But emission or absorption of a photon involves a change in momentum, and change in momentum with respect to time is force (as in classical mechanics); this is the characteristic force between electrically charged particles. According to *classical* electromagnetic theory charged particles emit photons only when they accelerate; in QED charged particles always emit photons, no matter what their state of motion. We should think of each electron as continually emitting and absorbing photons so that an electron is surrounded by a cloud of photons. These are known as *virtual* photons because they are not directly detectable, but must be taken into account in calculations. In order to understand the idea of a virtual particle we must consider Heisenberg's indeterminacy principle.

The indeterminacy principle tells us that there are certain *specific pairs* of physical properties that cannot *both* be determined simultaneously with

unlimited precision. The greater the precision with which one is determined, the less the precision with which the other can be determined. There is, for example, an indeterminacy relation between location and momentum. Using the symbols $\Delta x$ for the indeterminacy in a particle's location and $\Delta p$ for the indeterminacy in its momentum, the principle tells us that $\Delta x \cdot \Delta p \geqslant \hbar$.[4] There is also an indeterminacy relation between the energy and time involved in a specific interaction; this is the relation that concerns me at the moment. This indeterminacy relation can be interpreted as allowing for the existence of particles of a wide variety of energies – including energies that violate conservation of energy – provided the particles do not exist long enough to be detected. These are virtual particles, the higher a virtual particle's energy, the shorter its lifetime. EI takes place between two charged particles when one absorbs a virtual photon emitted by the other.[5] This picture allows us to understand, in a qualitative way, why EI drops off with distance. Photons have zero rest mass, so they travel at the speed of light (in all reference frames) in a vacuum. This speed is fast, but finite, so the maximum distance a virtual photon can travel is determined by its lifetime. The more energy a virtual photon has, the shorter its lifetime, so those that live long enough to reach distant objects will have low-energy. In principle the field has unlimited range – which is related to the fact that photons have no rest mass. Field carriers with zero rest mass are necessary (but, we will see, not sufficient) for unlimited range.

WI is responsible for a variety of transformations. The earliest evidence for this interaction came from beta radioactivity: emission of electrons from nuclei of atoms. Beta decay posed a substantial problem in the early part of the twentieth-century because energy and angular momentum seemed to be missing; neutrinos were postulated as particles that carried off the missing energy.[6] Neutrinos were originally proposed by Pauli rather tentatively in 1930, and then incorporated into a detailed theory by Fermi in 1933, although that theory has now been superseded. Recall (Sec. 2.1) that early twentieth century physicists believed that the nucleus is composed of electrons and protons: beta decay was thought to involve emission of one of those electrons. Once it was recognized that the nucleus consists of protons and neutrons a new account was required: beta decay occurs when a neutron disintegrates into a proton, an electron, and an anti-neutrino by means of WI.[7] This account does not require that the neutron was composed of an electron and a proton. Mass may transform into energy ($E = mc^2$) which may then transform into mass – often yielding different particles than were present initially. Let us look at how this case is integrated into SM; the key step is to determine the mediators of the interaction.

WI is mediated by three bosons with three different electrical charges (positive, negative, and neutral) symbolized $W^+$, $W^-$, and Z. These field quanta have mass, which accounts for the short-range of WI (approximately $10^{-18}$ meters). The two electrically charged bosons have the same mass; the neutral boson is slightly heavier. While these bosons occur as virtual particles

in WI, like photons, they can also appear as real particles under appropriate conditions and have been detected in experiments at accelerators. Although the SM account of WI is modeled on QED, there are substantial differences between the two theories. The existence of a single, electrically-neutral massless field is not retained in the theory that results from the modeling process. Moreover, since the photon does not have electric charge, there are no electromagnetic interactions between photons; but WI bosons carry weak charge, so there are weak interactions between these bosons. We will encounter further differences when we look at the mathematical structure of these theories.

WI mediates a variety of interactions in which the final particles are different from the initial particles. Many of these transitions can be thought of as decays because a particle is replaced by a set of particles of lower mass, but this is not the only case. For example, inverse beta decay occurs when a proton absorbs an anti-neutrino and emits a neutron and a positron. There are also cases (regularly produced at particle accelerators) in which WI mediates creation of new particles out of energy. Another case occurs when two leptons – say an electron and an electron neutrino – glance off each other in a kind of elastic collision. This interaction cannot be mediated by EI because one of the particles has no electric charge. Rather, the "colliding" particles exchange a Z. When WI involves composite hadrons an account must be given in terms of quarks. For example, beta decay takes place in two steps. First one of the neutron's down quarks transforms into an up (creating the final proton) and a virtual $W^-$. Then the $W^-$ transforms into an electron and an antineutrino. Similar accounts involving virtual W and Z intermediaries apply to other weak interactions.

This discussion of WI has been, in one respect, a bit misleading: I have mixed together ideas from an earlier stage in the understanding of this interaction with ideas from more recent accounts. In particular, I have been writing as if WI is distinct from EI, which was the view before SM was developed. But SM combines WI and EI into a single theory. Discussion of this unified theory is best pursued in its mathematical framework; I return to this topic below. For the moment I want to provide a qualitative picture of the SM account of SI.

In 1935 Yukawa attempted to integrate SI into the framework that was developing at that time. The distinction between the four fundamental interactions was in place, and Fermi had developed his WI theory that included the neutrino – but without any notion of a mediating particle. The notion of a photon has an extensive prehistory in particle theories of light that go back at least as far as Newton; the modern version of the photon was introduced by Einstein in 1905. But in 1935 QED (the model for WI and SI) was more than a decade in the future. In this context Yukawa proposed the existence of a single massive boson that mediates SI, along with the thesis that the field's range is inversely proportional to this mass (Pais 1986: 430).[8] For a time it looked as if this attempt might work, especially after a particle that

seemed to have the appropriate characteristics was discovered in cosmic radiation. But it became clear that this particle, while a boson, is not the required field carrier.[9]

While SI was introduced to account for the nuclear binding force, in SM strong interactions occur directly between quarks and are responsible for binding quarks into hadrons. The force between nuclear constituents is a secondary manifestation of the quark interactions. The bosons that mediate SI are known as *gluons*. These are electrically neutral but carry strong charge and respond to SI. According to SM there are eight *massless* gluons, although SI is a short-range interaction. The short-range is a consequence of *quark confinement*: as two quarks move further apart, the amount of energy required to move them even further apart steadily increases. One consequence of this increasing energy demand is that isolated quarks never appear. Consider how this works in the case of mesons. The amount of energy required to break the bond between the two quarks constituting a meson exceeds the amount needed to produce a new quark-antiquark pair. Thus as more energy is pumped into the system we end up with two mesons, rather than a pair of high-energy isolated quarks.

A number of factors contribute to the experimental determination of the type of interaction involved in a given case. These include the types of particles found in the inputs and outputs, plus the relevant conservation laws since different interactions conform to different conservation laws. One consideration that is often useful in identifying interactions is the time involved: particle decays mediated by SI and EI are generally much faster than those mediated by WI. A particle that decays by SI will typically have a lifetime in the neighborhood of $10^{-23}$ seconds. Weak decays that take as long as $10^{-10}$ seconds are common, and some may take much longer – more than 10 minutes on average for a free neutron. There is, however, considerable variation in decay times as well as variations in the interaction strength. For example, the decreasing strength of SI with shorter distance results in a longer decay time for more massive hadrons that have their quarks initially packed more closely together. In the case of WI there is considerable variation in the interaction strength with energy. Indeed, one basis for the

*Table 10.1* Relative Interaction Strengths

| *Interaction* | *Relative coupling strength* |
|---|---|
| Strong | 1 |
| Electromagnetic | $10^{-2}$ |
| Weak | $10^{-5}$ |
| Gravitational | $10^{-39}$ |

unification of the weak and electromagnetic interactions is that the two forces have the same interaction strength at energies above about 200 GeV.[10] For purposes of comparison we can take the coupling strength of the strong interaction as 1; Table 10.1 then gives the relative strengths of the four interactions for the typical energies at which we live.

## 10.2 Mathematical Framework

In quantum theory physical quantities are represented by operators (*A1–A2*). Consider spin, which is represented by an operator that is formally analogous to the operator for angular momentum. Since spin is always a multiple of a basic unit, it is sufficient to give that multiple to specify the spin; this multiple is an example of a *quantum number*. Other properties that are multiples of a specific unit can also be specified by giving the appropriate number. Originally quantum numbers were used for properties that occur in spacetime, such as energy and angular momentum. But physicists also use quantum numbers to specify properties that have no spacetime interpretation – although they do have consequences for measurements made in spacetime; these are known as *internal quantum numbers*. I want to develop this notion by considering *isospin*, the first case introduced into physics, and a case that is central to the following discussion.

In the 1930s Heisenberg noted that neutrons and protons respond in exactly the same way to the strong force even though their masses are not quite the same and the proton is charged while the neutron is uncharged. He suggested that, from the perspective of the strong force, the two particles can be considered two different states of a single entity – the *nucleon*. We can describe this situation in terms of a fictitious 2D nucleon-space in which one axis is the neutron and one the proton. A nucleon is a vector in this space. A vector that coincides with the proton axis represents a proton; a 90° rotation in this space transforms a proton into a neutron but preserves the shared property that characterizes the strong force. A parallel account applies to neutrons. In general, a vector in this space is a mixed state of the kind that is common in quantum theory. Using $n$ for the neutron axis and $p$ for the proton axis, such a state might be represented by $an + bp$, where "+" indicates vector addition. The numbers $a^2$ and $b^2$ give the probability of finding a neutron or proton, respectively, on measurement.[11] Heisenberg introduced a formalism for describing this situation that is mathematically identical to the formalism for spin. This involved the postulation of a new property, now known as *isospin*; it is characterized by an internal quantum number that is invariant when one state in isospin space is rotated into another (*A3*). Thus isospin is a conserved quantum number (see Sec. 10.3 for further details).

Heisenberg's treatment of neutrons and protons came together with *group theory* (*A4–A5*) in the early 1960s when Gell-Mann and Ne'eman used group theory to classify the various baryons and mesons that had been discovered. Mathematicians have studied and classified different types of

groups with different properties; the new idea entering into physics was that certain groups could provide a basis for organizing these particles into sets, and predicting new particles. Particles that share certain properties can be viewed as axes in an imaginary space, and a particular group of operations can be viewed as rotating vectors in this space. The number of dimensions in this space will correspond to the size of the matrices in an IRR (*A6*) of the group. For example, at the time in question there were eight known baryons that have spin (not isospin) 1/2. The group SU(3) has IRRs of sizes **1**, **3**, **8**, and **10**, and these spin-1/2 baryons can be represented by an 8D space operated on by the **8**. In addition, physicists knew of eight spin-zero mesons that could also be represented by an **8**, and nine spin-one mesons that could be represented by an SU(3) octet plus a singlet. The spin-3/2 baryons provided a striking case. They appear to fit a **10** representation of SU(3), but when this structure was originally proposed only nine such particles had been identified. Application of group theory to these particles thus led to the prediction of a tenth particle, the $\Omega^-$, a prediction that was confirmed in 1964 (see R 106–9 for details that I omit).

The general success of this approach led some to ask if there is a more basic set of particles out of which the newly discovered particles can be constructed. In particular, the fundamental representation of SU(3) (*A6*) is a **3**, which suggested the possibility that the hadrons could be constructed out of three basic particles (R 109–11); these particles are *quarks*. The original version of the theory contained three quarks with the same isospin, providing the axes of a 3D isospin space. Although this approach broke down as more quarks were discovered and other problems arose, SU(3) retains a fundamental role in quark theory. I will return to this topic when we consider SI, but there is other work to be done first. I want to describe the general mathematical approach to quantum field theory (QFT), and then consider the specific interactions.

The physical situation is described by a mathematical expression known as a *Lagrangian density*, symbolized $\mathscr{L}$. There is no automatic procedure for writing down $\mathscr{L}$, but existing physical theory provides guidance for a first approximation. Then a key constraint is applied: it is required that $\mathscr{L}$ be invariant with respect to a specific group of operations (*A3–A4*). Different groups are appropriate for different fields, and the introduction of a particular symmetry group is a testable hypothesis. Typically the $\mathscr{L}$ we begin with will not meet the invariance requirement, but an invariant $\mathscr{L}$ can be constructed if we make a specific *type* of change in the part of $\mathscr{L}$ that describes the interaction; we will examine this change in the following subsections. For historical reasons the change is known as a *gauge transformation*.[12] The gauge transformation determines the form of the interaction term in $\mathscr{L}$ and yields a field that is responsible for the interaction. This field is referred to as a *gauge field*, and the bosons that transmit the field are *gauge bosons*. The details of this procedure are importantly different for each of the three interactions; I will consider them in turn.

### *10.2.1 Electromagnetic Interaction*[13]

We begin with EI which is the easiest case and which provides the model for the other cases. For an electron moving in an electromagnetic field $\mathscr{L}$ is the sum of three parts: $\mathscr{L}_{\text{EM}}$ describes the field; $\mathscr{L}_{\text{DIRAC}}$ describes an electron in the absence of any field; $\mathscr{L}_{\text{INT}}$ describes the interaction between the electron and the field. Beginning with an initial version of $\mathscr{L}$ we are going to consider what happens when a particular transformation is imposed.

In quantum mechanics an electron is represented by a wave function $\Psi$; this expression occurs in both $\mathscr{L}_{\text{DIRAC}}$ and $\mathscr{L}_{\text{INT}}$. Consider two different ways in which we might change the *phase* of $\psi$. First, suppose we change the phase in the same way at every point in spacetime; for example, we do this when we shift or rotate the axes. This kind of transformation, known as a *global phase change*, yields our first invariance requirement: $\mathscr{L}$ must be invariant with respect to a global change of phase in $\psi$. Mathematically, a global phase change is introduced by multiplying the original $\psi$ by a complex exponential ($e$ raised to a complex power). In this case we find that, because of its mathematical structure, $\mathscr{L}$ is indeed invariant. So this step introduces nothing new.[14]

Second, consider a *local phase change* in $\psi$. This involves different phase changes at different spacetime points. Invariance with respect to a local phase change is a much more demanding requirement than invariance with respect to a global change. To implement the local phase change mathematically we multiply $\psi$ by $\exp[ik\phi(x)]$, where $\phi(x)$ is a function of spacetime location. Thus the value of $\phi(x)$ can be different at different spacetime points, although these changes are related. When we do this we find that $\mathscr{L}_{\text{EM}}$ and $\mathscr{L}_{\text{INT}}$ are invariant, but $\mathscr{L}_{\text{DIRAC}}$ is not invariant: the transformed $\mathscr{L}_{\text{DIRAC}}$ consists of the original $\mathscr{L}_{\text{DIRAC}}$ plus an additional term. We can, however, produce an invariant Lagrangian density by modifying the term in $\mathscr{L}_{\text{INT}}$ that represents the electromagnetic field. This modification is the gauge transformation (characterized by its mathematical form), and it cancels the troublesome term.

> The form of $\mathscr{L}_{INT}$ is *completely determined* by the requirement that this cancellation occur. Symmetries are ordinarily thought to restrict the *form* of the interaction, so that **the *determination* of the interaction itself by a gauge symmetry** is both extraordinary and intriguing. . . . [T]his noteworthy feature is present in all gauge theories.[15]
>
> (R 133, see also 142)

$\mathscr{L}$ is invariant, then, to a pair of transformations: a local phase transformation plus a gauge transformation of the interaction term. Moreover, the gauge transformation introduces a term that represents a massless boson: the virtual photon that transmits the interaction from one spacetime point to another. Thus the requirement of local-phase-change invariance *requires*

the existence of photons that carry the interaction. We will see this pattern repeated as we look at the other interactions that concern us.

One more item is needed to completely determine the interaction: we need to know its *strength* as well as its mathematical form. In the language of field theory, we need to know the strength of the coupling between the photon and objects that have electric charge. This coupling strength, which is determined experimentally, is a function of the familiar charge on the electron.

There was no explicit reference to group theory in this discussion of EI. Indeed, everything I have just described was known well before the integration of group theory into the mathematics of QFT. Mathematically, the phase change is multiplication by a complex number, which may be a function of spacetime variables; thus there is no need to talk of groups or matrices. However, when we look at this case from the later perspective it is straightforward to absorb it into that framework by treating a number as a $1 \times 1$ matrix.[16] When we do so, we find that the phase transformation has the characteristics of matrices that represent the group U(1). Thus we can say that the photon emerges out of the requirement that $\mathscr{L}$ be invariant with respect to a local U(1) transformation. U(1) is, then, the symmetry group for EI.

### *10.2.2 Weak Interaction*[17]

In order to clarify the use of QED as a model for WI, I want to say a bit more about U(1). Each element of U(1) is a $1 \times 1$ matrix that represents a phase shift. The single entry in such a matrix can be written as a complex exponential, exp(ik), where k is a real number. The adjoint matrix is exp(-ik) which is not equal to the original, so these matrices are not Hermitian and cannot represent physical quantities (*A2*). This does not generate a problem since the phase of a wave is not a physical property. However, the product of our little matrix and its adjoint is one, so the adjoint of a U(1) matrix is its *inverse* (*A5*). Matrices for which the adjoint is identical with the inverse are known as *unitary* – thus the label U(1). Unitary matrices have many important properties. In particular, if we think of a vector in space, the operator that rotates that vector (or rotates the axes) is unitary.[18] When a unitary operator rotates a set of vectors it preserves the lengths of the vectors and the angles between them. In effect, unitary operators implement changes of basis – where a *basis* can be thought of as an analogical extension of the set of coordinate axes used to describe a vector in 2D or 3D space. Unitary transformations change the way we describe a physical situation, but leave at least some physical properties unchanged (*A4–A5*). Thus unitary transformations are symmetry operations. Symmetries that are represented by unitary matrices are known as *unitary symmetries*; all symmetries considered in this discussion are unitary.

Unitary matrices take on greater significance when we consider larger groups and their representations by larger matrices. U(2) can be introduced as the set of all $2 \times 2$ unitary matrices. The subset of these matrices with the

additional property that their determinant is equal to one also forms a group, SU(2); it plays a key role in WI.[19] The SU(2) symmetry that concerns us is a symmetry with respect to an *isospin rotation*; it involves an extension of the concept of isospin to WI (known as "weak isospin," $I_w$). The idea is that certain particles are indistinguishable with respect to WI; these can be viewed as having the same value of $I_w$, with different particles in a set distinguished by values of one component of $I_w$. (I discuss the role of components in Sec. 10.3.) SU(2) operations "rotate" one of these particles into another. Recall that there are three families of leptons, each consisting of a massive particle and its characteristic neutrino; each family is distinct with respect to an SU(2) rotation. As a result, WI can turn an electron into an electron neutrino, and vice versa, but cannot turn an electron into a muon, or an electron neutrino into a muon neutrino, and so forth. Quarks also respond to WI, but there are additional complexities in this case; I will consider some of these complexities below. First, let us examine the mathematical framework of WI.

We can begin as we did in the case of QED. We write down a first approximation to the Lagrangian density, $\mathscr{L}_w$, and require that this expression be invariant with respect to a local SU(2) transformation of the state function. Restricting discussion for now to leptons, the relevant state function is a doublet consisting of a massive lepton and its associated neutrino. Again the invariance requirement fails, but can be restored by a gauge transformation on the field, which modifies the form of the interaction term. In this case the gauge transformation requires the introduction of *three* gauge bosons, one positively charged, one negatively charged, and one neutral, $W^+$, $W^-$, and $W^0$. Let me underline why SU(2) yields three gauge bosons. The fundamental representation of SU(2) consists of $2 \times 2$ matrices, which operate on the two-component state function. The operator may be any SU(2) matrix, but SU(2) has three generators (*A7*), and each SU(2) matrix can be expressed in terms of these generators. Each generator represents a conserved current, which is a gauge boson.

However, a complication now appears because the mathematics of gauge invariance *requires* that these bosons are, like $\gamma$, massless. (See R 135 for the electromagnetic case and R 142 for the case of SU(2) and the general result.) In general, massless field particles imply that the interaction has infinite range while the range of WI is quite short.[20] The problem, then, is how to introduce massive bosons; this will require some new ideas.

The key step is to maintain that the symmetry is not exact, but *spontaneously broken*.[21] Since the symmetry requires massless bosons, introducing massive bosons will disrupt the symmetry. However, the symmetry cannot be broken by introducing a mass term into $\mathscr{L}_w$ because it will not be possible to make meaningful calculations in the resulting theory.[22] Symmetry breaking is implemented by the *Higgs mechanism*. An additional *field* is introduced – the Higgs field – that pervades all of spacetime and interacts with the weak-field bosons in a way that yields the massive field carriers we need.[23] I will not

pursue the details of how all this is done (see CG Chs 10–11; R Secs 11.1–11.2), but it is important to note that the resulting theory requires the existence of a massive boson that mediates the Higgs field. So far this Higgs boson has not been observed, but its exact mass is unclear – which allows for the possibility that it is out of reach of the present generation of particle accelerators. However, if this particle is not eventually detected, the entire approach is in deep trouble. Thus introduction of the Higgs mechanism brings along new testable empirical consequences. In addition, the unified theory EW yields several confirmed predictions, including the existence of three WI field bosons and the existence of weak neutral currents, a previously unknown type of interaction. (See CG Ch. 13 for a summary of the empirical situation.)

Let us now turn to the unification of QED and WI in EW. Consider the four bosons of EI and WI: $W^+$ and $W^-$ have electric charge while $\gamma$ and $W^0$ lack electric charge. EW introduces two new electrically neutral bosons that are constructed out of combinations of $\gamma$ and $W^0$. An image that might be helpful is to think of the original, utterly distinct, $\gamma$ and $W^0$ as two orthogonal vectors. In EW these are rotated towards each other making the angle between them less than 90°, with the result that the two vectors "mix." The angle of rotation, which measures the degree of "overlap," is known as the *weak mixing angle*, $\theta_w$. The *new* electrically neutral bosons are mutually orthogonal constructs out of $\sin\theta_w$ and $\cos\theta_w$. One of these mediates EI and the symbol $\gamma$ is retained; the other mediates WI and is now labeled Z. According to TC, the concepts used to describe these bosons are different from, although continuous with, their predecessors, and I have just described ways in which they are continuous and different.[24] In addition, while the strengths of the two interactions are different in the world we live in, they are identical above an energy of 200 GeV. This is possible because the strengths of the interactions vary with energy. For the energy levels at which we live the two interactions appear to be quite different, so there are two distinct coupling constants in EW. But at sufficiently high energies only one coupling constant is required. Presumably, there was complete unification of the two interactions during the very early history of the universe, but that time has passed.

The symmetry group of EW is SU(2) × U(1), a combination (direct product) of symmetry groups we have already encountered. However, there are important differences between the way these groups appear in EW and their earlier appearance. In EW, U(1) is the symmetry group of the new electromagnetic part that results from the combination of $\gamma$ and $W^0$. Moreover, the original electric charge is replaced by a new quantity known as hypercharge (Y).[25]

The weak part of EW involves an additional complication. It had long been assumed that all interactions conserve "parity" – roughly, that nature does not distinguish right from left – so that if a given process occurs, its mirror image also occurs.[26] But a variety of evidence suggests that this is not true for WI, and this fact must be included in a theory of the interaction.

The point is particularly central for of neutrinos, which have a spin of 1/2. We can think of a neutrino as a particle that spins on an axis as it moves through space in a straight line (for caveats see Sec. 10.3.4). There are two possible spin states (say, clockwise and counterclockwise), and we can represent each by an arrow pointing in the direction that a right-hand screw moves when turned in the corresponding manner. We can represent a particle's direction of motion by another arrow. If the two arrows point in the same direction the neutrino is labeled "right-handed"; if the directions differ, "left-handed." Available evidence indicated that only left-handed neutrinos exist in nature. But the electron, muon, and tau occur in both right-handed and left-handed versions, and this difference was built into EW. As a result, each of the weak-isospin doublets contains the left-handed version of a massive lepton and its associated neutrino. Matrices of the fundamental representation of SU(2) operate on these doublets. The right-handed version of each massive lepton forms an isosinglet with $I_W = 0$.[27] As a result of the two features just described, the symmetry group of EW is often written as $SU(2)_L \times U(1)_Y$.

Since quarks respond to WI, weak isospin applies to quarks. As we have seen, quarks are also divided into three families, each consisting of two quarks. In order to apply WI to quarks, these families must be broken up into left-handed and right-handed components; the left-handed components of each quark family forms a weak-isospin doublet; each right-handed quark is a weak-isospin singlet.[28]

It should be clear that this is a limited unification. The two parts of SU(2) $\times$ U(1) are completely independent of each other (R 173). U(1) continues to serve as the symmetry group for the electromagnetic part of EW. It is an Abelian symmetry (*A5*) that holds exactly and is mediated by one massless field boson with infinite range. SU(2) is the symmetry group of the weak part. It is a non-Abelian symmetry that is broken in EW and is mediated by three massive field bosons that have very short-range. I noted above that it is useful to think of particles that respond to WI as carrying a weak charge in analogy to the electromagnetic charge. All quarks and leptons carry weak charge. The bosons that mediate WI also carry weak charge. Indeed, this is a consequence of the fact that the WI symmetry group is non-Abelian (R 141–42). As a result, there are weak interactions between these bosons. The photons that mediate EI do not carry electric charge, so there are no electromagnetic interactions between photons. In addition, the electromagnetic part of EW conserves parity while the weak part does not conserve parity. EW also involves two coupling constants, one for the EI part and one for the WI part. I return to the nature of this unification in Sec. 10.5.

### *10.2.3 Strong Interaction*

QED deals with electrically charged particles; a local phase shift is implemented through multiplication by a complex exponential, and the symmetry

group is U(1). WI deals with both doublets and singlets; it would deal only with doublets if parity were conserved. Leptons and quarks respond to WI so the doublets are of two types: each lepton doublet consists of a massive electrically charged particle and an uncharged massless neutrino; each quark doublet consists of two massive electrically charged quarks, one with charge $+2/3e$ and one with charge $-1/3e$. In both cases a doublet can be viewed as two states with the same weak isospin. Since the isospin operators are expressed mathematically as square matrices, the existence of doublets requires $2 \times 2$ matrices and the symmetry group for the weak interaction is SU(2).

We have seen that SU(3) was originally introduced as a tool for classifying hadrons (this preceded use of SU(2) as the symmetry group for WI). At the time in question only three quarks were known. Suppose these quarks form an SU(3) triplet, and that each baryon is made of three quarks. If we associate the three quarks with a **3** IRR of SU(3) (*A6*), then the allowed kinds of baryons would be determined by the possible choices of three quarks; these could be accommodated by a $\mathbf{3 \times 3 \times 3}$ representation. This is a reducible representation of SU(3) that can be decomposed into IRRs in just one way: $\mathbf{3 \times 3 \times 3 = 1 + 8 + 8 + 10}$. We have already encountered sets of baryons fitting **1**s, **8**s and **10**s. Now suppose that mesons consist of a particle and an anti-particle; we have: $\mathbf{3 \times \bar{3} = 1 + 8}$, again fitting the known particles.[29] While the discovery of more quarks undermined this use of SU(3), the accounts of baryons and mesons survive. SU(3) is the SI symmetry group, although for a quite different reason. Each quark must itself be considered a triplet and SU(3) is the symmetry group for these triplets. Let us examine the reasons for this multiplication of quarks.

Consider the thesis that baryons are made of three quarks. Baryons and quarks are fermions; they have half-integral spin and conform to the Pauli exclusion principle. Yet some baryons appear to violate these requirements. To see why note that each quark in a doublet has spin 1/2, and that the two quarks are distinguished from each other by a component $S_z$ that is either 1/2 or –1/2 (this is explained in Sec. 10.3.4). In a complex object values of spin add, as do values of spin components. The $\mathbf{\Delta}^{++}$ is made up of three u quarks and has $S_z = 3/2$; thus each quark must have $S_z = 1/2$. Since the three quarks are identical in all other respects, we seem to have three identical fermions bound together in a single particle, which violates the exclusion principle. Moreover, if the $\mathbf{\Delta}^{++}$ consists of three identical quarks its state function will remain the same when we interchange any two quarks; thus its state function is symmetric. But this violates the spin-statistics theorem which requires that particles with half-integral spin have anti-symmetric state functions (*A3*). Both problems can be solved by introducing a new quantum number which has three allowed values, and postulating that the three quarks differ in this quantum number. This new property is conventionally referred to as *color* and the three allowed values as red (r), green (g), and blue (b) – plus anti-red ($\bar{\text{r}}$), anti-green ($\bar{\text{g}}$), and anti-blue ($\bar{\text{b}}$). (There is no connection between this

terminology and our ordinary notion of color; color is an additional quantum number indicating an additional degree of freedom.) The theory that results is known as "quantum chromodynamics" (QCD). The six quark types are now called *flavors*, and each flavor comes in three colors. Color is an internal quantum number; it does not appear in any detectable particle. It is postulated that the three colors occurring together cancel, so that every baryon is made up of three quarks of three different colors, leaving the colorless baryons that appear in detectors. In the case of mesons, color neutrality is achieved if each meson consists of a colored quark and an anti-quark of the corresponding anti-color. For example, the $\pi^+$ could be made up of a red u and an anti-red anti-d. In addition to solving a theoretical problem, the color postulate yields testable predictions that have been confirmed (cf. CG 12–13, 126 and R 313–16, 324–25).

SU(3) is the *symmetry group for quark color*. The state function for each quark flavor consists of an SU(3) triplet encompassing the three distinct quark colors. The colored quarks making up this triplet are identical with respect to the strong force: the force between quarks is the same no matter what their color. Let us see how this works in the case of mesons. We can think of the quark and anti-quark that constitute a meson as continually changing colors, but doing so in tandem so that we always have a color and its anti-color. This gives three permitted color states of a specific meson: $r\bar{r}$, $g\bar{g}$ and $b\bar{b}$. As is typical in quantum theory, a meson should be thought of as a superposition of these three states. Each state can be viewed as an axis in "meson color space," and each meson as a vector in this space. An SU(3) transformation on this vector changes the components but leaves the vector (the meson) unchanged. Note again the analogy with vectors in 3D Euclidean space where rotation of the axes changes the components of the vector in a coordinated way, but leaves the vector's length and direction unchanged. There is a similar, although more complex, account of baryons.

The requirement that SI be invariant with respect to an SU(3) transformation implies, again, the existence of massless bosons that mediate the interaction. SU(3) has eight generators, and there are thus eight massless field carriers – the gluons (*A7*). In this case *the symmetry is exact so that gluons are massless*. The quarks within a hadron are continually exchanging gluons – so SI holds hadrons together. Color is the analog of charge for QCD and gluons mediate color charge much as photons mediate electric charge. The forces between composite hadrons – including the forces between nucleons that were the original subject of SI – are a residual effect of this underlying interaction between quarks. Moreover, since SU(3) is a non-Abelian group, it follows that gluons carry color charge: Every gluon has a color and an anti-color, which are not necessarily the same type. An example will illustrate how color integrates with gluon exchange: A blue quark might emit a gluon that is blue and anti-red. As a result, that quark is now red. However, the gluon will be absorbed by a red quark that becomes blue, keeping the composite quark colorless.[30]

Still, SI is a short-range interaction. Since the mediating bosons are massless we need another mechanism to account for this limited range. This mechanism is *quark confinement*, which I introduced in Sec. 10.1. Recall (note 5) that each fundamental particle is surround by a cloud of the virtual bosons required by the particle's charges: photons for electric charge, W and Z bosons for weak charge, and gluons for color charge. In addition, these virtual bosons produce other virtual particles that must be included in calculations. In the case of EI, virtual photons produce virtual electrons and positrons which shield the charge on the electron from measurements. As a result, the apparent strength of EI decreases as we measure it from greater distances. Virtual photons, which have no electric charge, do not contribute to this shielding. In WI and QCD the field bosons have the field charge – weak charge in the case of WI and color charge in the case of QCD. This must be taken into account in calculating the strength of the field, and it turns out that charged virtual bosons produce an *anti-shielding* effect: they *increase* the strength of the charge as the distance increases. In WI this effect is small compared to that of the field-boson masses, but in QCD the effect is crucial. As the distance between two quarks increases the number of virtual gluons produced between them increases; these have the effect of increasing the potential energy between these quarks. This increase is essentially linear and yields a basically constant force holding the quarks together. The result is quark confinement – the failure of isolated quarks to appear – and accounts for the short-range of SI. (See CG 149–53; R 148–54 for some quantitative discussion.) However, "There is no analytical proof of confinement. Confinement is not displayed in perturbation theory, but numerical simulations demonstrate convincingly that QCD has this necessary property for an acceptable theory" (CG 148, cf. Veltman 2003: 49).

When we include SI in SM we have a theory that is unified in terms of its general mathematical framework. The symmetry group of the combined theory, SU(3) × SU(2) × U(1), consists of three distinct groups along with the other variations noted above. We must also include a third coupling constant which unifies with the other coupling constants at temperatures above $10^{15}$ GeV ($10^{28}$ °F).

## 10.3 From Angular Momentum to Isospin

In this section I trace a series of transformations that take us from the concept of angular momentum in classical physics to the concept of isospin in SM.

### *10.3.1 Angular Momentum*

Angular momentum is conveniently introduced by considering a symmetrical rigid body spinning around its axis of symmetry. In the absence of friction or other external forces the spinning will continue forever. Thus there is a

conserved quantity involved in this phenomenon, *angular momentum*, which depends on the body's angular velocity, mass, and the way the mass is distributed – with mass that is further from the symmetry axis making a greater contribution. Whirling ice skaters provide the classic illustration of this last factor: their angular velocity is greater when the arms are close to the body than when the arms are extended. Since extending the arms moves mass away from the axis, while the total mass is unchanged, the angular velocity must be reduced to keep the angular momentum constant. There is, then, a limited analogy between the angular momentum and linear momentum of a rigid body: both depend on mass and velocity, but the mass distribution enters into angular momentum. Angular momentum is a vector quantity that can be represented by a vector along the symmetry axis; it is conventionally taken to point up for clockwise rotation and down for counterclockwise rotation.

The concept of angular momentum can be extended to cases that do not involve rigid bodies. One extension deals with a point particle moving in a circular orbit around some central force that keeps the particle in orbit. We need not be concerned with the detailed nature of this force; also, since we are dealing with a point particle, the mass distribution can be ignored. Since the particle tends to move off in a straight line tangent to the circle, at each moment the particle has a linear momentum **p**. Let **r** be the vector from the center of force to the particle; the particle's angular momentum is defined as **r** X **p**, where "X" stands for the vector cross product. For our purposes it is sufficient to note that this product yields a vector that is perpendicular to the plane defined by **r** and **p**, with magnitude equal to the product of the magnitudes of those two vectors.[31] Classical angular momentum is a continuous parameter. A quantized version entered physics in 1913 with Bohr's theory of the atom.

### *10.3.2 Bohr's Theory of the Atom*

Bohr postulated that electrons in an atom move around the nucleus only in specific circular orbits, each having a definite energy and angular momentum. Contrary to the predictions of classical electromagnetic theory, electrons in these stable orbits do not radiate energy.[32] Radiation, and thus spectral lines, are produced when an electron jumps from a stable orbit to another stable orbit with lower energy. Bohr assumed that the frequency of the emitted radiation equals the energy difference between the two orbits, $E_i - E_f$, divided by Planck's constant h. He also assumed that the angular momentum in an orbit is an integral multiple of $\hbar$, and that the electron is held in its orbit by the electromagnetic attraction between it and the nucleus. Quantized angular momenta plus the restriction to circular orbits imply quantized energy and a one-one correspondence between energy and angular momentum. The assumption that momentum is quantized constitutes our first departure from the classical concept of angular momentum.

In classical physics all values of angular momentum are equally permissible so the concept does not imply a privileged set of angular momentum values; a new implication is introduced with the Bohr atom.

While Bohr's theory had some success in accounting for the hydrogen spectrum, it did not account for the fine structure of that spectrum or generally for the spectra of more complex atoms. There were important attempts to elaborate the theory and improve the fit to the data. For example, Sommerfeld introduced elliptical orbits, an extension that is quite natural for a central-force field and that Bohr had noted as a possibility. The shape of an ellipse is determined by the ratio of its two axes so that ellipses permit a wider range of shapes and a wider range of angular momenta for a given energy. Angular momentum is still quantized, restricting the allowable range of orbital shapes and requiring a new quantum number to specify these shapes. This extension introduces new energies and improves the theory's ability to account for the hydrogen fine structure because an electron moving on an ellipse will approach quite close to the nucleus, and move at a sufficiently high speed to require including the relativistic mass change. Taking this factor into account led to a somewhat better fit with the data. Incorporation of relativity brings along a host of additional implications. Sommerfeld and others continued to pursue a better fit to experiment by introducing further quantum numbers, but these attempts were of limited success and were superseded by the development of modern quantum theory (Rigden 2002 Chs 4–6, Tomonaga 1997 Ch. 1).

### *10.3.3 Quantum Theory*

Now consider the status of angular momentum in the non-relativistic quantum theory of the hydrogen atom. For reasons that will become clear shortly, I want to consider how we could locate an electron in 3D space. This requires three independent coordinates. Since our present concern is to locate the electron with respect to the nucleus, we can place the origin at the nucleus. In Cartesian coordinates we construct three mutually perpendicular axes and locate the electron by giving its coordinates along these axes. In the present case spherical coordinates are more convenient. First we construct a line from the origin to the electron; its length $r$ is one of our coordinates. In addition, we use two angles: $\theta$ determined by projecting $r$ onto the x-y plane and taking the angle between the resulting line and the x-axis, and $\varphi$ equal to the angle between $r$ and the z-axis. For the moment, we can ignore $\varphi$ and treat the atom as essentially 2D with the electron moving in a plane around the nucleus; this parallels the treatment in Bohr's theory. In effect, we are using polar coordinates with the electron located by $r$ and $\theta$.

In quantum theory (as in Bohr's theory) $r$ determines an electron's energy and is quantized. The value of $r$ for the lowest-energy electron will serve as a unit since all other energies are integral multiples of this energy; they can be specified by giving just the multiplier $n$ which can have any integral value

from one on up. Angular momentum is determined by the rate of change of $\theta$; it is quantized and equal to an integral multiplier $l$ times this rate of change and some additional parameters that are the same for all electrons. In quantum theory for *each* value of $n$ (that is, for each energy level) there are $n$ permissible values of angular momentum. Thus there are $n$ permissible values of $l$ *running from 0 to* $n - 1$. The *magnitude* of the angular momentum associated with a specific value of $l$ is $\hbar[l(l + 1)]^{1/2}$. For example, at the third energy level $n = 3$ and $l$ has three permissible values, 0, 1, and 2, with corresponding angular momenta of 0, $\hbar\sqrt{2}$, and $\hbar\sqrt{6}$. Each of these values of $l$ specifies a different electron state, although all share the same energy. Note especially that for each value of $n$, one of the permissible values of $l$ is zero; $l = 0$ is the only permissible value for $n = 1$. Since a circulating electron cannot have zero angular momentum, we should no longer think of the electron as literally moving around the nucleus. We have, then, two changes in the implications associated with ANGULAR MOMENTUM as we move from the Bohr treatment of the hydrogen atom to its quantum mechanical counterpart: A given energy implies a specific range of allowable angular momenta, and an electron can have an angular momentum of zero.[33] There are more changes to come, but before pursuing them I want consider why this quantum-mechanical parameter should be thought of as angular momentum at all.

Recall a central thesis of this book: Concepts are not words and conceptual shifts are not always reflected in linguistic changes. Often, when the concept associated with a word is changed in a systematic way the term associated with the older concept is retained. Thus in studying conceptual change we should not focus on the language, but on specific ways in which the new concept is similar to and different from its predecessor. Systematic relations between the classical concept of angular momentum and the quantum-theoretical concept are particularly clear because it is standard procedure to derive the operators required for quantum theory (*A2*) from classical expressions by a specific set of substitutions. The quantum mechanical expression for the angular-momentum operator is derived from the classical expression in just this way, and this procedure allows a direct comparison between the classical and quantum-mechanical expressions. In addition, classical angular momentum is a conserved quantity; this aspect is partially taken over into both Bohr's theory and quantum theory. In the latter two theories angular momentum is a constant of the electron's motion in a given state; this is why angular momentum can serve as one of the parameters used to specify that state. However, in Bohr's theory and quantum theory – but not in classical mechanics – the state can change spontaneously (so "state" does not have the same meaning as it did for Descartes and Newton). Still, the units of angular momentum are the same in all three theories.

Next, consider how we assess whether electrons have angular momentum. We are dealing here with integrated theories in which various features

cannot be isolated for individual evaluation; we have reasons for attributing a property to electrons just in case that property plays a role in accounting for the relevant data. Thus the evidence for accepting a theory is also evidence for holding that the concepts embodied in that theory are instantiated. The explanation of spectral lines plays a common role as relevant evidence for the three theories we are considering. To be sure, classical mechanics cannot explain many features of spectral lines, but the fact that this inability is recognized as a failing of classical theory shows that spectra are among the relevant phenomena. Explanation of some spectral lines provided a major success for the Bohr atom, and the ability to explain these lines was also a major constraint on the development of quantum theory. The existence of a body of data relevant to all three theories provides a common feature of the concepts used in those theories.

There is also considerable overlap in the role that angular momentum plays in the Bohr and quantum theories. In both cases $\theta$ is one of the parameters needed to locate an electron in space, and angular momentum is determined by the rate of change of this parameter.

Now consider $\varphi$, the third parameter that appears in the quantum mechanical treatment of the hydrogen atom. Beginning with an analogy to classical mechanics, if we think of the electron as moving around the nucleus in a plane, we can construct an axis through the nucleus perpendicular to this plane. This would be the z-axis in Cartesian coordinates, and angular momentum can be represented by a vector pointing along this axis, with its length proportional to the magnitude of the angular momentum. Let us now construct another line through the origin at an angle $\alpha$ to the z-axis. Classically, we can calculate the projection of the angular-momentum vector on this axis – it is the magnitude of the angular momentum multiplied by $\cos\alpha$. For a given angular momentum, the angle determines the magnitude of this projection. Although such a calculation is not essential to the Bohr atom, it could be carried out.

In quantum theory $\varphi$ gives the magnitude of a component of the angular momentum on some axis; the number that results has the right units for this interpretation. However, the differences from classical mechanics are very substantial. First, this is an independent parameter that is essential for specifying the quantum-mechanical state of an electron; the parameter cannot be calculated from the magnitude of the angular momentum and some set angle. Just as the energy level determines a set of permissible values of the angular momentum, so the angular momentum determines a set of permissible values of the projection on an axis – which is customarily referred to as the z-axis.[34] For a given value of $l$, the $z$-component will be $m_l\hbar$, where $m_l$ must have one of the values ranging from $l$ to $-l$ in steps of one unit. Thus for $l = 2$, $m_l$ must have one of the values 2, 1, 0, –1, -2.

Second, the maximum value of this component, $\hbar l$, is *less* than the magnitude of the angular momentum $\hbar[l(l + 1)]^{1/2}$. This difference is required by the indeterminacy principle since the operators representing components of

angular momentum along any two mutually-perpendicular axes do not commute. Thus there is an indeterminacy relation between any two of these components.[35] But if a component of the angular momentum equaled the momentum then (as in classical mechanics) the values of the two remaining components in a system of Cartesian coordinates would be zero, and all three components would be simultaneously determined.

Third, the range of values for a component is independent of the angle between the z-axis and the angular momentum vector. The permissible values of these components will be the same no matter what the angle of the z-axis, although the *probabilities* of these values depend on details of the system.

### *10.3.4 Spin*

I will consider the introduction of this concept in three steps. First, one problem about spectra that arose in the old quantum theory was that available theory predicted single lines where the evidence indicated two closely spaced lines of slightly different energies. In 1925 Goudsmit and Uhlenbeck proposed dealing with the problem by considering the electron as spinning on its axis.[36] We can think of the electron has having two possible spin directions, clockwise and counterclockwise, each of angular momentum $\hbar/2$. This yields two possible spin vectors of equal magnitude and opposite direction. A spinning electron will have a magnetic field – also with two possible directions – that will interact with the magnetic field generated by the electron's orbital motion. Vector addition of this field with each of the fields generated by the spin yields two energy levels, and thus the required doubling of the spectral lines.

As originally conceived, this proposal involves an electron that is literally a ball of charge spinning around an axis; it was clear to many physicists that the proposal has unacceptable consequences.[37] For example, after submitting their paper Goudsmit and Uhlenbeck consulted Lorentz who argued that the proposal would require a diameter for the electron that was unreasonably large (approximately $10^{-12}$ cm where the diameter of the nucleus was estimated to be $10^{-13}$ cm). They tried to withdraw the paper, but it was too late (Pais 1986: 277–78). Tomonaga notes a related problem:

> if the size of the electron is $e^2/mc^2$ as H. A. Lorentz has considered, then so fast a rotation is needed to have a self-rotating angular momentum of 1/2 that the electron's surface reaches a speed ten times higher than that of light.
>
> (1997: 35)

Second, in 1927 Pauli reworked the idea of spin in the context of the new wave mechanics. Mathematically this requires introduction of appropriate operators. Still thinking of spin as a vector quantity, Pauli introduced three

such operators – one for spin around each of the three coordinate axes. In choosing these operators Pauli took his guidance from the quantum mechanical properties of angular momentum – a reasonable enough procedure since spin is presumably a form of angular momentum. As a result, spin shares the features of quantum-mechanical angular momentum noted above: The same indeterminacy relations hold among the spin operators as among the angular momentum operators, and (using $s$ instead of $l$ to specify spin) the magnitude of the spin is proportional to $[s(s + 1)]^{1/2}$. But new features appeared as well. While the basic unit of angular momentum is ħ, spin occurs in units of ħ/2. In addition, quantum-mechanical operators must operate on some function. Since Pauli's spin matrices are $2 \times 2$, he introduced a spin wave function with just two components: a spin eigenfunction with eigenvalue 1/2, and a spin eigenfunction with eigenvalue –1/2 (*A2*). But when Pauli examined how this new wave function behaves under rotations in space he found that it does not act like a vector. In addition, Pauli developed his account in terms of non-relativistic quantum theory; he explored a relativistic version but found it too difficult (Tomonaga 1997: 53). Both of these problems are solved at the next stage.

Third, in 1928 Dirac published the relativistically correct wave equation for the electron. From the point of view of relativity there is a problem with the Schrödinger equation since it includes second derivatives of the spatial terms but only the first derivative of the time term. Relativity requires that space and time terms be treated in the same way in this case. There are two different ways of constructing a relativistically correct equation. One way is to take second derivatives throughout; the resulting equation is generally known as the Klein-Gordon equation, although it was also discovered by several other physicists (Pais 1986: 288–89), and was already known at the time of Dirac's work. But Dirac objected to this equation on technical grounds (see Pais 1986: 289 and Tomonaga 1997: 56 for details) and found an equation that takes first derivatives throughout.[38] Electron spin appears as a consequence, although no attempt was made to build spin into the equation. In other words, spin is a relativistic phenomenon that does not involve any image of a spinning electron:

> Dirac has derived everything about electron spin through Lorentz invariance and that the wave equation must be first order without using a model at all. It may be since this work of Dirac's that we started not to think about self-rotation or rotation at all from the words *electron spin*.
>
> (Tomonaga 1997: 61–62)

This is particularly important because of the odd behavior that Pauli noted when considering rotations of the spin-1/2 wave function. When we use the Pauli matrices to implement a rotation of 360° we get the original wave function *multiplied by –1*; a second 360° rotation is required to recover the original wave function.[39] As Pauli noted, spin wave functions are not vectors. They are a new kind of mathematical object now known as *spinors*.

The Dirac equation requires an additional innovation: instead of Pauli's 2-component wave function, Dirac's equation requires four components acted on by $4 \times 4$ matrices that are extensions of Pauli's spin matrices. The additional components represent anti-matter – another consequence of Dirac's equation. Dirac's equation also gives the correct value for the spin magnetic moment, as well as other important results. Eventually particles were discovered with half-integral spin greater that 1/2 (3/2, 5/2, etc.). With this discovery it became clear that, as in the case of angular-momentum, a given spin $s$, corresponds to a class of states with one component running from $s$ to $-s$ in integral steps.

### *10.3.5 Isospin*

As noted in Sec. 10.2, during the 1930s Heisenberg introduced ISOSPIN as part of an attempt at a theory of SI. I want to describe Heisenberg's reasoning and the relation to spin a bit further. According to Tomonaga (1997: 164), at that time Heisenberg considered the force between neutrons and protons to be fundamental. He arrived at this view by noting that in the most stable atoms the number of neutrons and protons is roughly equal. If there were a comparable force between protons or between neutrons, there should be nuclei with just neutrons and just protons; thus he concluded that no such force exists. He then constructed his theory by analogy with what are known as *exchange forces* in the theory of chemical bonding.[40] Heisenberg considered an especially simple case: an $H_2^+$ *ion* (not an atom) which consists of two protons at a distance from each other, plus a single electron that moves back and forth between the protons. A quantum mechanical analysis of this situation results in a force that holds the protons together. This is the exchange force, so-called because it arises as a result of the electron exchanging its position between the protons. Reasoning by analogy, Heisenberg proposed that the neutron and proton are different states of a single particle – the *nucleon* – with an electric charge moving back and forth between them.[41]

Now another analogy comes into play. It was known at the time that there is an interaction between electrons in an atom that can be thought of as arising from a spin-spin interaction. If we think of *isospin* as an intrinsic property of nucleons, much as spin is considered an intrinsic property of electrons, there are only two possible isospin states, again paralleling the electron. Bringing together the idea of an exchange and the analogy to spin, Heisenberg adapted Pauli's spin formalism to his case. He introduced three isospin matrices of the same form as Pauli's – although these no longer represent spin about spatial axes, but rather *define* isospin space – and an isospin eigenfunction with just two eigenstates. Tomonaga (1997: 169–70) argues that the treatment of isospin as an intrinsic property is the key step in moving from the case of the ion with a charge moving through space to the present case which does not involve spatial transfer. Working with his

adapted Pauli formalism, Heisenberg was able to construct a theory of the nuclear force that yielded several testable consequences (Tomonaga 1997: 174–76). The notion of isospin introduces a new symmetry: The isospin of a nucleon remains the same as it is rotated in isospin space, so isospin is a conserved quantum number. The neutron and proton have the same value of isospin I but different values of a component $I_3$ (instead of $I_z$). This is referred to as an *internal symmetry* because it does not involve any transformation in spacetime.

Now let us jump ahead to the role of isospin symmetry in SM. When members of a set of particles respond in the same way to an interaction we have a respect in which they are interchangeable. This is reflected in the isospin formalism by their having the same value of I but different values of $I_3$. Consider the pion, a meson that comes in three varieties distinguished by their having positive, negative, or no electric charge ($\pi^+$, $\pi^-$, $\pi^0$). The masses of the three pions are almost identical: 140 MeV for the positive and negative pions, and 135 MeV for the neutral pion (Perkins 2000: 87–91). These form an isospin triplet with I = 1 and $I_3$ assignments of 1, 0, –1 for $\pi^+$, $\pi^0$, and $\pi^-$, respectively. Consider some empirical consequences of this assignment. First, because of the use of the spin formalism, the assignment of I = 1 to this multiplet *requires* that there be three types of pions; if the actual number of pions differed, the assignment would be incorrect. Second, treating pions as a multiplet with respect to SI requires that they all have the same coupling strength in strong interactions – for example, in those between any pion and any nucleon. Third, conservation of isospin places constraints on the possible interactions. Finally, isospin invariance places constraints on the Lagrangian for the pion-nucleon interaction (R 101–3).

Isospin is a genuine, full-blooded concept even though it has been constructed by a series of analogical steps beginning with the classical concept of angular momentum. We have seen that spin is already an extension and modification of the concept ordinarily associated with the word "spin" in everyday language and in classical physics. The familiar word is associated with a new concept even though there are clear differences between this concept and the concepts that preceded it. As with any descriptive concept, the content of this new concept is specified by its implicational relations with other concepts, its ties to its extra-systemic subject matter, and its systemic role. Some readers may want to describe these developments as "metaphorical" extensions of earlier concepts, and I have no objection to this terminology as long as it is not being used to suggest that the new concepts have some kind of second-class cognitive status. At this stage of our discussion it does not seem that introducing the label "metaphorical" adds anything to the discussion. Of course, the point of this label may be to invoke a different theory of conceptual development than that provided by TC. Once this alternative is specified and we are shown how it leads to a different detailed account of the developments we have been discussing, we can engage in the usual process of comparative theory evaluation.

## 10.4 Forces and Interactions

Much conceptual distance has been traversed in moving from the everyday concept of a force to the SM account of the fundamental interactions. In particular, the picture of a force as a push or pull has undergone significant change. While it might be possible to impose this image on some cases, it has no place in (for example) the particle decays that these interactions mediate. As a result, INTERACTION does not imply a push or pull. The mutual implication between INTERACTION and CHANGE OF MOMENTUM introduced by Newton is retained – although other changes in implications differentiate the twentieth-century versions of these concepts from earlier versions. For example, in Newtonian physics there is a mutual implication between force and the product of mass and acceleration (*ma*); in special relativity (SR) force is typically defined as change of momentum, but does not imply *ma*.

Consider another example: In everyday thought (and Descartes) interactions between physical objects require contact, while contact between bodies implies an interaction. Interactions between bodies at a distance from each other require a sequence of intermediate collisions – all occurring in ordinary space and time. In QFT, where every interaction is mediated by an exchange of field bosons, the notion of colliding bodies is irrelevant – field bosons are emitted and absorbed, they do not interact with other particles by "collision." The demand for something that mediates interactions between items at a distance remains, but the mediating bosons are virtual particles, not particles in ordinary space and time. Moreover, the concept of a virtual particle requires the indeterminacy principle. Thus the conceptual machinery of quantum theory is involved in the SM account of these interactions. In addition, Lorentz invariance is a basic constraint on allowable mathematical accounts of these interactions, so that SR is also built into these theories. As a result, claims about interactions in QFT carry a vast array of implications that are absent from everyday thought and from classical mechanics, along with implications that are at odds with implications of these older approaches. For example, virtual particles are not constrained by conservation principles that apply to particles in ordinary space and time. While only particles of the latter sort can be detected, virtual particles play a central role in the calculations that predict what will be detected.

Turning to the instantiation conditions for the SM account of fundamental interactions, there is little point to discussing whether an interaction exists apart from consideration of the detailed theory of that interaction. In SM the concept of a weak interaction cannot be divorced from its role in EW, where WI is partially integrated with EI. Testing whether these concepts have instances requires testing predictions of the integrated theory. As a result, such predictions as the existence of three WI bosons with different electric charges, weak neutral currents, and variations of measured values of the interaction coupling constants with distance, are constitutive of SM concepts. Failures of these, or other, predictions that lead to modification or

replacement of SM would generate conceptual change. Moderate changes of this sort would likely result in a change of the concept that physicists associate with the terms "weak interaction" and so forth. If a sufficiently drastic change occurred, the terminology itself might be dropped.

We find a higher degree of long-range continuity when we consider the systemic role of INTERACTION in thought about the physical world. It is a part of our common experience that physical objects do not exist in isolation. Even Leibniz, who denied that there are any interactions at the most fundamental level, had to provide an explanation of why objects appear to interact. Physicists seek an account of the kinds of interactions that exist and the relations between them, and develop testable quantitative theories of these interactions. (See Franklin 1993 for an account of a recent failed attempt to identify a new fundamental interaction.) In the history of physics there have been important changes in the understanding of the number of distinct interactions that exist in the world and in how to theorize about these interactions, but recognition that one goal of physical theory is to provide such an account remains a constant.

## 10.5 Unification

The search for a single unified theory that includes all interactions is a version of a successful research program that has been pursued for several centuries, but the understanding of what counts as unification has undergone changes as physicists learned more about the kinds of unification that are possible and appropriate in various domains. The project begins with the seventeenth-century rejection of Aristotle's bifurcated universe and reaches a major plateau with Newton. In this case unification meant eliminating the division of the universe into terrestrial and celestial realms made up of different kinds of matter following different laws. It was replaced with the view that there is a single physical universe with same material and the same laws throughout. Thus in the context of seventeenth century physical theory, UNIFICATION implies that there is only one domain to be studied where previously it was held that there are two distinct domains. Evidence that this concept is instantiated comes from the successful deployment of the same laws to account for the behavior of objects in the heavens and on the earth, and from evidence that the same materials are found in both realms. The major systemic role of UNIFICATION lies in the guiding assumption that there is only one body of physical science that applies throughout the universe – a revolutionary thesis in the early seventeenth century. Adoption of this guiding theme has ramifications throughout science. For example, by rejecting the view that physics and astronomy are different sciences concerned with different domains, unification entails that observations made on earth are relevant to claims about the heavens, and conversely.

The next major step in the pursuit of unification was Maxwell's theory of electricity and magnetism.[42] By Maxwell's time pursuit of unification was no

longer radical and the particular unification in question was much more limited: it concerned two phenomena that had been considered distinct. Empirical research had established pervasive relations between the two, so it was reasonable to suspect that they could be encompassed in a single theory. Moreover, by Maxwell's time the dominant role of mathematics in physical theory was standard, and the central feature of the unified theory is a single *mathematical* framework. Maxwell provided a set of four equations for electricity, magnetism, and the relations between them.[43] Two of these equations describe the field around a stationary electric or magnetic charge; I will consider these equations first.

Both electricity and magnetism occur with two different kinds of charge: positive and negative for electricity, north and south poles for magnetism. In each case the two kinds of charge are opposite in that a given quantity of one cancels the effect of an equal quantity of the other. For both electricity and magnetism like charges repel and unlike charges attract, and this attraction and repulsion conform to an inverse square law. *But*, there is an important empirically established difference between them that must be included in Maxwell's equations: Isolated positive and isolated negative electric charges occur; isolated north and south magnetic charges have not been found. A north pole and a south pole are always associated so that the net magnetic field at a distance from a magnet is always zero. This difference limits the unification so that the equations for electricity and magnetism are not completely parallel. Using **E** for the electric field, **B** for the magnetic field, and $\rho$ for the electric charge density, we have:[44]

$$\text{div } \mathbf{E} = \rho \tag{M1}$$

$$\text{div } \mathbf{B} = 0 \tag{M2}$$

The equations have the same mathematical form in that the same function of the field around a charge is set equal to the total charge. But in the electrical case the total charge depends on the specific situation, while in the magnetic case the total charge is always zero.[45]

There is also a connection between the two phenomena that is described by the remaining equations (**j** is the current density):

$$\text{curl } \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{M3}$$

$$\text{curl } \mathbf{B} = \mathbf{j} + \frac{\partial \mathbf{E}}{\partial t} \tag{M4}$$

Again there is some formal similarity: M3 relates a specific function – curl **E** – to the rate of change of **B** while M4 relates curl **B** to the rate of change of **E**. Since there are also clear differences I will consider each term on the right-hand sides of these equations. In M3 the term on the right describes a

*changing magnetic field*, so M3 specifies the relation between an electric field and a changing magnetic field. In M4 the first term on the right describes a *steady electric current*. There is no parallel term in M3 because there is no such thing as a steady magnetic current. The second term in M4 relates the magnetic field to a changing electric field. Maxwell introduced this term (although not in the form given here) because of empirical constraints. In his version the term was tied to a new concept – DISPLACEMENT CURRENT. Maxwell conceived of this current as the result of the polarizing effect of an electric field on the charges in a medium – including the ether. Thus, in Maxwell's original version, the formally similar terms in M3 and M4 have very different physical content. Maxwell's understanding of the M4 term has not survived: "The name stems from a certain mechanical picture of free space which Maxwell had in mind but which has since been found to be unnecessary and misleading" (Konopinski 1981: 25).[46] Still, M4 has endured; on the modern interpretation the term in question just describes a changing electric field. Thus on the modern interpretation there is greater unification of electricity and magnetism than in Maxwell's version, but it is still a limited unification.

A further development in electromagnetic theory will illustrate another kind of unification. Ampère maintained that electricity is more fundamental than magnetism in that magnetism always results from moving electric charges; this view eventually prevailed. Magnetism has thus been reduced to electricity, providing an instance of REDUCTIVE UNIFICATION. This kind of unification implies that on a sufficiently fundamental level only one of the phenomena exists. Thus REDUCTIVE UNIFICATION implies a difference in the ontological status of the items involved in the reduction. Where physicists previously thought in terms of two distinct (although related) phenomena that have the same ontological status, these are no longer considered distinct and only one is fundamental. The seventeenth-century unification of the heavens and earth does not involve this kind of shift; after unification the celestial and terrestrial realms have the same ontological status. SR provides another perspective on the reductive unification of electricity and magnetism. The balance between electrical and magnetic fields is different in different reference frames so that, in effect, what appears as part of a magnetic field in one frame will appear as part of an electric field in another frame, and conversely. *But*, it is always possible to choose a reference frame in which a given charged particle is stationary and thus has no magnetic field. There are no frames of reference in which the particle has a magnetic field but no electric field.

Einstein's work resulted in a number of unifications that are worthy of detailed study, but I will consider only the unification of space and time in SR, which introduces another variation on our theme. One way of thinking about this case is that the apparently different phenomena of spatial and temporal gaps are shown to be "abstractions" from a more basic item. As an often-quoted remark of Minkowski's puts it (e.g., Taylor and Wheeler 1966:

37): "Henceforth space by itself, and time by itself, are doomed to fade away into mere shadows, and only a kind of union of the two will preserve an independent reality." More precisely, our familiar space and time are projections of the spacetime interval on different axes. To see what is involved consider the pre-Einsteinian understanding of space. Suppose we specify a set of three mutually perpendicular coordinate axes and determine the distance of a point P from the origin. This distance can be expressed in terms of the three projections of this point on the axes. If we rotate the axes in 3D space, keeping the origin unchanged, the distance of P from the origin remains the same, although its projections on the rotated axes (its coordinates) will typically be different from the previous coordinates. If we take the three coordinates in either of these frames and apply the Pythagorean theorem, we get the same value as before for the distance from the origin (see also *A3*). The moral is that the coordinate values on a particular set of axes have no isolated physical significance; they are only a means of keeping track of the invariant distance from the origin. There is a formalism for describing this situation in which we can say that the original coordinates mix when we rotate the axes, so that the value of a coordinate on the rotated axes is made up of contributions from the values on each of the original axes – and vice versa.[47] In this classical framework time provides a distinct 1D system to which the notion of a rotation does not apply.

The unification of space and time in SR consists of considering time to be one more axis that mixes with the spatial axes on rotation.[48] Thus, what appears as a temporal gap on one set of axes becomes part of a spatial gap on a rotated set, and conversely. However, a new invariant is introduced: the four-dimensional spacetime interval that gives the same value when calculated from the coordinates as measured on any of the various sets of rotated axes. This is a new kind of unification with a different function and different implications than those we have examined previously. In the seventeenth-century unification of the heavens and earth, and in Maxwell's unification of electricity and magnetism, the unified domains are on the same ontological level. In a reductive unification one domain is taken to be more fundamental than another. In Minkowski's account the unified domains have the same ontological status, but both are seen as aspects of a more fundamental item. This can be viewed as another form of reductive unification, where both of the original domains are reduced to a new item that had not been conceived of in earlier frameworks.

However, the relativistic unification is not as complete as the above description suggests. Note, first, that the formula for determining the spacetime interval is not a straightforward four-dimensional extension of the Pythagorean theorem; rather, the space terms and the time term have different signs.[49] This difference in signs is crucial for SR. For example, it is required for the conclusion that the spacetime interval is always zero on a light ray. Thus a residual distinction between space and time is tied to the

special status of the speed of light – a central thesis of the theory. Another consequence of this difference in signs is that the mixing of space and time terms is limited in a way that mixing of spatial coordinates is not limited; this limitation is directly tied to considerations of causation. If it is physically possible for A to cause B – that is, for a signal moving at the speed of light to leave A and arrive in time to influence B – then no rotation of the axes will place B temporally before A.

The residual distinctions between space and time also appear when we consider how these concepts are tied to the world. In this case the key consideration is how we measure spatial and temporal gaps. These measurements require different procedures, which is reflected in the units of measurements. This difference in units also affects the formula for the space-time interval. Since items that are added or subtracted in the formula must have the same units, a conversion factor is required. The relevant conversion factor is the speed of light, so it is $ct$, not $t$, that enters the formula. While this multiplication of $t$ by $c$ has the required effect of replacing units of time with units of length, its full significance is worthy of substantially greater exploration than I will pursue here since $c$ is the most fundamental physical parameter in SR, not just a conversion factor arising from conventional definitions of units (such as the conversion between meters and feet).

I want to stress that these limits in the unification of space and time should not be considered a failure of the unification project. Rather, the massive empirical confirmation of SR indicates that *the kind of unification we find in SR is the appropriate kind for the actual world*. We return here to a persistent theme of this book: As we learn about the world we rethink our concepts. There is no good reason for holding that once we lay down a set of projects embodying a particular set of concepts; any departure from that starting point constitutes a failure. Rather, as research proceeds we learn what the appropriate projects are, and what concepts we need to formulate those projects. The seventeenth-century concept of unification embodies a project that may not apply in many domains; continuing use of the same word is not fundamental – it is the underlying concepts that are basic.

Now consider EW, which implements yet another unification concept. Two aspects of this unification merit our attention. One is the mathematical aspect – the respects in which the electromagnetic and weak parts of EW are combined in a unified mathematical theory. As in the case of Maxwell's equations, there is a common mathematical framework – gauge theory in the present case – although it occurs along with differences in mathematical details for the two parts of EW. These differences – Abelian vs. non-Abelian symmetry, exact vs. broken symmetry, and so forth – are more drastic than in the Maxwell case. The second aspect concerns the strengths of the two interactions, which differ in our world – requiring two different coupling constants in EW – but are identical above an energy of 200 GeV which presumably occurred in the distant past. Note how a new systemic role has

been introduced. All of the earlier unifications applied to the world as a whole: as we find it now, as it was in the past, and as it will be in the future. In the case of EW the weak interaction is weaker than the electromagnetic interaction at some energies, but has the same strength at other energies. Moreover, the differences in the coupling strengths are an essential feature of the theory.

This difference in role brings along differences between what is implied by this unification concept and by earlier versions. The central implications of the EW form of unification concern mathematical similarities on an extremely abstract level. The unified theories are both gauge theories in which a symmetry group determines the mathematical form of the interaction; we have already discussed the differences that appear when we look at the specific symmetry groups involved. In addition, the mixing of $\gamma$ and $W^0$ in EW involves a new kind of connection between physical parameters. Neither QED nor WI is reduced to the other, nor do we have each of them reappearing as aspects of a single phenomenon. So this is a new version of unification with a new set of desiderata associated with an old word. (See Morrison 2000 Ch. 4 for an illuminating discussion.)

### *10.6 Conclusion*

In Chs 9 and 10 I have examined a few important cases of conceptual change in physics. There is considerable detail in each discussion, although less detail than could have been supplied, and less than a specialist in any of these cases would prefer. I want to emphasize that this detail is central to a major thesis of this book: Detailed analyses are required to understand conceptual change. Attempts to give simple, one-dimensional answers to the question of whether conceptual change – or epistemically significant conceptual change – has occurred are, I urge, unilluminating. I have also attempted to show how TC provides a basis for organizing such studies.

## Appendix: Some Mathematical Concepts

In this appendix I sketch some key mathematical concepts used in quantum field theory (QFT). None of the concepts will be discussed with anything approaching rigor. I begin with some elementary notions, but move fairly quickly to material that is not usually found in introductory courses in quantum mechanics.

### *A1 Operators*

Operators, as the name suggests, represent operations that we carry out on some item. Let us first consider operations on mathematical expressions. Suppose we want to add three to algebraic expression $s$. We could create the

operator [+3] to represent this operation and write $[+3]s = s + 3$. We could also write $[\times 7]s = 7s$ to indicate multiplication by seven. These simple examples will serve to illustrate an important feature of operators. Suppose we apply an operator to $s$ and then apply a second operator to the result. Sometimes the order in which we carry out the operations does not matter. For example, adding $a$ to $s$ and then subtracting $b$ gives the same result as first subtracting $b$ and then adding $a$. In this case we say that two operations *commute*. Addition and multiplication do not commute: If we first add three to $s$ and then multiply the result by seven we get $7(s + 3) = 7s + 21$. If we reverse the order of the operations we get $7s + 3$. Thus $[\times 7][+3] \neq [+3][\times 7]$. Turning to calculus, consider the derivative of $s$ with respect to time, which we can symbolize as $D_t s$. It is basic to calculus that multiplying by some number and taking a derivative commute: $a(D_t s) = D_t(as)$. But taking a derivative and adding a number do not commute: $D_t s + a \neq D_t(s + a)$.

The distinction between commuting and non-commuting operations applies also in geometry. Consider a circular disk with a mark near the circumference at one point. If I rotate the disk around its center by 27° clockwise and then by 32° counterclockwise, the mark ends up in the same place as it does if I carry out the two rotations in the reverse order. Suppose, however, that I travel 100 miles east and then 100 miles north on the surface of the earth (assumed spherical). When I walk north I move along a meridian of longitude; walking east I move along a line of latitude. All meridians are the same length on a sphere, but lengths of lines of latitude differ: they are very short near the poles and a maximum at the equator. As a result of this variation, carrying out the two operations in two different orders will never leave me at exactly the same place; the difference may be considerable if I am close to a pole. (Suppose I start 100 miles from the north pole, keeping in mind that when I am at the north pole I can move only south?) Commuting operations always give the same result when the order is reversed; if there is a single exception the operations do not commute.

We will be concerned here with one important class of operators – *linear operators* – which have two defining characteristics that are individually necessary and jointly sufficient. First, consider an expression $s$, a linear operator [L], and a number $a$: $[L]as = a[L]s$. That is, if we both operate with [L] and multiply by $a$, the order in which we carry out these operations does not affect the result. Second, if we have a second expression $r$ and a suitably defined addition operation (such as vector addition), then $[L](r + s) = [L]r + [L]s$. We get the same outcome independently of whether we first add the expressions and then operate on the result, or operate on each expression and then add the two results. The basic operations of calculus, differentiation and integration, are linear; the operation of adding a constant to an expression is *not* a linear operation.[50]

Matrices are linear operators. The operations are implemented by matrix multiplication; carrying out two operations in succession will be represented by successive matrix multiplications. Matrix multiplication is non-

commutative; we will explore non-commuting linear operators further in the next section.

### *A2 Operators in Quantum Mechanics*

In classical physics properties of a system are typically described by mathematical expressions; we calculate other properties by means of mathematical operations on these expressions. For example, if we know the momentum of a system $p$ and its mass $m$, we can calculate the system's kinetic energy by first squaring the momentum and then dividing by twice the mass ($E = p^2/2m$). Let us explore a more complex case: a particle moving along the x-axis. We make no assumptions about the nature of its motion: it may be moving at a constant velocity, or undergoing a constant acceleration, or what have you. Suppose, however, that we have a mathematical expression that describes its distance from the origin at any given time; in the usual notation, $x = f(t)$. The time derivative of this expression yields a new expression that gives the particle's velocity at any time. The result will be zero if the particle is stationary, a constant if it is not accelerating, or a more complex expression if it is accelerating. The time derivative of this new expression yields a third expression that gives the relation between the particle's acceleration and time. Note also that given the expression for a particle's velocity, we can multiply this by its mass to get an expression for its momentum as a function of time. We can think of $x = f(t)$ as describing the particle's *state*, and the various operations as the means by which we extract information from this state description.[51]

In quantum theory this relation between state descriptions and operators is changed. Physical properties such as energy, momentum, location, and angular momentum are represented by *linear operators*. To describe a quantum system a physicist begins by writing down appropriate operators and uses these to determine the expression describing the system's state (usually written $\psi$). In elementary cases we first determine the system's energy operator (which, for historical reasons, is known as the *Hamiltonian*). Two different forms of the Schrödinger equation then come into play: the time-independent form allows us to calculate an array of possible states from this operator; the form that includes time allows us to calculate how states change with time. The significance of the quantum-mechanical state function has elicited considerable debate; its correct interpretation is one of the cluster of difficult issues involved in understanding what quantum theory tells us about nature. For present purposes we can ignore these debates and just focus on the point that mathematical operators play a different role in quantum theory than they do in classical physics. *In quantum theory every physically significant feature of a system is represented by an operator*. Once the state function for a system is known, the operators for various properties can be used to extract information about those properties, but this does not change the fact that the relation between

operators and state descriptions is different in quantum theory than in classical physics.

Now consider the relation between an operator that represents a physical quantity and specific values of that quantity. There is a common type of equation in which a linear operator acts on an expression and gives back the same expression multiplied by a number. A typical case in calculus is the derivative of an exponential: $D_t e^{at} = ae^{at}$; the time-independent Schrödinger equation is another example. Put abstractly, if $A_{op}$ is an operator, $\psi$ a state, and $a$ some real number, we have an equation of the form $A_{op}\psi = a\psi$. Equations of this sort are known as *eigenvalue equations*: $\psi$ is an *eigenstate* of $A_{op}$, and $a$ is an *eigenvalue* of the operator. Typically an operator will have many different eigenstates, each associated with an eigenvalue. Each eigenstate has one associated eigenvalue, but different states can have the same eigenvalue. In this case the eigenvalue is described as *degenerate*.

I am now using a capital letter with the subscript *op* to stand for an operator; I will use the same letter without the subscript to stand for the physical quantity that the operator represents. In quantum theory, if $A_{op}$ represents the physical quantity A the operator's eigenstates are possible states of the system; each eigenvalue is the value of A in the corresponding state. If we measure A the result will be an eigenvalue of $A_{op}$. But it is important to distinguish two different kinds of cases. First, suppose that a system is already in an eigenstate of $A_{op}$. Measuring A will not affect the state of the system. Mathematically, we have $A_{op}\psi = a\psi$; physically, we have measured A and found the associated eigenvalue. Second, a system may not be in an eigenstate of $A_{op}$. In this case solution of the mathematical problem will give a set of eigenstates, their associated eigenvalues, plus the probability that we will find the system in each state when we carry out a measurement. The physical act of measurement will change the system so that it is now in a specific eigenstate of $A_{op}$; the measured value will be the associated eigenvalue.[52]

Given that measurable outcomes are eigenvalues of operators, a bit more must be said about the operators used in quantum theory. These are linear operators that can be represented by matrices; typically these matrices include complex numbers. Corresponding to every linear operator is another linear operator called its *adjoint*. Without giving a formal definition of an adjoint we can note that when the operator is represented by a matrix, the adjoint is constructed by interchanging rows with columns and replacing each number in the matrix by its complex conjugate – that is, each occurrence of $i$ is replaced by-$i$. A real number is equal to its complex conjugate. In quantum theory physical properties are represented by operators that are identical with their adjoints; these are known as *Hermitian* operators. It is a key feature of these operators that they have only real eigenvalues, even if complex numbers occur in the matrices. Thus while the operators used in quantum theory typically involve complex numbers, the eigenvalues of these operators – which represent physical quantities – are always real numbers.

Consider two quantum-mechanical operators $A_{op}$ and $B_{op}$. For some choices the two operators commute; for other choices they do not commute. Moreover, two operators commute if and only if they have the same eigenstates. Suppose $A_{op}$ and $B_{op}$ do not commute. After we measure A on a system, the system will be in an eigenstate of $A_{op}$; measuring B on this new system will put it into an eigenstate of $B_{op}$. But if we measure B first the system will initially be put into an eigenstate of $B_{op}$, then into an eigenstate of $A_{op}$. Since non-commuting operators have different eigenstates, changing the order of the operations will leave the system in a different state; this will not happen if $A_{op}$ and $B_{op}$ have the same eigenstates. We have here a mathematical reflection of the indeterminacy principle. Operators that share eigenstates commute and there is no indeterminacy relation between them; operators that do not share eigenstates do not commute and there is an indeterminacy relation between them.

Historically quantum theory began with the notion that certain physical quantities occur only in discrete units. As quantum theory developed it also came to deal with continuous quantities, so it would be a mistake to now define quantum theory in terms of that original idea. A more accurate general description of quantum theory is in terms of the replacement of classical parameters by operators. QFT carries this procedure one step further: the state functions of quantum theory are also replaced by operators. This is known as *second quantization*.

### *A3 Invariance*

Suppose you and I are using different stop watches to find the time between two events. Our watches run at the same rate, but when I reset my watch it always reads "2 seconds" while your watch resets to zero. As long as we are interested in the time *difference* between two events, we both get the same result. If we label the events *a* and *b*, you get $a - b$ while I get $(a + 2) - (b + 2) = a - b$. In other words, the time between two events is *invariant* with respect to uniform addition of some constant to each individual time measurement.

Rotation of a set of axes is analogous, although more complex in detail. If I locate two points by their *x* and *y* coordinates $(x_1, y_1)$ and $(x_2, y_2)$ I can use the Pythagorean theorem to calculate the distance between them. Now suppose I rotate the axes, keeping the origin unchanged.[53] The coordinates I read off my new axes will be different than those on the old axes, but if I use these new values to calculate the distance I get the same result as before: distance is invariant with respect to a rotation of the axes.

Now consider a simple example from calculus. If I add a constant *c* to *s* and then take the derivative of $s + c$, the result is the same as we get from taking the derivative of *s* alone: $D(s + c) = Ds$.[54] Thus differentiation of *s* is invariant with respect to addition of a constant to *s*. Another example of invariance is provided by the non-linear function $f(x) = x^2$. Replacing *x* with $-x$ leaves $f(x)$ unchanged; this holds for any even power of *x*, or any

function that consists of a sum of even powers of $x$. A parallel point holds if we add 360° to $\theta$ in $f(\theta) = sin\theta$, where $\theta$ is an angle. In all these cases, a different substitution need not leave the expression invariant.

Invariance plays a major role in mathematical physics. When two descriptions of a situation are related by some operation, properties that are invariant with respect to that operation give the same result in either case. Thus we can view the descriptions as alternative ways of describing the same properties. Invariance of distance with respect to rotation of the axes provides a prototype of this case. One of the axioms of the special theory of relativity – known as *the principle of relativity* – states that laws of nature are invariant with respect to all frames of reference moving relative to each other with constant velocity. If we find that some phenomenon is not invariant in this way, then we have not yet reached a law of nature. One key outcome of relativity is that these invariants are located in different properties than they seemed to be according to classical mechanics. For example, in Newtonian mechanics a particle's mass is invariant with respect to velocity, but this is not the case in relativity physics; we must move to a more complex property involving energy and momentum to find an appropriate invariant.

The pervasive occurrence of *conservation laws* is a manifestation of the importance of invariants in physics. Classical laws such as conservation of energy or mass specify quantities that are invariant in certain circumstances – such as in an isolated system. Particle physics introduces several new conservation laws.

### *A4 Symmetry*

Operations that leave some property invariant are described as *symmetries*; an operation may be a symmetry with respect to some properties of an item, but not with respect to others. This is a fairly straightforward extension of the everyday notion of symmetry. We can, for example, describe the bilateral symmetry of a plane figure in terms of operations that leave a property invariant: First we introduce a pair of mutually perpendicular axes in which the y-axis coincides with a figure's axis of symmetry. Then if we interchange x and -x for every point on the figure, we end up with the a figure that looks the same as the original. In a familiar pattern (recall the discussion in Sec. 2.5), we now have a more general concept of symmetry with the everyday notion as one specific case. Let us generalize further.

Suppose I have a rectangular book and an outline of that book on my desk; some operations on the book will allow it to fit back into the original outline. Clearly operations such as moving it into another room or burning it will not do. We define the *center of the rectangle* as the point at which the two diagonals meet. Any rotation of 180° around this point is a symmetry operation. (For a square, rotations of 90° are symmetries; for a regular hexagon, rotations of 60° will do.) In 3D space, 180° rotations around each of the lines that connect the midpoints of opposite sides are symmetries.

Next consider rotations of a circular disk – not the disk with the mark discussed above, but an undifferentiated disk. Every rotation about the center of the disk is a symmetry. For present purposes I want to draw one contrast between this case and the rectangle. Restricting ourselves to rotations in a plane around an object's center, in the case of the book all the symmetries are rotations of some integer times 180°; these are known as *discrete symmetries*. In the case of the disk, any rotation is a symmetry. These are *continuous symmetries* and are the only kind we will consider when we tie this discussion up to QFT.

Since symmetries always involve some property that is conserved, one might suspect a connection between symmetries and conservation laws. The connection, established by mathematician Emmy Noether, is known as *Noether's Theorem*: roughly, in a dynamical system, every continuous symmetry (of the Lagrangian density, which is discussed in the main text) implies a conservation law. Symmetries and related conservation laws are central to SM.

### *A5 Groups*

Consider sets of operations, which we can refer to as A, B, C, etc. The juxtaposition AB means that two operations are carried out in succession: B first then A on the result of B (operator sequences are read from right to left). To make the discussion more concrete suppose we are operating on our unmarked disk. Each operation is a turn by a definite number of degrees, where operation A consists of turning the circle clockwise by $a°$. The set of all rotations of a circle form a mathematical structure known as a *group*.[55] A group is any set of operations that has the following four properties.

Closure: AB = C, where C is a member of the set.

Successive application of two operations is equivalent to another operation in the set. In our example, C will be a single turn of $(a + b)°$.

Associative: (AB)C = A(BC).

The result of applying A to the result of the sequence BC is the same as applying the result of AB to the result of C. In both cases our example will give a total turn of $(a + b + c)°$.

Identity operation (I): AI = IA = A.

For reasons that will appear momentarily, we include an operation that does nothing.

Inverse: $(A^{-1})$: $AA^{-1} = A^{-1}A = I$.

The inverse of an operation undoes the result of that operation. The inverse of a clockwise turn of $a°$ is a counterclockwise turn of $a°$. Thus the result of an operation followed by its inverse is the identity operation. Properly speaking A and $A^{-1}$ are inverses of each other; each is an equal member of the group. I is its own inverse since II = I.

As we have seen, the sequence AB need not have the same result as the sequence BA although it does in our example. Groups in which AB = BA are called *commutative groups*, (or *Abelian groups* after mathematician Henrik Abel). The identity operation commutes with every member of the group, and the sequence consisting of an operation and its inverse also commutes. Most – but not all – of the groups that are relevant to QFT are non-Abelian. Groups provide a powerful tool for studying symmetry; the symmetries that interest us are invariants of specific groups.

### *A6 Representations*

A group of linear operators can be represented by a set of square matrices that has the following property (I use "$\times$" for matrix multiplication and "$M_A$" for the matrix that represents operation A): if AB = C, then $M_A \times M_B = M_C$. Note three points. First, matrix multiplication is not commutative. Second, while each member of the group being represented must be correlated with a single matrix, more than one group member may be correlated with the same matrix. Indeed, every group has the *trivial* representation in which we correlate every member with the identity matrix $M_I$. Since $M_I \times M_I = M_I$, the defining condition for a representation is met. (This representation has physical significance; see in Sec. 10.2.2.) When there is a one-one correspondence between a group of operations and the matrices of a representation, the representation is described as *faithful*. A group may have many representations that are neither trivial nor faithful. Third, there is another sense in which a group can have many representations – a sense that applies even to faithful representations. Nothing said so far implies anything about the size of the matrices in a representation ($3 \times 3$, $7 \times 7$, or what have you). Typically a group of operations will be represented by many sets of square matrices of different sizes, although all matrices in a particular representation are the same size. In addition, the faithful representation with the smallest matrices – known as the *fundamental representation* – plays a special role. For example, for SU(2) these will be $2 \times 2$ matrices and, more generally, for SU($n$) the matrices will be $n \times n$. Representations consisting of larger matrices are constructed from the fundamental representation. Some groups have two or more distinct representations in terms of matrices of a given size.

Sometimes a representation in terms of matrices of a particular size can be broken down into a set of representations using smaller matrices. Representations that can be broken down in this way are described as *reducible*; representations that cannot be reduced are called *irreducible* (IRR). Any reducible representation can be reduced to a set of IRRs such that the

sizes of the matrices in the reduction add up to the size of the original reducible matrix. In addition, if the elements of group G commute with the elements of group H, we can form the larger reducible *direct-product group*.[56] Each element of this new group consists of the product of one element from G and one from H; all possible pairs are included. Consider, for example, SU(2) which has, among others, an IRR constructed out of $3 \times 3$ matrices. Two of these IRRs can be compounded into a reducible representation in terms of $9 \times 9$ matrices which can be decomposed into a set of IRRs giving: $\mathbf{3} \times \mathbf{3} = \mathbf{1} + \mathbf{3} + \mathbf{5}$ (this use of boldface is standard notation). Each of the groups on the right hand side of the equation is a different IRR of SU(2). The fact that a group can have many different IRRs will be of special concern to us.

### *A7 Generators*[57]

The transformations that concern us are continuous and *unitary* (discussed in Sec. 10.2.2). Any continuous unitary transformation can be written in terms of a set of Hermitian matrices called the *generators* of the group. To construct the matrix for a specific transformation in the group, each generator is multiplied by a real number, which is a value of a group parameter (e.g., angles for rotations), and the resulting matrices are summed.[58] In other words, the generators form the basis "vectors" of a vector space; every transformation in this space can be expressed as a linear combination of these basis vectors.[59] Consideration of how we determine the number of generators will help bring out the point of this account. As a first step let us examine square matrices in which all the elements are real. This space will have $n^2$ basis matrices. For example, each basis matrix could have 1 in just one slot, and 0 in every other slot. Every matrix in the set can then be written as a linear combination of these basis matrices. There are many different sets of basis matrices, but their number remains constant.

Now consider the number of generators in the groups that concern us. If each slot is occupied by a complex number we need two parameters per slot. (Recall that a complex number can be written as $a + bi$; real numbers are complex numbers with $b = 0$.) This suggests that we need $2n^2$ generators, but the limitation to Hermitian matrices yields two constraints. First, the main diagonal (from top left to bottom right) must contain only real numbers, so only $n$ parameters are needed for the diagonal. Second, there are $n^2$ – n off-diagonal slots containing $2(n^2$ – n) parameters. But in a Hermitian matrix corresponding elements across the diagonal are complex conjugates of each other; thus only half of these are independent. So we have $n^2$ - $n$ independent off-diagonal elements, plus $n$ independent diagonal elements, giving a total of $n^2$ independent generators. For the *special unitary groups* we have one more constraint: the determinant must be 1. The additional constraint reduces the number of independent generators by 1 so there will be $n^2 – 1$ independent generators. Let us apply this approach to the $3 \times 3$ matrices

that represent SU(3). The following procedure will not yield generators in the form in which they are actually used in QCD (see Sec. 2.3.3), but it will give the correct number of generators (R 91–93). There are six complex off-diagonal elements in a $3 \times 3$ matrix, but only three of these are independent, so let us focus on the slots above the main diagonal. We can construct three basis matrices by putting 1 in one of these slots, and letting all other slots be 0. We can construct another three by putting $i$ in one of these slots, the rest again all being zero. On the diagonal, where all entries are real, only one parameter is required per slot. In general we could proceed by considering the three cases in which we have 1 in one slot of the diagonal and 0 in the remaining slots. However, the requirement that the determinant be 1 implies that the trace (the sum of the diagonal elements) is zero. So only two of the three possibilities just mentioned are independent, yielding a total of eight generators.

Next consider why the generators are important for physical theory. First, the generators of a continuous group form the basis of an IRR for that group. Second, in quantum theory operators express symmetries of a system if and only if they commute with the system's Hamiltonian. Each generator meets this condition. Third, in accordance with Noether's theorem each generator is associated with a conserved quantity – that is, each generator yields a conservation law. Fourth, a continuous current corresponds to each conservation law, and in QFT a continuous current requires a particle. These are the particles that mediate the fields, and the number of generators thus determines the number of field particles: one for the U(1) symmetry of QED, three for the SU(2) symmetry of WI, and eight for the SU(3) symmetry of QCD.

Note one further point about these groups and their IRRs. In the application of SU(2) to WI we are dealing with a group of $2 \times 2$ matrices that act on 2-component vectors (e.g., doublets consisting of a massive lepton and a neutrino). These $2 \times 2$ matrices also constitute a vector space with three basis vectors – the generators. There is another IRR of SU(2) consisting of $3 \times 3$ matrices that acts on this vector and describes how the generators mix under "rotations." This is analogous to the mixing of vector components when we rotate the axes. I do not consider this case in our main discussion, but two points are worth noting. First, the fact that the generators mix is directly related to the point that the WI bosons carry weak charge, and thus enter into weak interactions. Second, the existence of these interactions is a consequence of the fact that SU(2) is non-Abelian. There is a parallel situation for SU(3). The fundamental IRR of this group consists of $3 \times 3$ matrices that act on the 3-component vectors composed of the quark colors and describe how these colors mix in a specific quark. SU(3) has eight generators, and there is also an $8 \times 8$ IRR of SU(3) that acts on a vector consisting of the generators, with consequences analogous to those for SU(2).

# 11 Conceptual Change, Incommensurability, and Progress

> I intend to discuss how progress leads to confusion leads to progress and on and on without respite.
>
> (Pais 1986: 4)

I have argued throughout this book that human history displays a widespread introduction of new concepts and abandonment of older concepts as new thoughts appear over time and older ways of thinking are abandoned. While I have focused mainly on the development of science, the point applies to new technologies, new forms of social organization and economic activity, and I suspect every other area of human endeavor. Considerations of conceptual change raise problems about incommensurability, but this term has been used to cover a variety of situations, thus it is important to be clear on just what problems are relevant.

Let us begin with a theme from Kuhn that becomes more focused in his later writings: there is a problem of *translation* when we compare conceptual frameworks.[1] It will help pin down the exact nature of the problem if we recall why no such problem arises on the prevailing view in philosophy of science circa 1962. On that view there is a set of basic concepts that are, sufficiently closely for our purposes, universal (see Secs 1.7, 3.4, 3.5), but most scientific and everyday thought takes place in terms of auxiliary concepts that we construct out of basic concepts. In the case of auxiliary concepts we encounter all the reasons for conceptual change that we have considered because the construction of *projectible* auxiliary concepts – concepts that can be used to formulate sustainable generalizations – is a fallible, creative process. However, there is no translation problem because all empirically significant auxiliary concepts can be translated without loss of content into basic concepts. Once this translation has been carried out, differences between competing theories, and the appropriate tests for deciding among them, are clear. The translation problem arose when philosophers challenged the existence of this basic framework and the possibility of such translations. Quine took one step in this direction when he rejected the analytic-synthetic distinction because translation of auxiliary concepts into basic concepts is supposed to occur by means of analytic

propositions. Such conceptual reduction is Quine's second dogma. Kuhn and Feyerabend attacked the prevailing view from a different direction that is independent of the status of the analytic-synthetic distinction. They challenged the existence of a universal basic framework and proposed that meaning – at least for key terms in a scientific language – is determined by interrelations among terms. This thesis eliminates a single framework for all translations; it seemed to many that this view raises a deep problem about how we can compare competing theories and make a reasoned choice between them.

We can further clarify what is at issue by considering two possible ways of avoiding the translation problem without invoking a universal framework. One option is that competing claims can be adequately stated in one of the competing frameworks. We have seen that new fundamental theories typically introduce new concepts, so it is clear enough that we will not be able to express the new theory in the older framework. The key issue, then, is whether we can express an older theory in a newer framework without distortion and without biasing the choice. Kuhn denies that this is possible; I want to consider the reason for this claim in his later account of conceptual systems, which is better developed than his earlier remarks.

The centerpiece of Kuhn's later account is the notion of a *lexicon*.[2] This is a set of terms that are both interrelated and attached to experience in a particular way. Kuhn typically approaches these terms by commenting on how they are learned, but he also notes that this learning process is not necessary for acquiring a lexicon: "The consequences would be the same if, for example, the lexicon were a genetic endowment or had been implanted by a skilled neurosurgeon" (1989: 66, n. 11). Thus I will abstract from the learning process and focus on its presumed result. Kuhn continues his earlier practice of writing as if a conceptual framework is a language although he also expresses doubts about this metaphor: "By now, however, the language metaphor seems to me far too inclusive" (1989: 92). Kuhn also denies that he is concerned primarily with language in another respect – one in which language is not sufficiently inclusive:

> Throughout this paper I shall continue to speak of the lexicon, of terms, and of statements. My concern, however, is actually with conceptual or intensional categories more generally, e.g., those which may be reasonably be attributed to animals or to the perceptual system.
>
> (1989: 60, n. 2; cf. 1991b: 94, 1993: 229–30)

This is in accord with the view I have taken throughout the present book, and I will generally couch the discussion in terms of concepts, although I will slide into writing about language when quoting Kuhn and discussing these quotes.

A lexicon consists of a set of kind-concepts that get content both from the ways they relate to each other and to their characteristic instances. Kind-concepts are projectable so that their content includes "some generalizations satisfied by their referents" (1993: 230); these generalizations generate

expectations. When required to classify items into kinds all members of a conceptual community arrive at the same results, although they may do so on the basis of different expectations (1993: 239). In his last publication Kuhn distinguishes two types of generalizations: *nomic* generalizations are exceptionless, while *normic* generalizations admit of exceptions; Kuhn gives "liquids expand when heated" as an example of a normic generalization. This results in two types of kind-concepts. Those governed by normic generalizations are "the most populous part of the lexicon"; they fall into contrasting sets (1993: 239) and are governed by the *no-overlap principle*: their referents may not overlap "unless they are related as species to genus" (1991b: 92). Encounters with overlaps are generators of conceptual change:

> What should one have said when confronted by an egg-laying creature that suckles its young? Is it a mammal or is it not? . . . Such circumstances, if they endure for long, call forth a locally different lexicon, one that permits an answer but to a slightly altered question: "Yes, the creature is a mammal" (but to be a mammal is not what it was before). The new lexicon opens new possibilities, ones that could not have been stipulated by the use of the old.
>
> (1989: 72)

Before Kuhn drew the nomic/normic distinction he focused on concepts governed by exceptionless generalizations; these provided some of his most developed examples. Since part of the content of these concepts derives from generalizations, such concepts come in sets of related concepts rather than contrasting sets. Kuhn gives a particularly detailed discussion of force, mass, and weight in Newtonian physics, where he emphasizes that Newton's gravitation law and three laws of motion are included in the content of these concepts. Newton's first law is central to the concept of a force, since it specifies the only instance of force-free motion in Newtonian physics (1989: 68). Weight and mass involve Newton's second law and gravitation law. Kuhn illustrates the role of these generalizations by imagining two subsets of a community with different views of these laws.[3] One group considers the second law a necessary truth and the gravitation law an empirical generalization; the other group reverses this viewpoint. This difference will be neither apparent nor significant in normal practice, but it will come out if they encounter an empirical challenge that requires a revision in one of these laws. Kuhn seems to suggest that this would lead to disagreement on how to proceed, but no deep conceptual change. A challenge that requires altering *both* will force conceptual change (1989: 73–74).

Generalizations relating these concepts give only part of their content. In addition, part of their content is determined by the way we identify instances. Kuhn notes that in the Newtonian framework weight is a relational concept – a measure of a local force; mass is not relational (recall Sec.

9.4). He relates this distinction to the advent of the spring balance, which did not exist before Newton's time (1989: 69–70), and which can give different values for the same object at different locations; this cannot occur with the older pan balance. But use of the spring balance requires two additional laws – Newton's third law and Hooke's law – which are therefore also implicated in WEIGHT, its distinction from MASS, its relation to FORCE, and thus its relation to MASS via the second law. Given the many examples of this sort in the present book, we need not belabor Kuhn's point: There is no term in the Newtonian framework that translates the pre-Newtonian term "weight," and similarly for other key terms involved in the comparison. For similar reasons, the Newtonian terms "force" and "mass" are not translatable into relativity since Newton's version of the second law does not hold in this later theory (1989: 74; 1983: 44). Kuhn maintains that such *untranslatability equals incommensurability*, where translation is to be understood as "a quasi-mechanical activity governed in full by a manual which specifies, as a function of context, which string in one language may, *salva veritatae*, be substituted for a string in another language" (1989: 60). In the same passage Kuhn acknowledges that such straightforward substitution is "not quite the activity of professional translators"; we will return to this point. Kuhn also considers a variation on this theme: the possibility that we may work in an expanded framework which includes both the older concepts and their successors. He notes that historians work in such a framework, but holds that it cannot describe a coherent world, and that its use requires constant attention to which of the incompatible parts of the framework is being used at a particular time (1989: 74–75, cf. 1983: 54). Such a framework will not help with the problem of theory choice, but it suggests another approach – one that Kuhn does not discuss.

Perhaps we can find a framework that is neutral with respect to specific competitors, and into which the relevant parts of the competitors can be translated. I think it is clear how Kuhn would respond to this option: terms would enter into new relations to each other and to their referents in this framework, and thus would not be *translations* of the older terms. Moreover, depending on the exact differences induced, we might well end up with a biased choice. This "neutral" framework may have (or come to have) competitors, and translation into one of these competitors might yield a different outcome. So this option also fails to provide a neutral basis for evaluating incommensurable frameworks.

All of this is in accord with TC, which underlines an additional aspect of the generalizations implicated in conceptual content. Recall a key theme of SSR: that logic and observation are not *sufficient* for choosing between scientific theories; additional methodological criteria are required, and these are internal to specific frameworks. These additional criteria are just the GAs that, according to TC, provides part of the content of descriptive concepts. While this theme fell into the background in Kuhn's later writings, it is reflected in the role that generalizations play in the lexicon. But this theme brings out a *second incommensurability problem*: different frameworks

include different criteria for evaluating scientific theories, criteria that come into play in order to close the gaps that remain once we have appealed to logic and empirical evidence.[4] As a result, any attempt to approach theory choice from within one of the competing theories introduces an additional source of bias into the evaluation process.

There is a *third incommensurability problem* that also emerged circa 1962, and that has been central to discussions of theory choice: incommensurability of the "data" used for evaluating competitors. It is again important to set the issue in the context of the view that prevailed at the time: Terms of the basic language get their meaning from a direct correlation with observables that occur to our senses independently of any beliefs we hold. Thus observables play a double foundational role: they are the ultimate source of meaning for all terms in the language, *and* they provide the touchstone for theory choice since all disagreements concerning empirical matters can be expressed as disagreements about what observables will occur under specified circumstances. Disputes can thus be settled by the relevant observations. Hanson (1958), Kuhn, and Feyerabend challenged the existence of such theory-free data arguing, instead, that observation is *theory-laden*. However, as discussion developed several different claims were included under this rubric, some sustainable and some not sustainable. In Brown 1995 I discuss six versions of this claim, but only one need be considered here: For a body of sensory experience to be relevant to the evaluation of a theory, that experience must be described in terms of the concepts of that theory. A central theme from logical empiricism will underline the import of this claim. For logical empiricists the philosophical problem of theory evaluation concerns logical relations, and logical relations hold between propositions. As a result, we do not confront theoretical predictions with experience, but with *propositions that describe experience*.[5] But, it was argued, qualitatively identical perceptual experiences can receive quite different descriptions when approached from different conceptual systems. Another theme from latter-day logical empiricism will underline the import of this claim. By 1962 the prevailing account of theoretical concepts was in terms of axiom systems and correspondence rules, where empirical content flows from experience to a theoretical structure (Sec. 3.5). The challengers began from this image, but reversed this relationship, maintaining that meaning flows from theory to sensation. As a result, they argued, competing theories that embody fundamentally different conceptual systems will not be confronted with the same evidence statements. In other words, there are no theory-neutral observation reports that provide an independent touchstone for comparing fundamentally different theories. Shortly, I will consider some ways in which TC provides a new perspective on these issues, but before doing so I want to integrate several other themes that have played an important role in the literature.

Scientific theory choice deals with frameworks that are genuine competitors. This requires that there be some items for which the competing theories offer genuinely different accounts. We need not choose, for example, between

the standard model in high-energy physics and the rules of baseball as accounts of the fundamental constituents of the universe.[6] This constraint does not require agreement on the complete domain under consideration; competing frameworks can disagree on what is to be included in a single domain. This occurred in the dispute between Aristotelians and Galileo about the status of the celestial and terrestrial realms, and (more narrowly) in the disagreement between Descartes and Newton over whether change of speed and change of direction should be considered instances of acceleration. It does require that there be identifiable items that are common to the competitors. Fall of an object on a moving ship provided one such example for Aristotelians and Galileo. Note especially that Galileo's correct prediction might provide an opportunity for an opponent to wonder how Galileo had arrived at his result, and undertake to learn his approach. The shape of the planetary orbits provides another example for Descartes and Newton. Since they both required that their theories explain these shapes, Newton could argue that his theory gets them right while Descartes' theory cannot get them right. In other cases an older view might require that two phenomena be explained by the same account, while a later approach separates them. For Aristotle, falling and rising objects near the earth are treated as instances of the same phenomenon – motion to a natural place. For Descartes they are viewed as instances of quite different phenomena that receive different kinds of explanations (Sec. 9.3). But in spite of this bifurcation, Aristotelians and Cartesians both recognized that objects rise and fall, and agreed that physics must give an explanation of why these occur.

There is another respect in which an account of scientific theory choice must involve genuine competitors: they must have been in competition in the actual historical development of the subject. Kuhn generated considerable confusion by missing this point, although he eventually recognized it. In SSR Kuhn placed much emphasis on a comparison of the conceptual frameworks of Aristotelian and Newtonian physics but we saw in Ch. 9 that these were never genuine competitors; Newton's opponents were Descartes and, to a lesser degree, Brahe. For a long time Kuhn approached conceptual incommensurability in terms of his initial encounter with the phenomenon: his first attempt to understand Aristotelian physics from the perspective of his own training in Newtonian physics. He prefaces one report of this experience with this remark: "The road I traveled backward with the aid of written texts was, I shall simply assert, nearly enough the same one that earlier scientists had traveled forward with no text but nature to guide them" (1987: 15). Two years later he was clear that this was a mistake: "In recent years I have increasingly recognized that my conception of the process by which scientists move forward has been too closely modeled on my experience with the process by which historians move into the past" (1989: 87). The gaps that must be closed in typical cases of scientific theory choice are much smaller than many gaps that a historian must cross.

However, while the historian's problem is different from the scientist's, it is a genuine problem. One approach to getting a grasp on a superseded framework is to work back through the history following the links through changes that took place. TC provides a guide to what we should look for in following such a path. A problem similar to the historian's occurs when we consider two cultures that developed largely independently of each other. Sometimes such frameworks come into competition, as has happened with Chinese and Western medicine (cf. Wang 2002). But here too *competition* occurs only because of mutually recognized areas of overlap. Both frameworks recognize the existence of disease states that they seek to ameliorate and agree, in at least some instances, that a particular set of signs and symptoms (cough, abnormal pulse, loss of blood) are indicators of disease.

One more theme must be integrated into this discussion: realism. The terms "realism" and "scientific realism" are used to describe many issues; here I am interested in just two of these (cf. Brown 1990):

> Whether discovery of a correct description of features of reality as they are in themselves is a pursuable goal of scientific research . . .;
>
> (R1)

and:

> Given an affirmative answer to R1, whether the fact that a particular theory prevails in a given competition implies that it is a better candidate for this realist aim.
>
> (R2)

Kuhn has consistently held that the history of science does not provide a series of closer approaches to a correct description of any aspect of nature, and that the aim itself is confused:

> I am not suggesting, let me emphasize, that there is a reality which science fails to get at. My point is rather that no sense can be made of the notion of reality as it has ordinarily functioned in philosophy of science.[7]
>
> (1991c: 115)

The two issues are not independent of each other since an affirmative answer to R2 implies an affirmative answer to R1. In addition, many hold that a negative answer to the R2 implies a negative answer to R1 since there is no point to claiming that we are pursuing a goal unless we have grounds for assessing how well we are doing in that pursuit. Indeed, we may ask how we can assess whether we are approaching a goal in the absence of an account of that goal. Yet in the present case we do not have such an

account – it is what, on a realist view, we are seeking. Later in this chapter I will offer a limited response to this challenge. For the moment I want to note two points.

First, challenges to scientific realism do not arise *only* because of the absence of a universal framework for evaluating competitors. Even if we had such a framework, it would not guarantee that we can pursue the realist goal; this depends on how universality is achieved. For example, in a Kantian approach universality is guaranteed by concepts that are necessarily shared among (at least) human cognizers. But Kant's approach to justifying this universality claim undermines the realist goal. Consider another example. In the logical empiricist framework universality is a consequence of shared sensory experience, but this does not get us to any reality beyond our senses. Logical empiricists recognized this and were generally anti-realists. Here is Hempel's classic statement of this view: "Scientific systematization is ultimately aimed at establishing explanatory and predictive order among the bewilderingly complex 'data' of our experience, the phenomena that can be 'directly observed' by us" (1965: 177). He acknowledges "the remarkable fact" that postulation of non-observables has led to "the greatest advances" in pursuit of this goal, and views this as a puzzle. Hempel's solution of the puzzle leads him to expand his account of the aim of science from just establishing "deductive connections among observation sentences" to also establishing "inductive explanatory and predictive" connections plus "systematic economy and heuristic fertility" (1965: 222). Learning about items we cannot sense is not included as an aim. Van Fraassen's anti-realism (1980) is a more sophisticated version of the same idea. One might think that the existence of a universal framework is at least necessary for realism, but this is not correct; I will return to this claim shortly.

Second, a theory's winning a competition on the basis of agreed upon evidence and standards is not sufficient to justify the claim that the winner is a better account of the domain in question than any of the losers. They may all be equally poor in this respect, and later competitions may lead to replacement of the current winner by another theory that embodies a very different conceptual system. Thus we might have a series of theory replacements running from $T_1$ to $T_n$, where $T_1$ and $T_n$ have little in common, even though there is a great deal in common between any two adjacent members of the series. Kuhn makes this point when discussing an historical narrative describing a series of changes in scientific belief: "By the end of the narrative those changes may be considerable, but they have occurred in small increments, each stage historically situated in a climate somewhat different from that of the one before" (1991c: 112). This leads Kuhn to distinguish between questions concerning the rationality of *belief* and the rationality of *change of belief*. In effect, the former is the question of realism, the latter that of theory change. In his later work Kuhn clearly holds that in science change of belief is rational. For example, after reviewing the traditional demand for observations that are neutral with respect to all beliefs, Kuhn writes:

> From the historical perspective, however, where change of belief is what's at issue, the *rationality* of the conclusions requires only that the observations invoked be neutral for, or shared by, the members of the group making the decision, and for them only at the time the decision is being made. By the same token, the observations involved need no longer be independent of all prior beliefs, but only of those that would be modified as a result of the change. The very large body of beliefs unaffected by the change provides a basis on which discussion of the desirability of change can rest. It is simply irrelevant that some or all of those beliefs may be set aside at some future time. To provide a basis for rational discussion they, like the observations the discussion invokes, need only be shared by the discussants.
>
> (1991c: 113)

At this point it looks as if incommensurability is *irrelevant* for questions of theory choice. Kuhn still maintains that there is incommensurability between these competing views since, as noted above, he holds that incommensurability equals failure of quasi-mechanical, truth-preserving translatability (1989: 60). "Incommensurability thus becomes a sort of untranslatability, localized to one area or another in which two lexical taxonomies differ" (1991b: 93). But in these later papers Kuhn insists that translation is *not required* for communication and rational theory choice. A different cognitive process – which he describes as *interpretation* and as *language learning* – is required. Discussing the case of the historian (which, we have seen, often involves larger conceptual gaps than that of the scientist involved in actual theory choice), Kuhn writes:

> Faced with untranslatable statements, the historian becomes bilingual, first learning the lexicon required to frame the problematic statements and then, if it seems relevant, comparing the whole older system (a lexicon plus the science developed with it) to the system in current use. Most of the terms used in either system will be shared by both, and most of these shared terms occupy the same positions in both lexicons. Comparisons made using those terms alone ordinarily provide a sufficient basis for judgment.
>
> (1989: 77)

A few years earlier Kuhn wrote:

> Translation is, of course, only the first resort of those who seek comprehension. Communication can be established in its absence. But where translation is not feasible, the very different processes of interpretation and language acquisition are required. These processes are not arcane. Historians, anthropologists, and perhaps small children engage in them every day.
>
> (1983: 53, cf. 1993: 238)

Kuhn also tells us that "anything which can be said in one language can, with imagination and effort, be *understood* by a speaker of another. What is prerequisite to such understanding, however, is not translation but language learning" (1989: 61). And, "with sufficient patience and effort, [one can] discover the categories of another culture or of an earlier stage of one's own" (1991a: 220). Kuhn has also backed off from his metaphor of a scientific revolution as a gestalt shift (although this may still be an appropriate analogy for particular historians). "To speak, as I repeatedly have, of a community's undergoing a gestalt shift is to compress an extended process into an instant, leaving no room for the microprocesses by which the change is achieved" (1989: 88). Kuhn even clams that the possibility of significant comparisons of competing modes of scientific practice "was never for me in question" (1983: 55).

Leaving aside questions of whether Kuhn is correctly reporting his earlier views, once we have acknowledged the possibility of mutual understanding, there is no residual problem of the rationality of theory comparison. This applies not only to the use of different concepts, but also to differences in evaluation standards and in conceptualization of the data. In all these cases there may be genuine disagreements – disagreements that are more severe than those acknowledged by logical empiricists – but there is no reason why failures of communication need occur. To be sure, such failures may occur among those who do not approach the problem with sufficient effort, patience, and imagination, but there is no problem of theory comparison that transcends rational mediation.

It is particularly important to keep in mind that innovators and early adopters of a new framework are often masters of the previous view and thus able to find means of catching the attention and interest of those they would convert. I have already noted Galileo's use of the rock dropped on a moving ship as one example. But note also that a major task in *Dialogue* is to show that Aristotelian arguments against the motion of the earth commit specific logical fallacies that would be familiar to Aristotelians. (Finocchiaro 1980 provides detailed analyses of many of these arguments.) We have seen that Newton argues directly against Descartes, bringing out specific problems with Cartesian physics and offering testable alternative solutions. Newton's editor Cotes reiterates and expands these arguments. In a similar way, in presenting special relativity Einstein starts off from two well-known problems: a problem in the interpretation of Maxwellian electrodynamics, and a problem of consistency between two postulates that others had already found attractive. He resolves these problems while working within the established mode of mathematical physics, and in a way that preserves Maxwell's equations and explains why Newtonian physics – which is superseded – works as well as it does. From this perspective, Kitcher's (1978) account of *reference potential*, which allows for the flexible identification of some items countenanced by a later theory with items invoked by a predecessor, is one technique that can be used by both historians and innovators to build bridges between conceptual systems.

A crucial feature of this approach to theory comparison is that it depends on human cognitive abilities, and thus introduces *scientists* into an account of theory evaluation in addition to abstractly formulable linguistic structures. This introduction of scientists into philosophy of science was a central theme of SSR, although it dropped into the background in much of Kuhn's later work. I want to review the role this theme played in SSR. Again, the discussion is best set in the context of the situation in 1962.

Recall the logical empiricist distinction between *context of discovery* and *context of justification*, where the latter deals with logical relations between observation statements, on the one hand, and those generalizations and theoretical claims that go beyond observation statements, on the other. Logical empiricists held that philosophical analysis of the epistemic status of science is concerned only with these logical relations. Any considerations of the psychology of the actors in the development and acceptance of scientific claims were held to be irrelevant to epistemic evaluation and were relegated to the context of discovery. The context of discovery was not rejected as unimportant, but only as irrelevant to the task of *philosophical* analysis. The psychology of discovery, for example, is a legitimate field of scientific research. But the evaluation of its results depends (it was held) on their meeting the appropriate criteria for the evaluation of scientific theories, criteria that must be established independently of any particular scientific results.

In SSR's introductory chapter Kuhn suggested that he would be challenging the distinction between the two contexts (8–9). This challenge is captured especially in Kuhn's thesis that observation and logic are not sufficient to account for revolutionary theory change, and his attempt to close the gap by taking into account aspects of the psychology of scientists and social interactions in scientific communities.[8] Many rejected this as an inappropriate intrusion of psychology and sociology into epistemology,[9] but Kuhn's move is more accurately interpreted as a recognition that such psychological and social factors are relevant to epistemology. In this regard SSR is continuous with the naturalistic approach to epistemology that was emerging at that time. Put differently, the claim is that in order to understand (and evaluate) scientific theory choice we must attend to the scientific *process* as well as the scientific *product*. But the relevant aspects of the scientific process are not psychological quirks of the scientists involved; they are the skills that scientists develop through their training and continuing scientific work. (Recall the discussion in Sec. 5.8.) Skills are lodged in individual scientists, but they are no more "subjective" in a pejorative sense than is the ability to drive a car. It is these skills that Kuhn is invoking when he writes of the need for – and availability of – patience and imagination in understanding a competing theory. The point, then, is that *human* theory evaluation is dependent on human psychology and we cannot give an adequate account of this process without taking human psychology into account. Note especially that this dependence on our

psychology is not just a limiting constraint on the prospects of human knowledge. It is also a feature that *enables* the development of knowledge. Our ability to respond with intelligence and sensitivity overcomes the gaps left by failures of translation – gaps that are inevitable given that early conceptualizations are often quite inadequate, and that scientific progress requires both the introduction of new concepts and the elimination of older concepts which no longer have a role to play in the researcher's repertoire.

We are, however, not completely finished with incommensurability. Returning to Kuhn's late works, the significant impact of incommensurability appears in the evaluation of beliefs – that is, the question of realism. If the development of science requires the introduction of new concepts that are not translatable into existing concepts, then it seems impossible to assess whether successive frameworks are moving closer to a correct description of items in their domain. Commenting on "the question of science's zeroing in on, getting closer and closer to, the truth," Kuhn contends that such claims are "meaningless," and that this "is a consequence of incommensurability" (1993: 243–44). The basis for this claim lies in the systemic character of scientific concepts, which Kuhn treats as incompatible with realism. In what I take to be his clearest statement of this position, Kuhn begins with the untranslatability of a lexicon into its successor. As a result, the earlier statements are "immune to an evaluation conducted with [the later] conceptual categories." But, he immediately adds:

> The immunity of such statements is, of course, only to being judged one at a time, labeled individually with truth-values or some other index of epistemic status. Another sort of judgment is possible, and in scientific development something very like it repeatedly occurs.
>
> (1989: 76)

The passage (quoted earlier) on becoming bilingual follows, and Kuhn continues: "But what is then being judged is the relative success of two whole systems in pursuing an almost stable set of scientific goals, a very different matter from the evaluation of individual statements within a given system." Elsewhere he writes:

> Evaluation of a statement's truth-values [*sic*] is, in short, an activity that can be conducted only with a lexicon already in place, and its outcome depends upon that lexicon. If, as standard forms of realism suppose, a statement's being true or false depends simply on whether or not it corresponds to the real world – independent of time, language, and culture – then the world itself must be somehow lexicon dependent. Whatever form that dependence takes, it poses problems for a realist perspective, problems that I take to be both genuine and urgent.
>
> (1989: 77)

I suggest that, with the help of some ambiguity in the notion of evaluation being *dependent* on the lexicon, Kuhn has confused three different issues. First, the *conceptual content* of a theory is determined by the system in which it occurs; TC is in complete agreement on this point. Second, familiar arguments that Kuhn does not give here, and that I will not recount, strongly support the view that only entire theoretical systems are subject to epistemic – in particular, empirical – evaluation. But neither of these speaks to the third issue: what it *means* to attribute truth-values to individual statements in a system. Suppose we have a scientific theory that embodies a conceptual framework, and that the evidence supports this theory. It makes good sense to hold that each of the sentences constituting the theory is true – where this means that it correctly describes the items it speaks about. What we must avoid is the all-too-common confusion between the meaning of a claim, the evidence for it, and what it means to say that a claim is true. Note especially that while the evidence may support only the theory as a whole, this does not block attribution of truth to the individual sentences in that theory. Moreover, if a theory is replaced, a bilingual historian or scientist aware of the new evidence can – thinking in terms of the older theory – conclude that certain claims in the theory are false, and explore which claims carry over to the new theory, or have close successors in that new theory. *Furthermore*, there is no reason why realism – understood as a quest for the correct account of things-in-themselves – must be tied to the view that scientific knowledge is apportioned to individual sentences. There is no bar to a version of realism which holds that a conceptual system is the minimum unit of correspondence. Thus even if one rejects attribution of a truth-value to individual sentences, a robust form of scientific realism remains a possibility. *Finally*, rejecting the claim that science pursues correct accounts in a linear fashion is not the same as rejecting the claim that science pursues correct accounts, nor does it eliminate all grounds for thinking that, as science develops, our ability to pursue this goal improves. I want to consider such an alternative approach.

We must not forget a central feature of scientific research that is well attested in the sections of Ch. 2 that deal with empirical science: the role of experience in driving research. This occurs in two respects. First, much research is directly elicited by experience. In recent decades many have argued for a central role of theory in driving research. I think that this is basically correct, but the point was often overstated because it emerged as part of a critique of logical empiricism which focused mainly on experience, giving theory only a secondary role. By now we can see that experience and theory are more nearly equal partners in generating scientific problems. Consider a well-worked example: at the beginning of planetary astronomy the wandering motions of the planets need not have seemed problematic; they could have just been listed among the observed facts. It required the hypothesis that all true planetary motions are circular to generate theoretical research. But let us not forget the other side. The hypothesis of circular

motion would not have generated a research problem without the observation of celestial items that appear to violate this hypothesis. The point is especially dramatic when nature impinges on researchers in unexpected ways – such as in the initial observations of sperm and radioactivity. To be sure, none of these phenomena would have seemed surprising without some theoretical background that indicated what to expect. But it was the new observations that drove further research. In more or less dramatic ways, this same interplay holds throughout the history of scientific innovation. One side or the other may dominate in a particular case, but both are required to generate new lines of research.

But while the theoretical and experiential sides may be roughly equal in generating new research, the decision to accept a theory depends ultimately on its ability to handle the results of our interactions with nature. This takes us to the second role of experience noted above: it is the final arbiter of scientific acceptability. Sometimes, in a well-developed science, a piece of research may be primarily driven by theoretical considerations. Dirac's determination to construct a relativistically correct quantum theory that uses only first derivatives is as clear an example of successful theory-driven science as we are likely to find. The quest carried him through a significant mathematical innovation – introduction of square matrices where previously standard practice would require numbers or vectors – and the introduction of a new fundamental concept – antimatter. But the work also had empirical consequences – some already known, some new. It is only because of its empirical successes that the theory prevailed. In the face of empirical failures, any theoretical principle – circular celestial motions, conservation of energy, direct proportionality of force and acceleration, stability of species, total separation between space and time – can be reconsidered and replaced no matter how well founded it may once have seemed in experience and reason.

However, a proper understanding of empirical evidence in science requires another break with the classical empiricist tradition – one that, I think, Kuhn never fully made.[10] The epistemic significance of empirical evidence does *not* derive from its dependence on our senses. Rather, we pursue evidence pertaining to presumed items in the world by attempting to interact with those items. We evaluate claims about items in a domain by attempting to probe them in various ways, and the greater the variety of probes at our disposal, the richer the body of evidence we have for these claims. The development of instrumentation – beginning with Galileo's use of the telescope and exploding in the twentieth century – has greatly increased the variety of ways in which we probe nature; it has also increased the precision of the results of these probes.[11] On this view our senses play a pragmatic role – not a foundational role – in gathering empirical evidence: our senses are the means by which information about items in the world enters into our cognitive systems. But the information does not reduce to the sensations that provide this access.

Put differently, we seek descriptive concepts that have referents – items that exist independently of our theories. These referents form the subject

matter of our theorizing; they are the items we seek to learn about when we construct and test theories. With this in mind we can understand one main attraction of the appeal to stable referents as the means of eliminating incommensurability. On one version of this view, we pick out items (such as Mars) or kinds of items (such as gold) that remain the focus of research as our ideas about them change. Even if the concepts we use to think about these items undergo radical change, it is clear that we are discussing the same things throughout. No fundamental problem arises about comparing successive theories because they are all theories of the same items. There is no doubt that research of this sort takes place, and in such cases evidence derived from interactions with the referents provides the main grounds for evaluating competing theories. Moreover, according to TC, as long as the means of identifying such items remain stable, these items provide a bridge across conceptual systems. This can hold even as those means of identification are enriched – as occurred, for example, when electrical resistance and spectroscopy came to play a role in identifying chemical elements. But this is not the only – or even the dominant – mode of scientific research. It is not the kind of research that led to quarks, gluons, and weak-interaction bosons; nor is it the kind of research that drives the search for the Higgs boson, or Newton's identification of change of speed and change of direction as instances of the same phenomenon, or the limited unification of space and time in special relativity. Nor does it apply to cases such as the unexpected darkening of Becquerl's photographic plate that led to the discovery of radioactivity. The last example suggests a familiar variation on the notion that scientists study stable referents: that one introduces a new entity as whatever caused an observed phenomenon, and that research proceeds to seek out that cause. We are dealing, again, with a kind of research that does occur, but there is a great deal of research that does not fit this pattern either. It gives no insight into the introduction of isospin and its role in modern accounts of the stability of atomic nuclei, or into the role of the weak interaction in understanding why some nuclei undergo radioactive decay. Nor does it help us understand the development from the Aristotelian search for the cause of the continued motion of projectiles, to the (rather different) Cartesian and Newtonian claims that there is no cause.

Both of these approaches are strongest when the object of study is an individual item, but considerably weaker when it is extended to kinds of items. As we have seen (Sec. 6.1), our understanding of what items to classify as members of the same kind shifts over history. In Ch. 9 we examined some of the key steps from the view that rest and uniform motion are fundamentally different to considering them as instances of the same kind. Recall also the fate of the Aristotelian and Chinese elements – none of which are to be found on modern lists of the chemical elements. Nor are any of the modern chemical elements on these ancient lists – even though some of these elements were already familiar. The unification project discussed in Ch. 10 provides a supply of examples in which items that were put in distinct

classes at one stage of research were combined – or partially combined – into a single class as research proceeded. Meanwhile research into the structure of atoms has continued to multiply fundamental kinds, moving from a compact account in terms of just electrons, protons, and neutrons to the standard model which includes roughly fifty kinds of entities (six leptons, six quarks, the twelve field bosons we have discussed, the graviton, the Higgs boson, and an array of anti-particles). Theories that go beyond the standard model add additional kinds of entities. Most of these – as well as some of the key properties by which they are characterized and differentiated – were not conceived of in, say, 1900. In some contexts the question whether two items are to be treated as of the same kind, *simpliciter*, is downright misleading since what counts as the *same kind* varies with specific research contexts. Protons and neutrons are the same with respect to the strong interaction, but not with respect to the electromagnetic interaction. Consider again: Are isotopes of an element items of the same kind? What about isomers of a compound, and ionization states of an atom?

We must recognize that there are multiple forms of scientific research, and of theory change, with different features providing continuity through different changes. TC provides an account of the dimensions on which conceptual change can take place, and a highly flexible account of ways that features from different dimensions can be mixed and matched to provide continuity in particular cases. TC also provides an account of the role of referents in ongoing research through the role of instantiation conditions in determining conceptual content along with the account of a theory as the hypothesis that a particular conceptual system provides a basis for understand a specific domain.

We are now ready to return to the question of whether we have any good reasons for believing that we are making progress towards the correct description of items in the world as they are apart from any of our theorizing. Given the scope of the conceptual changes that occur as science develops, an argument on behalf of such progress must allow for a highly non-linear approach in which we may be on the wrong track for substantial periods of time – perhaps for most of the history of a subject. Sometimes a new theoretical development actually moves us further away. An account of progress that is compatible with this kind of development can be built on the above remarks about the development of instrumentation yielding a wider variety of means of interaction with nature, and results of much higher precision than in the past. All of these interactions provide constraints on our theorizing – constraints that come from nature. A theory that meets contemporary constraints has passed tougher tests than were available in the past, while the range and precision of such tests continues to grow. As a result, we have reasons for believing that contemporary theories provide a better account of nature than their predecessors, even though we cannot measure how close we are. Moreover, the process of theory testing never ends, so that the constraints on successful theories continue to grow.[12] Note

especially that the theory-dependence of observation – in the sense that observational results must be interpreted in terms of the concepts of the theory being evaluated – supports this project. It is the pursuit of such interpretation that allows us to recognize cases in which empirical results are incompatible with a particular theory, and to consider other theories with which they are compatible. Incommensurability – understood as the inability to translate newly introduced concepts into a previously available framework – does not undermine this project. This is fortunate because once we recognize that humanity did not begin its intellectual journey already possessing the ability to formulate all concepts that would ever be required, incommensurability becomes a requirement for progress.

A key question that now emerges is whether these constraints ever become sufficiently powerful to require acceptance of a single theory. There is no simple answer to this question. The answer may be different in different domains – such as the cause of polio and the fundamental constituents of the material world. Moreover, we must not forget that the elimination of specific theories, or classes of theories, from serious consideration, is an important form of progress in our knowledge of the world.

Note how the issue of translating concepts from one framework into another framework has dropped out of this discussion. Empirical evaluation of a theory can take place within the framework of that theory. Failures of the theory can be recognized, and attempts to construct or learn an alternative can begin. As Kuhn has emphasized, this is a different process than translation, but it is a process well within human capability. We can now also get beyond two further points that have generated some confusion in the literature. First, given that we need a theoretical framework to carry out coherent research, Kuhn, Lakatos (1970), and others have maintained that theory evaluation is not just between nature and a theory, but always involves two competing theories. However, this thesis runs together two quite different points. We can agree that scientists do not reject an established theory – leaving themselves with no basis for organized research – unless they have an alternative to adopt. In this sense, theory evaluation is comparative. But scientists do not need an alternative theory in order to *recognize* that the prevailing theory is empirically or conceptually defective, and thus seek an alternative. The empirical failings and internal inconsistency of Bohr's theory of the atom were well-known, but it took some time to find a successor.

Second, the kind of incommensurability that remains at this point in our discussion does not involve even a hint of relativism. It does involve a large dose of fallibilism: recognition that science proceeds by means of theories that are subject to reconsideration and replacement by radically different theories. But, we have seen, the replacement process is based on specific comparisons between theories – including their ability to handle results of our probes of nature. This does not mean that evaluations will be simple, straightforward, or algorithmic – only that there will be sufficient grounds

for coherent debate, which may include specification of further tests that could lead to a decision. Most importantly, as long as we are doing science, we accommodate theories to the results of empirical probes and it is not the case that any theory can be defended come what may.

There is one more form of incommensurability that remains to be considered: the psychological problem that arises for many people in adapting to new concepts. Three points are worth making in this regard. First, human cognitive history shows that – as a species – we are capable of carrying out this task. To be sure, some people are better at making such adjustments than others, and the number of people who introduce new concepts is considerably smaller than the number of those who can learn them. No doubt some are left behind in the process; such is life. Second, the gaps that must be crossed to introduce and learn new frameworks are considerably smaller than have sometimes been supposed. Even the transition to a strikingly new framework can result from relatively small, systematic changes in an available framework. As a result, there are conceptual bridges that can take us to the new framework, and TC gives an account of where to seek those bridges. *Third*, it is worth repeating that innovators and early adopters of a new framework are often masters of the previous view and thus able to find means of generating interest in the members of an existing community.

Beyond these three observations, there remain such problems as a detailed understanding of how people adapt to new concepts, and why some adapt more easily than others. More generally, there is a problem that Kuhn maintains was always his central concern: "What was and is at issue is not significant comparability but rather the shaping of cognition by language, a point by no means epistemologically innocuous" (1983: 55). These, however, are empirical questions to be pursued by the appropriate sciences.

# Notes

## 1 Studying Concepts

1 Sellars suggests that the nature of language learning changes once we have acquired a first language – much as learning games changes once we understand the concept of a game (SRLG 348).

2 In Illinois an elected official called "The Secretary of State" is in charge of licensing automobiles and drivers.

3 Throughout this book I use "item" as a neutral term that involves no commitment as to whether I am discussing objects, processes, or whatever. Sellars often adopts a similar practice, e.g., SRTT 97; SM 40ff; SK 298, n.1.

4 For the original psychological work see, for example, Rosch 1973a, 1973b, 1978; Rosch and Mervis 1975. For reviews of the literature and its impact, see Lakoff 1987 Part I, Smith and Medin 1981; Weitz 1988. See Bishop 1992 and Ramsey 1998 on the significance of this work for conceptual analysis.

5 Some will wonder why I did not quote a philosopher here. In response I note that although Sen is an economist, he is influenced by, and has influenced, philosophers. These passages provide an especially clear statement of the view I am describing and also illustrate one place where this common philosophical view has – for good or for ill – influenced thinkers outside of philosophy departments.

6 Cf. "The concept of *logical consequence* is one of those whose introduction into the field of strict formal investigation was not a matter of arbitrary decision on the part of this or that investigator; in defining this concept, efforts were made to adhere to the common usage of the language of everyday life. But these efforts have been confronted with the difficulties that usually present themselves in such cases. With respect to the clarity of its content the common concept of consequence is in no way superior to other concepts of everyday language. Its extension is not sharply bounded and its usage fluctuates. Any attempt to bring into harmony all possible vague, sometimes contradictory, tendencies which are connected with the use of this concept, is certainly doomed to failure. We must reconcile ourselves from the start to the fact that every precise definition of this concept will show arbitrary features to a greater or lesser degree" (Tarski 1983: 409).

7 See Sankey (1994, Ch. 4) for discussion of some of the problems that arise when philosophers ignore the distinction between the language of a specific theory and the total language in which that theory is expressed. Thagard (1992: 113–17) provides an illuminating comparison of learning a second language and learning a new scientific conceptual system.

8 Representations of objects in the world exemplify one important type of concept; we will see in Chs 4 and 5 that this is not the only type.

9 Putnam also argues that meaning involves the social environment. I discuss this view in Sec. 6.2.
10 My distinction between a physical and an abstract description partially parallels Hooker's distinction between a causal and a functional description (1995 Sec. 2.1.1). Hooker's causal descriptions can be identified with my descriptions from a physical perspective, but my notion of an abstract perspective allows for a wider range of cases than we get on Hooker's account of a functional description. Hooker identifies a functional description with an input/output map; a flow chart gives a great deal more information than this about a program.
11 Biological and psychological studies of neural processes may also make use of different concepts. Many discussions of concepts in the psychological literature are carried out from an abstract perspective. For example, psychologists offer abstract descriptions when they draw diagrams showing part-whole relations or genus-species relations among concepts. I will discuss these in Sec. 6.6.
12 I am using Peacocke's language in talking about *psychological* and *philosophical* studies of concepts; it is clearly part of Peacocke's view that philosophical studies of concepts proceed from an abstract perspective.
13 Hooker treats empiricism as a liberalized version of positivism, where the positivist account limits logic to finite truth functions (1987: 66).
14 It is often argued that naturalistic epistemology is circular. I will not pursue this issue here. For replies see Brown 1994a and Shogenji 2000.
15 The thesis that observation is theory-laden had already entered the literature in Hanson 1958.
16 This formulation allows for two possibilities: that one theory can be translated into the language of the other, or that both are translatable into some third language.

**2 Conceptual Journeys**

1 In this section I will use the following abbreviations for frequently cited works:
B: Brock 1993
P: Pais 1986
Ra: Romer 1964
Rb: Romer 1970
T: Trenn 1977
Ra and Rb contain several key papers from this period. Rb includes a valuable "Historical Essay" by Romer (3–60) covering the period I am discussing.
2 Ancient Chinese doctrine also admitted five elements – air, water, earth, metal, and wood – but did not distinguish celestial from terrestrial elements (B 6, Leicester 1971: 53–55).
3 Its current use depends on additional concepts that would have made no sense to the ancients. For discussion of some contemporary versions see Kostro 2000; Wilczek, 1999.
4 I do not consider the phlogiston theory in any detail since it is widely discussed in the literature. Briefly, phlogiston was viewed as a substance that is emitted in combustion, respiration, and a number of chemical transformation such as the rusting of iron; phlogiston was believed to be absorbed in other chemical reactions.
5 The "Voltaic pile" was invented in 1800 and the dissociation of water into hydrogen and oxygen soon followed.
6 In particular, I will say nothing about the development of organic chemistry, although some of the developments I will mention took place in that context.
7 See P 41–42 for discussion of some early attempts to decide the nature of these new rays.

8 Four of Becquerel's papers on this research are translated in Ra.
9 Fluorescence occurs when a material radiates with a characteristic color after being struck by some other radiation.
10 Also in 1900, Villard discovered another type of penetrating rays that were not affected by a magnetic field. Rutherford studied these new rays in 1901 and 1902, and labeled them *gamma rays* (P 9, Rb 23–24). These rays do not play a role in the particular set of developments I want to recount.
11 See also Romer's note (Ra 116–17) and Becquerel's paper that follows.
12 Rutherford and Soddy suggested two possible hypotheses to account for the residual activity, but concluded that the one discussed in the text is more probable.
13 By 1903 Thomson (and others) had concluded that "The atom of hydrogen contains about a thousand electrons" (quoted in P 179).
14 I noted above that Becquerel believed alpha rays to be a kind of secondary X-ray caused by the beta rays, but by 1903 Rutherford had strong evidence that alphas are independent particles with high mass and positive charge (P 61; Ra 151). Still, the situation was sufficiently unsettled so that in a paper of 1905 Rutherford described alphas as "groups of electrons . . . in rapid motion, and held in equilibrium by their mutual forces" (Ra 218–19).
15 Pais (1986: 118) notes that we need this concept to understand the phenomena once addressed by the notion of a metabolon.
16 The discussion in the remainder of this paragraph relies heavily on Rb 52–58.
17 As Soddy noted, the same proposal was made slightly earlier by van den Broek (1913), although his concerns were different: van den Broek was attempting to bring the periodic table into accord with the thesis that all elements are built up out of halves of alpha particles.
18 In this section I will use the following abbreviations for frequently cited works:
BA: Baron 1969
BO: Boyer 1959
BM: Boyer 1991
D: Descartes 2001
Ka: Kline 1972
Kb: Kline 1980
M: Maor 1995
N: Nahin 1998
19 See Kitcher 1983, Chs. 7–9 for an important general approach to mathematical change. Kitcher considers several kinds of mathematical change and argues that they are rational, but there is no particular emphasis on conceptual change – although the issue does arise in some of his discussions. As far as I can see, there is nothing in Kitcher's account that is incompatible with my approach to conceptual change.
20 Many contemporary writers avoid the term "imaginary," talking instead only of "complex numbers." It will promote clarity to use "imaginary" at the present stage of our discussion and reserve "complex" for numbers that involve both a real and an imaginary part. This usage will sometimes be overridden by the usage of authors I discuss.
21 Note how this equation was written in a form that avoids use of a minus sign; this was standard practice during the period in question. For the history of this formula and discussion of Cardan's method see BM 282–86; Ka 253, 263–65; N 8–17.
22 See N 20–22 for a general discussion of the roots of the equation in question, although using techniques not available in the sixteenth century. Cardan worked with specific values of $p$ and $q$ and knew that 4 was a root of the resulting equation (B 286). More generally, it is common to find problems that are posed in

terms of real numbers and have real solutions, but that make use of imaginary numbers in arriving at the solution; see N Chs 4–6 for examples. The mathematician Hadamard (1865–1963) is supposed to have observed that, "The shortest path between two truths in the real domain passes through the complex domain" (quoted in N 70).

23 An isomorphism is a one–one correspondence that preserves structure under a particular operation. For example, consider two equi-numerous sets $\{A_1, A_2, \ldots\}$ and $\{B_1, B_2, \ldots\}$, an operation $a$ defined on the first set, and an operation $b$ defined on the second set, where $A_i\ a\ A_j = A_k$, and $B_i\ b\ B_j = B_k$. We have an isomorphism under these operation if and only if whenever $A_i$ corresponds to $B_i$, and $A_j$ corresponds to $B_j$, then $A_k$ corresponds to $B_k$.

24 Although I will not pursue the matter here, mathematicians would add that such generalizations must have some independent mathematical interest; any *ad hoc* construct that meets the conditions stated in the text will not do.

25 This construction is do to Hamilton in 1837 (Ka 775–76).

26 Integers, rationals, and reals can all be represented geometrically on a line. A geometric representation of complex numbers requires that we move to the plane. We can use an analog of Cartesian coordinates with one axis representing the real part of the number and the other representing the imaginary part. Given this representation, complex numbers are analogous to 2D vectors, and vector addition is an appropriate analog for addition of complex numbers.

27 If we think of a complex number as a vector, then the positive square root of this special square gives the length of the vector. Complex roots of algebraic equations always occur in conjugate pairs.

28 Properly speaking, "exponent" refers to the notation used to indicate powers to which we raise some expression. However, exponents are so familiar in contemporary mathematics that I will treat "power" and "exponent" as synonyms.

29 See Sec. 2.2.4. Newton discovered the theorem c. 1665; it was published by Wallis, with due credit to Newton, in 1685 (BM 393).

30 Other bases can be used. When logarithms are introduced into calculus a different base becomes convenient. This base, signified by $e$, is the limit of $(1 + 1/n)^n$, as $n$ grows without limit; logarithms to this base are called "natural logarithms." Introduction of $e$ by this route requires the concept of the limit of an infinite series, but students of calculus learn this concept before they encounter the reasons for taking $e$ as the base for logarithms.

31 While this yields a formal account of the introduction of irrational exponents, in actual calculations we use approximations for the irrational numbers. A precise definition requires that we introduce limits.

32 If we rewrite the equation in the form $e^{i\pi} + 1 = 0$ we have a simple relation between five numbers of fundamental importance; the definitions of these numbers would not lead us to expect any such relation. Some mathematicians consider this result utterly amazing, even mystical.

33 See N 67; M 175–77 or an appropriate textbook for discussion and proofs.

34 See Ka 407–11 for discussion of an early seventeenth century debate over the existence of logarithms of negative and complex numbers.

35 The gamma function also applies to complex numbers with the restriction that the real part of the complex number be greater than one. See Hassani 1999: 309–10.

36 For discussion and examples see BA Chs 4–6, BO Ch. 4, Ka Ch. 17.

37 "Infinite series were in the eighteenth century and are still today considered an essential part of the calculus" (Ka 436). Many functions could be handled only by expanding in a series and integrating or differentiating term by term. "Moreover, it seemed clear, as Euler and Lagrange believed, that every function could be expressed as a series" (Ka 436).

38 See M 158–59 for some examples of Euler's practice in handling infinite series.
39 This was published in 1638 although Fermat claimed to have worked it out "some eight or ten years" earlier (BO 155).
40 Newton introduced the term "fluxions" for what we call "derivatives" in 1671.
41 The three dates cited are the accepted dates at which Newton wrote these works. He withheld publication for many years and the order in which items were published is not the same as the order of composition. See BO 190–202; Ka 359.
42 Hamilton still considered pure imaginary numbers absurd. He emphasized that treating complex numbers as ordered pairs eliminates any need to make sense out of $\sqrt{-1}$ (Kb 177–78).
43 For a formal development see Lightstone 1978, Ch. 13. Accessible introductions will be found in Davis and Hersh 1972 and Stewart 1992: 107–10. Non-standard analysis also leads to the introduction of another class of numbers – hyperreals – that have the familiar real numbers as a subset.
44 In this section I will use the following abbreviations for frequently cited books:
FA: Farley 1982
FK: Fenwick 1998
G: Gasking 1967
RI: Ritvo 1997
RD: Rowland 1992
S: Singer, *et al*. 1993
45 At this stage of research the subject was described as the study of "generation" and included both the development of new individuals and the regeneration of tails, claws, and such in some animals. It took time and research to recognize that these are distinct phenomena. (Farley 1981a: 163; G 8, 86).
46 Newton believed that there are multiple kinds of attraction in nature in addition to gravitation, electricity, and magnetism; see Sec. 9.4.
47 There were many related issues that also entered into these nineteenth century debates. Churchill (1979) provides a brief summary.
48 The use of "germ" is worth further comment. It was "in biology a label for the material of heredity, popular in the 18th and 19th centuries and sufficiently ambiguous to cover a wide range of ideas. For example, the egg and sperm were sometimes called germs, as were the contents of the cell nucleus. Frequently, it was an abstract unit supposed to be passed on to offspring regulating their development. There is no modern equivalent, although the concept of germ plasm led to the theory of the gene" (Maienschein 1981d).
49 Philosophers will want to speak a bit more precisely and distinguish between the concept of telegony and the doctrine that this concept is instantiated. I will not press this point in the many cases where looser language does not cause confusion. Concepts that are found not to be instantiated often disappear from our active repertoire. Telegony stands alongside phlogiston and pangenesis in this respect.
50 Kearney (1998) provides a detailed discussion of alternative procedures and risks.
51 As is the case throughout this chapter, I am taking only a selection of possible examples, with the focus on Western societies. Different issues and further distinctions might well arise if we looked at other societies.
52 Historical and anthropological study are liable to yield conceptual diversity in surprising places. Lon Becker informs me that hypothetical cases in which a fetus is moved from one woman to another are discussed in the *Talmud*. The motivation for the discussion is the religious status of the child, since being Jewish passes from mother to child. So some people have thought about some aspects of the distinctions we are considering in the past – albeit in scientific, social, and legal frameworks quite different from those of the present discussions.

53 Some object to the term "surrogate" on the grounds that the woman who gives birth to a baby is not a surrogate from the baby's point of view (e.g., Nelson 1992: 297, cf. Purdy 1992: 305–6). One alternative term that has been proposed is "contract mother," but this is also subject to objection since cases of surrogacy do not always involve an explicit contract. The unsettled state of the terminology is an indicator of the unsettled conceptual situation. I will use "surrogate" because it seems to be the term that is most commonly used.
54 IVF must not be confused with cloning which raises further conceptual issues since in that case a single parent contributes a child's entire genetic endowment.
55 Theological considerations led to one variation on IVF that was developed to meet Catholic objections to fertilization in a dish as "not natural." In the procedure known as "gamete intrafallopian transfer" the egg and sperm are brought together in the laboratory, but do not merge. Instead they are separated by an air bubble and transferred to the woman's fallopian tube where fertilization takes place. Clearly, this is not feasible in some of the cases we have noted.
56 There are also techniques in which a woman's natural cycle is tracked and a single egg removed when it develops normally (Kearney 1998: 96–100). This requires that the woman's egg production be normal. It reduces the chances of pregnancy, but also reduces (sometimes eliminates) the need for drugs, the chances of multiple births, and the cost.
57 Unfrozen embryos are viable for up to 14 days; "Estimates of viability for cryopreserved embryos range from two to ten years" (FK 191).
58 Compare cases in which various countries refuse to extradite accused murders to the US because they will face the death penalty.
59 Note the term "preembryo." Introduction of this term is part of the process of figuring out how to classify these items; see Shevory 1992: 235–38 for discussion.
60 S 4–6 provides a brief summary of the recognized stages.
61 See Levi 1949 on creative adaptation in the development of US product-liability law during the nineteenth century.
62 Multiple distortions were introduced by early telescopes and Galileo's ability to separate these from the genuine phenomena is impressive. Every one of the phenomena he reported has stood the test of time. The value of the new instrument was rapidly recognized by astronomers – including those associated with the Catholic church. Christopher Clavius, head of the mathematics department (which included astronomy) of the Jesuit Collegio Romano, was writing to Galileo and sharing his own telescopic observations less than a year after Galileo's first publication of his telescopic results. Clavius and the other three mathematicians of the Collegio Romano provided a favorable report on the telescope to the Roman Inquisition just thirteen months after Galileo announced his first results.
63 Ritter's motivations are of some interest. He was a *Naturphilosoph* who believed that the spectrum must form a symmetric structure with red and violet as poles (analogous to the north and south poles of a magnet) and green as a neutral axis. The discovery of radiation beyond the red required the existence of a parallel radiation beyond the violet to maintain this symmetry.
64 For detailed examples see Brown 1987; Franklin 1986, 2001; Galison 1987, 1997; Shapere 1982.
65 For detailed discussions see Brown 1987, 1995; Kosso 1989; Shapere 1982.
66 For discussion of this side of Descartes' work see Buchdahl 1969; Garber 1992a, b, and Gaukroger 1995. I discuss Descartes' physics in Sec. 9.3.
67 Presumably this second kind of change does not occur in physical science where it is assumed that – at some level – the physical world remains constant, and scientists seek concepts that will describe this level. It is likely that the second kind of change does occur in the domains that social sciences study.

68 Thagard (1992, Ch. 3) provides a list of kinds of conceptual change that includes many of the varieties I discuss below, and some that I do not discuss. Thagard's discussion is set in the context of a specific view of conceptual content that I will discuss in Sec. 6.4.
69 On a falsificationist account being false does not remove a theory's scientific status; only the refusal to reject a theory in the face of *appropriate* falsifying instances removes scientific status.
70 Kitcher (1983: 207–12) provides a parallel discussion using different examples.
71 I have ignored some subtleties concerning commutivity.
72 Some write G5 as $d^2 = -x^2\,y^2 - z^2 + t^2$. The two versions are equivalent since it is important only that the space terms have the same sign and the time term have the opposite sign.
73 The quote at the head of Sec. 2.2 refers to a case in which the concept of the dimension of a space was generalized to allow for non-integral values.
74 For example, polynomials of the form $a_n x^n + a_{n-1}x^{n-1} + \ldots + a_0$ can be viewed as vectors in an $n + 1$ dimensional space

**3 Some Theories of Concepts**

1 All Locke references are to Locke 1984. References will be given in the sequence book, chapter, section, indicated respectively by uppercase roman, lowercase roman, and Arabic numerals – e.g., II.viii.1. Page numbers will be given for quotations.
2 When I am actually imagining it is not altogether clear, in Locke, whether I am introspectively aware of the ongoing act, or aware of an impression that is distinct from the act in the same way that the impression of solidity is distinct from the actual solidity in the physical world. On my reading, none of the classical empiricists ever became clear on how to assimilate awareness of the activities of our own minds to the doctrine of ideas.
3 Locke's main discussion of abstract ideas occurs in his account of general language (III.iii).
4 Some commentators hold that Locke has a *selective-attention* account of abstraction of the sort we will encounter in Berkeley. I do not find this in Locke, and the interpretative debate is not important for present purposes. Those who disagree with my reading of Locke can take the present discussion as an exploration of one option in the theory of concepts that has been discussed by some philosophers.
5 Weitz (1988: 113–14) makes this point in regard to the idea of an idea.
6 See Berkeley 1948 for all references. I use the following abbreviations for specific texts. In vol. I, EVI*n*: *Essay Towards a New Theory of Vision*. In vol. II, PHK*n*: main text of *Principles of Human Knowledge*; PHKI*n*: *Introduction to PHK*; TD*n*: *Three Dialogues Between Hylas and Philonous*. In vol. III, ALC*n*: *Alciphron*. In EVI*n*, PHK*n*, and PHKI*n*, *n* is the section number; in TD*n* and ALC*n*, *n* is the dialogue number. Where Berkeley's texts are divided into brief sections page numbers are omitted.
7 Berkeley emphasizes that we need not always call up the associated idea when we understand a word (PHKI*19*) and that language has other functions besides that of indicating ideas (PHKI*20*). Still, for epistemological purposes the case in which a word is associated with some idea is basic and Berkeley agrees that as long as we are dealing with a meaningful word, rather than an ink blob, there must be an idea that we could call up.
8 It seems to follow that notions would also exist unperceived, but Berkeley does not discuss this point. This could be taken as an argument against the interpretation I am proposing, but it could also be taken as one of many indications that

Berkeley never fully worked out his doctrine of notions in print. I think that the following account provides enough insight into Berkeley's thinking to justify the later option. I will omit the property of existing unperceived in my discussion of notions.

9 Thus if there were material objects with causal powers, they would be active.

10 Unless otherwise stated, references are to Book I of Hume's *Treatise* (2001). References will be given as part.section, indicated respectively by Roman and Arabic numerals – e.g., II.1. Page numbers will be provided for quotations.

11 Weight is also problematic. Arguably, in everyday experience weight appears to be a simple quality although we will see in Sec. 9.4 that after Newton weight is understood as a relation.

12 Hume also follows Berkeley in rejecting simple ideas that have more than one source.

13 Hausman (1988) arrives at a similar interpretation as a result of comparing Hume's initial account of simple ideas with his discussion of the "formal distinction," which I consider next. In Hausman's terminology, color-cum-shape is a psychological simple, but color and shape alone are logical simples. I avoid "logical" here because it is not clear what the term means in this context. The important point is that psychologically simple ideas do not provide the basis on which Hume builds his epistemology.

14 Plato's view of the ultimate source of these innate concepts indicates that they are indeed copied from experience, although not sensory experience.

15 This is Kant's strategy. He argues that we have a small number of central concepts that are not derived from experience – and attempts to justify their application to experience.

16 Locke does not share this view since he holds that ideas of secondary qualities represent properties of material objects without resembling those properties.

17 Some empiricists, including Price (7, n. 1) and Russell (1959, Ch. V) express doubts about whether sensation and introspection are the *only* forms of acquaintance available to us. I will not pursue this issue here since I am concerned only with the thesis that all meaningful language is built on a primary vocabulary that must be introduced by ostension; this is independent of questions about the domains in which ostension can occur. I will continue to focus on sensory acquaintance, but it is worth noting that the variety of forms of acquaintance available to us seems to be an empirical question about human beings.

18 The same requirement extends to theories of concepts that are distinguished from theories of linguistic meaning. In the case of my own theory of concepts, I will address this question in Sec. 5.10.

19 Even those who held that sense data are parts of physical objects still considered the items of acquaintance to be qualities, and required the logical construction of material-object language out of quality language.

20 There were attempts to eliminate the need for theoretical terms by means of specific logical maneuvers via Craig's theorem and Ramsey sentences. These are further attempts to tame theoretical terms, not criticisms of physics for introducing them. These attempts are not germane to the present topic. For discussion see Brown 1979: 44–45; Hempel 1965: 210–17; Nagel 1961: 134–37, 141–42; Scheffler 1963: 193–216.

21 No time is associated with "soluble" because the disposition is taken as a permanent feature of an item.

22 Twenty years later Carnap acknowledged the possibility of dropping the material conditional, but still opposed this alternative. After noting that we might be able to establish explicit definitions using non-truth-functional logical or causal connectives he writes, "the exact form of definitions of this kind is not yet sufficiently clarified, but still under discussion" (1956b: 64).

23 For discussions of some of the historical background see Feigl 1956, 1970; Hempel 1952: 33–39. This proposal was developed independently of the logical empiricists in Campbell 1957, originally published in 1920.
24 It has been (unreliably) reported that Lewis described himself as a Kantian "who disagrees with every sentence of the *Critique of Pure Reason*" (Beck 1968: 273).
25 Strictly speaking, implications relate propositions not concepts. When I write here – and in later chapters – of concepts implying concepts, this should be understood as an abbreviation for implications between propositions of the form "*x* is a *C*."
26 Lewis notes that there is empirical evidence supporting the claim that sensations vary among individuals (MWO 111).
27 Here are two examples. The seventeenth century colonists at Plymouth (as well as the natives) hunted small whales that came close to shore. The colonists called them "blackfish," and in a description of the practice written in 1793 they are described as "fish of the whale kind" (Deetz and Deetz 2000: 248). Meanwhile, on the other side of the Atlantic, in 1774 a critic of Linnaeus' new system of classification, wrote: "What will the plain man think of a manner of classing, that denies a whale to be a fish?" (Home 1996: 17).
28 Lewis distinguishes four modes in which terms and propositions have meaning: denotation, comprehension, signification, and intension. The distinction I am discussing concerns two aspects of meaning as intension.
29 Those who do not include formal logic and pure mathematics in the analytic domain still hold that any philosophical knowledge is analytic.
30 For Carnap see especially 1956a, but the point is clear in the measured retreat we examined above. Carnap was prepared to give up several aspects of the original logical empiricist theory of meaning, but not the thesis that meanings are expressed in analytic propositions.
31 Antony (1993) points out that this thesis is implicit in Quine's recognition that we can protect selected propositions from refutation.
32 This term was introduced in Laudan, *et al*. (1986) as a neutral term for protected propositions that play a key role in the work of several contemporary philosophers of science.
33 Just what those circumstances are is a major research subject in contemporary philosophy of science.
34 For detailed discussions of Kantian elements in Kuhn see Brown 1975, 1979 Ch. 7; Hoyningen-Huene 1993; Kuhn 1983.

**4 Sellars: Exposition, Interpretation, and Critique**

1 For discussions of Sellars' philosophy that include material on his theory of concepts see Bernstein (1966–67); Brandom (1994); Brown (1986, 1991); Burian (1979); Delaney *et al.* (1977); Pitt (1981).
2 Van Fraassen takes Sellars to be the paradigmatic realist; van Fraassen's title *The Scientific Image* (1980) is an explicit allusion to PSM.
3 "Alas" lacks conceptual status because it lacks such implicational relations; including "alas" in a sentence does not generate any implications not already supplied by the original sentence. "Alas" underwrites inferences about the speaker's state of mind, but that is a different issue.
4 LABEL is itself a concept and labeling is a sophisticated linguistic activity (cf. NO 120). The present discussion assumes that the reader already has the concept of a label.
5 This discussion raises the crucial question of how wide a range of our beliefs about a type of item are included in its concept – that is, where we draw the line between statements that express conceptual content and those that use a concept to make additional claims about items that the concept describes. I will develop

Sellars' view in the next section. It also raises the question of how we individuate distinct conceptual systems; I discuss this issue in Secs. 5.8 and 6.1. Bonevac (2002: 12–14) attributes total holism to Sellars on the basis of his discussion of color concepts that leads to the remark, "there is an important sense in which one has *no* concept pertaining to observable properties of physical objects in Space and Time unless one has them all – and a great deal more besides" (EPM 148). Bonevac criticizes both Sellars' argument for this view and the view itself, and notes that "conceptual localism" provides an intermediate position between total holism and the kind of atomism that Sellars is attacking. Rather than debating the import of this passage, I note the passages cited in the main text suggesting that Sellars already adopts local holism. Moreover, I will defend local holism in Ch. 5 independently of whether this view is properly attributed to Sellars.

6 Occasionally Sellars mentions other types of concepts, e.g., EAE 460 where he describes "semantical terms" as a distinct class (cf. SAP 315, n. 1). But these other classes are not analyzed in any detail, and I will limit discussion here to the three types that I am about to introduce.

7 CONCEPT is an important exception. I will discuss self-reference in Sec. 5.8 and CONCEPT in 5.9. Still, the vast majority of our descriptive concepts are not self-referential; for now I will focus on these.

8 Recall that my use of the phrase "change the concept" is neutral between the notion that we alter a concept, and the notion that we replace it with a more or less similar concept; I will discuss Sellars' account of similarity in Sec. 4.5.

9 Material rules are central to Sellars' thought beginning with his earliest papers, although they appear under different rubrics. In ENNW, PPE, and LRB they are referred to as "conformation rules" which stand alongside formation and transformation rules in determining the content of non-logical axiom systems. In CIL they are "material invariances."

10 On the basis of a passage in SRLG Pitt (1981: 25–26) concludes that, for Sellars, laws imply material rules but the converse does not hold. After noting that accepting a law amounts to accepting a material rule, Sellars adds:

> an important qualification. Obviously, if I learn that in a certain language I may make a material move from 'x is C' to 'x is D', I do not properly conclude that all C is D. Clearly, the language in question must be the language I myself use, in order for me to assert 'All C is D'. But with this qualification we may say that it is by virtue of its *material* moves . . . that a language embodies a consciousness of lawfulness of things.
>
> (SRLG 331)

The passage is not crystal clear, but I take it that Sellars' qualification is that I may infer a universal generalization from a material rule only in my own language.

11 SAP provides Sellars' most detailed and explicit account. Sellars also describes these as necessary truths that are dependent on a particular subject matter, e.g., SM 68; SRTT 104.

12 Sellars attempts to extract a great deal of philosophical juice from this doctrine. He attempts to move from the need for GAs in any descriptive language to an account of causal necessity and an approach to the problem of induction. I will not spell out the details here since they are complex, controversial, and not germane to the project of this book. However, in Ch. 7 I will apply the theory of concepts I arrive at to the concept of a causal relation, something that Sellars never attempts – perhaps because he considers causation to be a special concept that plays a unique role in our thinking. See especially CDCM and IV.

13 See Feigl (1956: 17–19) for some historical background on the notion of implicit definition, including Sellars' role in its development. Sellars' most detailed discussion of this theme is SAP.
14 See, for example, LRB 310 and SRII 179 for the claim that, at the most fundamental level, a language (i.e., a conceptual system for a specific subject) should be viewed as an axiom system.
15 Cf. EPM 148, n. 1 where Sellars distinguishes between a rudimentary concept of green and a richer concept.
16 Sellars' exact account of what he is seeking to show varies. Often he seems to claim that his account gives the correct analysis of "means," although sometimes he tempers this view. For example, he notes that "means" has various everyday senses and that his objections are aimed only at "those elements of everyday usage which are reconstructed by the semantics of Carnap and Tarski" (EAE 460, n. 38, cf. SM 77, 82). In SM (114–15) Sellars describes ETs as a species of semantical rule. This is a striking shift in terminology, but we must keep in mind that for Sellars finding the correct concepts for describing a subject matter is part of the process of learning more about that subject. As a result, changes in terminology and modifications of earlier views are closely related.
17 Logical concepts play a central role in Sellars' account of "means." With regard to another logical concept Sellars writes: "'Not' stands for an *operator*, rather than an *attribute*, but understanding its status will turn out to be the key to understanding the status of all other senses and intensions . . . " (SM 70–71).
18 However, Sellars is not committed to the view that identity of meaning occurs across languages in the way that identity of knights occurs across chess sets. Rather, he is laying the groundwork for an account of how terms in different languages may be more or less similar. "One can also make sense of the idea that bishops are more like castles than they are like knights. Indeed, we are all accustomed to making judgments of this kind. 'The bowler in cricket is like the pitcher in baseball'. We decide similarity of 'pieces' with reference to the roles they are given by the rules" (MFC 434).
19 This applies equally to children learning a first language and adults learning an additional language.
20 Sellars does not actually argue for this claim, although he quips that the better is the enemy of the best (P 97). In an extended discussion (SRII 187–93) Sellars strongly supports Feyerabend's claim that the framework of commonsense has no ontological priority and can be replaced by a scientific framework, while adding assurances that he does not hold that we should make the replacement now. Churchland (1979) proposes that such replacement should occur as science proceeds, and that at each stage we should respond to experience using the concepts of our best available theories. He also provides some detailed examples of what such a replacement would look like.
21 Sellars' account of observables is essentially the same van Fraassen's. This agreement on what counts as observables sharpens their disagreements on the role of non-observables in science.
22 See Staley (2004) for a detailed discussion. Throughout this discussion I use "data-pattern" to describe a type, not a token.
23 Detection of weak neutral currents is another example. See Galison 1987 for discussion.
24 Brandom (1994: 223–25) describes how the Sellarsian account can be extended to particle detection using modern instrumentation; I once proposed a similar account (Brown 1986). This extension works for some cases, but not for the case discussed in the text. I return to this topic in Sec. 5.3.
25 In his theory of practical reasoning Sellars argues that "shall" – which he uses to express an intention – is basic, and that "intentions imply intentions just as

beliefs imply beliefs"(SM 182). Moreover, "in their primary use, 'ought' and 'good' are special cases of 'shall'" (TA 106). However, Sellars does not consider "shall" a normative term.

26 Sellars turns to teleological considerations when he considers the ultimate justification of moral claims; cf. IILO 206–12 and Solomon 1977: 150, 180–86. I will not discuss Sellars' account of this justification, but ends will enter into our discussion of prescriptions in Sec. 5.5.

27 In IILO (162) Sellars says that the connection between thinking and doing must be "*analytic*, a matter of strict logic." Elsewhere Sellars tells us that "emotivism was on to *something*" when it held that the "connection between *believing something to be good* and *being positively concerned about x*" is analytic (SE 406). But at this point in our discussion we need not take Sellars' use of "analytic" literally. The important point is that there is a necessary connection between thinking and doing in the sense that an *actual motivation* for doing what we believe to be prescribed is part of the content of prescriptive concepts.

28 If valid inferences were required we would have to spend our entire lives making trivial deductions.

29 This example introduces a new, but related, theme: The use of analogy as a means of comparing conceptual systems; I will return to this topic shortly.

30 I take it that when Sellars writes of a "logistically contrived deductive system" in a discussion of descriptive concepts, he is referring to an *interpreted* formalism meeting both of the conditions just stated.

31 Elsewhere, in response to the claim that material in the "order of discovery" is not relevant to epistemic justification he writes: "But reflection on the fact that to answer a question of the form 'Is $x$ justified in $\varphi$-ing?' requires taking $x$'s historical situation into account should give one pause" (MGEC 174).

32 Commenting on the outcome of his discussion Sellars writes, "I have used a myth to kill a myth" (EPM 195) – the myth of the given; see also SM 71.

33 In EPM Sellars discusses thoughts first, then sense impressions. In P (which is a later paper, see the "Acknowledgments" at SPR vii-viii), Sellars discusses impressions only, and he placed this paper before EPM in SPR. Sellars also discusses impressions first in SM. I think this order is appropriate because introduction of impressions is the less complex of the two examples. While the account applies to all of our senses, Sellars focuses on vision. The entire project is reviewed (along with other themes) in SM Ch. VI.

34 The causal hypothesis does not occur in EPM. In the case of sensations it is introduced in P; in the case of thoughts, in SM.

35 Above I raised some questions about the sense in which the model is *logically* implicated in concepts introduced by analogy. In the present case this sense is clear because of the causal relation that is postulated to hold between physical objects and impressions. But this is a peculiarity of the present example and does not transfer to the full range of cases Sellars considers. For example, while billiard balls are the model for molecules, billiard balls do not cause molecules, nor do they cause us to think of molecules.

36 In SM, where Sellars' emphasis shifts to the role of sense impressions in explaining how conceptual content enters into perception, he holds that these states are neither purely conceptual nor purely physical (16–17).

37 According to a widespread myth there are two kinds of non-Euclidean geometry: Lobachevskian and Riemannian. In fact, the phrase "Riemannian geometry" refers both to a specific geometry and to a general approach that includes an infinite class of geometries, but not all geometries. Minkowski's 4D geometry for special relativity is not in the class covered by Riemannian geometry, although it is a member of a wider class that is produced by relaxing one constraint on Riemannian geometries (that intervals be positive definite). Recall the discussion in Sec. 2.5.

**5 Reconstruction**

1 Definitions in this discussion come from *The Random House Dictionary of the English Language*, 1966. Typical dictionary definitions do not contain lists of analytic sentences. Rather, dictionary definitions provide a great deal of empirical information – which is what we should expect on a Sellarsian view of descriptive concepts. The information that "chair" is a noun provides constraints that are reflected in intra-systemic relations.

2 Bubble chambers ran (they are no longer used) under cryogenic conditions which required that they be completely sealed; they also cycled too quickly for human perception. See Galison 1997, Ch. 5 for a detailed account.

3 Sellars, I take it, would agree since this is a variation on case (b) in the passage from SRLG 357 quoted in Sec. 4.2.1.

4 Sellars adds, "and no one of these types of roles makes sense apart from the others." If "language" is equivalent to "conceptual system," Sellars is suggesting a tighter integration of roles within a system than we will find when we consider examples.

5 SYSTEMIC ROLE is a concept in the conceptual system I am proposing for describing concepts. I will return to this topic in Sec. 5.10.

6 In other words, even if JTB is necessary and sufficient for knowledge, considerations of the pursuit of knowledge brings out a respect in which they are not the same.

7 I take it for granted that explanation and prediction are also major functions of scientific theories, and that explanatory and predictive success provides the main basis for evaluating theories. An adequate discussion of these roles would require accounts of explanation and prediction, but I will not pursue these large topics here. Nor will I pursue the question of whether all scientific theories have a descriptive function.

8 Those familiar with the *semantic view of theories* will recognize a similar distinction drawn by its proponents. I consider this approach further in Sec. 5.8. For detailed discussions see Giere (1988); Lloyd (1994); Suppe (1989); van Fraassen (1980).

9 There are also important interactions between our epistemic theories and other normative theories – especially our moral theories. We rule out many ways of pursuing epistemic ends on moral grounds.

10 Galileo held that in mathematics the *quality* of our knowledge is equal to that of God's.

11 For discussion and references see Kellert 1994; Rueger and Sharp 1996.

12 Lewis, who distinguishes analytic propositions that contribute to content from synthetic propositions that do not, still arrives at a massive holism, but does not find this to be obviously problematic. Perlman (2000) provides a recent survey of the issues, as well as a radical response.

13 There is a reflection of this phenomenon in some scientific terminology that was originally used literally but is now misleading if taken in its literal sense. For example, "lepton" (which etymologically means "small") was introduced for a class of low-mass fundamental particles, but as research proceeded a heavy version of these particle was discovered. They are still considered members of a single fundamental class (because they do not respond to the strong force) but mass is no longer the basis for the classification, although the rubric "lepton" has remained. Other examples include "quantum" theory, which includes continuous changes; "chaos" theory, where scientists originally thought they were dealing with situations that are chaotic in the everyday sense, but have since learned that these cases involve complex forms of organization; and "atoms" where the term is no longer used to label items that are non-composite.

14 Jackman (1999: 363) points out that this is not the same as holding that every change in a concept changes every other concept. A letter grade in a course may depend on the average of the test scores, but every change is a test score does not yield a change in the final grade. Recall also intuitionistic logic which rejects the implication from $\sim\sim p$ to $p$, but retains the converse implication.
15 Originally the semantic view required an isomorphism between the model and the real world domain to which it is applied. But Giere (and others) argue that only similarity is required. Teller (2001) provides an account of similarity that is contextual in ways that parallel the ways in which the individuation of theories and conceptual systems is contextual.
16 There is considerable controversy over just what constitutes circularity and when it is problematic. One impediment to clarity is lack of agreement on terminology. Some use "circular" as an evaluative – indeed, a pejorative – term; others use it as a descriptive term and reserve such expressions as "viciously circular" for negative evaluations; see Brown 1994a for discussion and references. In that paper I adopted the latter of the two terminological options because one of my concerns was to discuss this literature. My concerns are narrower in the present book, and I have adopted the alternative terminology: Here I am using "circular" as a negative evaluative term. Shogenji (2000) argues that in many cases circularity in the descriptive sense is only apparent, and disappears on a Bayesian analysis.
17 Fitch proposed a means of avoiding paradoxes that is not based on eliminating self-reference.
18 In accordance with the account of theories I have given, my theory of concepts consists of the claim that this system is instantiated in human cognition.
19 While all descriptive concepts have a descriptive role, they do not all have an explanatory role.
20 PRESCRIPTIVE CONCEPT belongs in this system to the extent that we consider the recognition of obligations to be a cognitive function.
21 Glock (2000: 44) notes that attributing concepts to animals does not require that we attribute to them the same concepts that we have. Glock also surveys a variety of different views on animal concepts.

**6 Clarifications, Responses, and Refinements**

1 Aspects of content that are in the mind are called "narrow content"; external aspects are called "wide content." Externalism comes in several versions (see Sankey 1994, Ch. 2 for a critical study); I focus on versions that are widely held at the time I am writing.
2 Others have extended this argument to conceptual content – either because they think that it applies in that case as well, or because they do not distinguish between conceptual content and word meaning.
3 Crane (1991: 10) and Lau (2003) note that they could not be identical in every respect since a twin-earther's body would contain XYZ where ours contains $H_2O$; but we can let this pass.
4 There are other elements besides extension and stereotype in what Putnam describes as the "meaning vector" but these need not detain us.
5 Quine (1969) invokes this ability as the starting point for our inductive knowledge of the world, but holds that the kinds we initially pick out typically fall away as a science develops, and that in a mature science the notion of a natural kind will be superseded. Kornblith also considers our ability to select natural kinds as the ground of induction but differs significantly from Quine about the eventual fate of these kinds. Kornblith views science as pursuing deeper knowledge of natural kinds, and considers the kinds we initially pick out as rough indicators of the kinds at which science will eventually arrive (1995: 78). Science explains why

specific clusters of properties form natural kinds and thereby places our knowledge of these kinds on a firmer basis than we could arrive at through study of superficial features alone.

6 Ether provides an interesting example of one regularly cited procedure for fixing reference: one points to the sky and says, "Ether is that stuff up there." If this is a case of successful reference fixing, then we should note two points: a) "that stuff" is an extremely heterogeneous mixture; b) no member of this mixture has any of the key properties the ancients associated with ether – e.g., "that stuff" is not made of material that occurs only in the heavens, and does not have a natural circular motion. If one insists that the act of pointing is still a significant stage in the development of knowledge, then so is any act in which someone points and says "ugh."

7 The same reactions occur, but their rates are different.

8 Mellor (1977: 310–11) also challenges the distinction between essential and non-essential properties. I may have been too generous when I let pass the difference between $H_2O$ and XYZ in human and twin-earth bodies. Just how casually can we treat natural laws when postulating a world that is like ours except for one or two features? Surely we should look at actual science, not just at untutored intuition. Even a retreat to bare logical possibility will not do. Kuhn's point is that a complex consisting of physical laws (as we currently understand them) plus the claim that XYZ has the same observable properties as $H_2O$ is formally inconsistent.

9 This is a central theme of Mayr 1982. Other important critiques of KP include Churchland 1985; de Sousa 1984, and Shapere 1984.

10 Burge's arguments are not limited to concepts:

> the arguments of "Individualism and the Mental" suggest that virtually no propositional attitudes can be explicated in individualistic terms. Since the intentional notions in terms of which propositional attitudes are described are irreducibly non-mentalistic, no purely individualistic accounts of these notions can possibly be adequate.
>
> (1982: 117)

I will limit discussion here to concepts, which are at the heart of his arguments.

11 This appeal to the dictionary underlines the complexity of Burge's view of the relation between concepts and words. Elsewhere Burge lists several principles of a view of concepts deriving from the Aristotelian tradition. He notes that there are multiple readings of each principle listed, and that there is some reading under which he accepts all but one (1993: 309). The principle Burge unequivocally rejects is: "Concepts are prior to language in the sense that language is to be understood as functioning to express thought; but thought is never fundamentally individuated in terms of language" (1993: 312).

12 Burge does not attempt to elicit intuitions about CONCEPT by asking for them. This is in accord with the recognition (at least as old as Plato) that our response to cases is often more reliable than our attempts at formulating the content of a concept.

13 BRISKET does not exist in France. Burge notes that BRISKET refers to the lower part of the chest. Harrap's *New College French and English Dictionary* translates "brisket" as "*poitrine*" and "*poitrine*" as "chest." Child, Bertholle, and Beck (1983) sometimes use "brisket" as a synonym for "chest" and sometimes for "middle of the chest." More significantly, they note that "French and American methods for cutting up a beef carcass are so dissimilar that it is rarely possible to find in America the same steak cut you could find in France"; in France "there is

neither short loin nor sirloin left intact, and consequently no T-bone, porterhouse, or sirloin steak" (1983: 289–90).

14 Consider another example. Suppose, following Putnam (1975), "water" was introduced by reference to a particular body of liquid – say, the main constituent of the oceans. "Ice" may have been introduced as the solid that forms on the surface of lakes in winter, and "steam" as the vapor we dimly see above boiling water. The discovery that these are all $H_2O$ would be a unification of the kind I am considering. I will consider such cases further in Ch. 10 where I explore the highly successful unification project in the history of physics.

15 IA also allows for verbal, technological, and inferential means of access to the relevant content (76–80).

16 In earlier avatars Fodor held that concept learning is inductive, that induction always requires prior concepts, and thus that primitive concepts must be innate. Fodor now considers the possibility that primitive concepts are learned – but not inductively. Concept learning requires an innate basis, but this basis may consist of mechanisms rather than concepts (142). A full account of concept learning requires further knowledge of our kind of mind, knowledge which Fodor does not claim to have, but he does insist that no reflective mental processes are involved in acquiring any primitive concept. Fodor has vigorously opposed neural-net accounts of mind, but one clear result of research on neural nets is that they extract patterns from noisy data (Clark 1997: 59–60). Thus neural nets might provide at least part of the required mechanism.

17 Fodor's version of IA introduces a new, presumably theoretical, concept: MENTAL PARTICULAR THAT IS NOMOLOGICALLY LOCKED TO A PROPERTY IN THE EXTRAMENTAL WORLD The questions just raised about theoretical concepts apply in this case.

18 Contemporary literature in analytic philosophy indicates that this aspect of Wittgenstein's work has had little lasting effect in this field.

19 I do not find Barsalou's discussion of the distinction between structural invariants and constraints to be particularly clear or helpful. Barsalou tells us that constraints of one type "often represents statistical patterns or personal preferences, which may be contradicted on occasion" (1992: 37). But note the qualification "often" since the inverse connection between speed and duration is included in this class. After introducing another type that includes the connections between surfing and oceans, and between skiing and snow, he writes that these constraints "may often represent statistical patterns and personal preferences, rather than necessary truths" (1992: 39). However, structural invariants were described as being only "relatively constant," while the speed-duration constraint is quite universal. Barsalou also holds that structural invariants are normative, and that constraints are not normative (1992: 37), but I have not been able to make consistent sense out of this distinction. Structural invariants and constraints can themselves be represented by frames (1992: 36, 40).

20 Barsalou takes it for granted that any theory of concepts must exhibit typicality effects. Thus he notes that "Like object taxonomies, attribute taxonomies exhibit typicality. For example, many people egocentrically perceive legs to be more typical of *means of locomotion* than *fins* or *wings*" (1992: 32). The philosophers I will discuss in the remainder of this section all reject NS accounts of concepts, and hold that concepts are open-ended. Andersen, Barker, Chen, and Nersessian explicitly adopt a Wittgensteinian family-resemblance view of concepts. Typicality effects follow, as a matter of course, from these views.

21 More recently Barsalou has integrated frames into a new kind of empiricist theory of concepts. Chen (2001) applies this version to a study of conceptual change in taxonomy; Prinz (2001) includes an empiricist aspect (but without frames) in a more complex theory of concepts. I will not pursue this topic here

for two reasons. First, the empiricist element depends on an account at the level of neural processing: "Perceptual symbols are *not* like physical pictures; nor are they mental images or any other form of conscious subjective experience. . . . Instead, they are records of the neural states that underlie perception" (Barsalou 1999: 588). As indicated in Ch. 1, I am not going to consider the neural embodiment of concepts in this book. Second, discussion of this approach has so far proceeded mainly in terms of concepts that are close to experience. Barsalou recognizes the need to address what he calls "abstract concepts" (among which he includes truth, falsity, negation, anger, and disjunction), but his discussions are limited. In the cases of truth, falsity, negation, and disjunction Barsalou emphasizes that he is addressing only a core intuitive sense of each notion, and that he omits the formal aspects. This new empiricism is at an early stage (Barsalou notes unsolved problems throughout his 1999 paper); its further development is worth watching.

22 Computer modeling of conceptual change is a central feature of Thagard's work, but I will not pursue that theme here.

23 Thagard mentions only first-order logic, but this is because of the specific examples he is considering. I noted above that he is prepared to incorporate implications based on higher-order properties into his framework.

**7 Conceptual Analysis I: Causation**

1 Bishop 1992 and Ramsey 1998 provide useful discussions of the significance of typicality effects for conceptual analysis.

2 Plato's dialogues include such interactions at least in their presentation.

3 Of course, Burks may be right about the prevailing concept. In this case, the remark in the text would be an example of Carnapian explication (cf. Sec. 1.3).

4 Torretti (1999: 131–32) discusses a similar distinction between determinism and the ordinary causal concept, although without the claim that there is something especially proper about the everyday version. Pearl (2002: 26–27) maintains that deterministic causation is more in tune with human intuition and everyday thought. Suppes does not consistently hold to this preference for everyday talk over scientific usage. For example, he maintains that the causal relata are events (see Sec. F) and, in response to Armstrong's objection that the concept of an event presupposes that of a cause, Suppes maintains that even if this is true in our ordinary experience, "There is a long scientific tradition that makes a clear and sustained effort to use the concept of event without introducing the concept of cause." This usage is particular clear in kinematics (1970: 69).

5 Some might think that $Pr(y|x) > Pr(y|\text{-}x)$ would be more appropriate, but this is equivalent to PSR (Eells 1991: 56).

6 It is unclear why determinism is a metaphysical, rather than a physical, thesis.

7 Eells says "very roughly" because, he argues, contextual factors must be considered, and we must be careful about how we understand change of probability. Suppes – along with many other advocates of probabilistic causality – defines "probability-change" in terms of conditional probabilities, but Eells argues that this holds only for populations, not for individuals (1991: 1–5, plus Chs. 2 and 6).

8 Mackie actually says that the full cause is "both necessary and sufficient for the effect (in the field)" (1980: 64) where the parenthetical remark refers to the notion of a causal field introduced earlier (1980: 34–36): a set of standing conditions that we do not specify in a particular case. The causal field is important when we are concerned with what is typically called the cause; it is not relevant to my concerns here.

9 More examples are provided by Armstrong (1999: 177) and Pearl (2001: 314–15).

10 At this point in his discussion Suppes is considering only what he calls "prima facie causality." There are cases in which it appears that $x$ boosts the probability of $y$, but an earlier event $z$ accounts equally well for the probability increase. In such cases $x$ is a "spurious cause" of $y$. Suppes defines a "genuine cause" as a prima facie cause that is not spurious (1970: 21–25), so all properties of a prima facie cause are also properties of a genuine cause.
11 Hume begins his discussion with the remark that this condition is subject to controversy (2001: 54).
12 Tooley also holds that, "There are a number of causal concepts" (252), and that while some of these can be reduced to others, which ones we take as basic depends on our views on several other topics, such as the nature of causal laws and whether we think that causal facts supervene on non-causal facts.
13 Mackie (1980: Ch. 7) provides a critical survey of proposals up to the early 1970s (the book was originally published in 1974).
14 In his 1987 book Tooley advocates a view that is intermediate between the type and token views, although he later rejects this view on grounds of simplicity (1990: 274).
15 Hume often writes of causal relations between *objects*, and it is an interesting question what he means (or, given his overall view, should mean) by "object." Still, there are many places in which he discusses causation in terms of events, and I will limit discussion to this reading, which is common.
16 Dowe (2000: 92) essentially agrees: "As in Salmon's theory, causality is treated fundamentally as a property of processes and interactions."
17 This view is motivated by Hume's remark: "*if the first object had not been, the second never had existed*" (1975: 76). David Lewis once proposed such an account of causation for event tokens (1973: 563).
18 Since terminology varies somewhat let me note that I describe a relation as "asymmetric" when C10 is valid, although some prefer "anti-symmetric"; in either case, PARENT is an example. I describe a relation as "non-symmetric" when $xRy$ implies nothing about $yRx$. Thus SISTER is non-symmetric; SIBLING is symmetric.
19 Sellars observes that we may have a high probability that a necessary relation holds (CDCM 270–71).
20 Pearl goes beyond other writers I have been considering in constructing a formal calculus for handling interventions; he sees this as an extension of standard probability calculus. Still, Pearl views his work as tracking the "ordinary conception of cause and effect" (2001: 35).
21 We should not forget the complexity of Hume's position. He does insist that causation implies a necessary connection and eventually traces the idea of necessary connection to an impression of reflection. This introduces a sense of "necessary" that is different from logical necessity.
22 In the following discussion I will be using "entailment" in two contexts. Thus far I have been considering what is entailed by premises that include $xCy$. In that context entailment is a property of an argument; entailment claims do not appear as a premise or conclusion in any of the labeled arguments examined above. In the present section entailment claims will also appear as premises and conclusions since I will be concerned with what is entailed by expressions of the form "$p$ entails $q$" (which I abbreviate as $pEq$). As long as this is kept in mind the double usage should not generate any confusion.
23 However, ENTAILMENT is not free of controversy. E2 holds in the most common logical systems, which include the principle that a contradiction entails every proposition. But this claim is rejected in paraconsistent logics. Advocates of these alternative logics place various restrictions on E2.
24 Clendinnen (1999: 187–89) offers a view of the status of causal concepts that has much in common with my view, although in this paper his main concern is "to make explicit the criteria by which we presently make causal distinctions" (191).

25 Even if Salmon and Suppes are right in their claim about ordinary discourse, the concepts of causation they discuss are not just analyses of that ordinary concept. It is most unlikely that an everyday concept includes the sophisticated use of probability calculus that they deploy

**8 Conceptual Analysis II: Epistemic Concepts**

1 Thus I would modify the passage from Sellars quoted above by inserting a single word so that it reads "we are not *just* giving an empirical description. . . . ".
2 "Logic" without a modifier will always refer here to standard formal logic, never to transcendental logic.
3 Kant's restriction of the epistemic role of synthetic a priori propositions to phenomena derives from his attempt to establish these propositions a priori. Since GAs are not established a priori, Kant's arguments do not extend to them. Among those who recognize a central role for GAs, there is substantial disagreement about whether they are subject to parallel limitations. Kuhn, for example, thinks that the role of paradigms in science implies an antirealist conclusion (see Ch. 11); Sellars deploys similar ideas in a realist program.
4 These are late texts that make explicit a view that was taken for granted in much earlier work.
5 It is worth asking whether the same circle applies to a priori justification. Many presumably a priori truths are not obvious – as the history of logic and mathematics clearly show. Indeed, if conceptual analysis is an a priori discipline, the disagreements among analysts provide pretty dramatic evidence of the lack of obviousness of a priori truths. We need principles of justification here too – principles that also need a justification. If the justification is a priori, the same kind of circle would seem to follow.
6 Kyburg (1977) disagrees. In an article that acknowledges Sellars and is reminiscent of C. I. Lewis, Kyburg argues (in effect) that GAs are analytic.
7 Whether this dependence of norms on decisions can by extended to logic and other fields are important questions that I cannot pursue here. Recent scholarship (e.g., Friedman 1999; Giere and Richardson 1996) has shown that at least some of the logical positivists held this view of logic. As Carnap put it, "In logic, there are no morals. Everyone is at liberty to build up his own logic, i.e., his own form of language, as he wishes" (1959: 52).
8 In case anyone is curious, in my view Plato is *not* seeking a definition of "knowledge" in this dialogue; he gives us that definition at the beginning: knowledge is a epistemic state that is infallible and of the real. The bulk of the dialogue is concerned with examining various epistemic states and considering whether they meet the definition. Perception is found to be infallible (according to the account of perception that Plato provides) and also of the real in one sense: its objects are real. But perception fails to meet the second criterion in another sense: it is not "of the real" because perception alone does not inform us *that its objects are real*; this requires that we go beyond perception. Three other states are then explored: belief and true belief fail because they are fallible. True belief plus an account never gets tested against the two criteria because all interpretations of "an account" that Plato considers are rejected as inadequate.
9 Although my aim is to focus on underlying concepts, not on what words mean in some body of discourse, practical considerations of communication make it unrealistic to introduce a new label for every concept. We must, for the most part, discuss these issues in existing language while seeking to avoid purely verbal matters.
10 Even these minimal conditions are subject to controversy. In *The Republic* Plato denied that knowledge implies belief – that is, he denied that knowledge is a kind

of enhanced belief. Rather, he argued, knowledge and belief are distinct cognitive states with distinct objects. We will consider a different objection shortly.

11 Here "theory" is a synonym for "analysis of a concept."

12 This type of definition is common in logic and mathematics. For example, a recursive definition of number might include as a base clause "zero is a number," and as a recursive clause (given the concept of a successor) that the successor of every number is a number.

13 Such complex clauses are common in analyses of everyday concepts. Pollock (1986: 181) considers it unlikely that the concepts we acquire as we grow up in a society and learn a language have such complexity.

14 Goldman is aware that we have the option of improving on available concepts. In a footnote (1992: 117, n. 8) Goldman considers his claim that reliable processes need not be perfectly reliable, and notes that he may face an analogue of the lottery paradox: a belief that arises from a series imperfectly reliable processes may be justified according to his account even though there is a very good chance that it is mistaken. He offers two alternative responses to this situation:

> we might simply indicate that the theory is intended to capture our ordinary notion of justifiedness, and this ordinary notion has been formed without recognition of this kind of problem. The theory is not wrong *as* a theory of the ordinary (naive) conception of justifiedness. On the other hand, if we want a theory to do more than capture the ordinary conception of justifiedness, it might be possible to strengthen the principles and avoid lottery-like analogues.
>
> (1992: 125)

15 Goldman considers this option in "Strong and Weak Justification" (1992: 130–31).

16 This view has clear affinities with contextualist accounts of justification, although I will not pursue the subject here. See, for example, Annis 1978; Cohen 1987; Henderson 1994. There is also work on contextualist theories of knowledge by philosophers who deny that knowledge requires justification.

17 For the remainder of this section I use "information" only in the sense in which there is no such thing as false information; cf. Dretske 1981.

18 Correspondence is a symmetric relation: where model and prototype correspond, we can also learn about the model from the prototype.

19 Consideration of implications indicates that a proposition may carry a great deal more information than appears at first glance – especially when taken in conjunction with other propositions. An axiom system that carries all the information about some subject in just a few propositions is old ideal which has been superseded (in most cases) since Gödel.

20 Clearly there are reflexive cases, but as I argued in Sec. 5.9, these are generally no more problematic than discussing English grammar in English. A limited number of cases produce paradoxes that must be handled. But, as Fitch (1946) argued, these should not be dealt with in ways that rule out a vast array of innocent and important cases.

21 Hooker, who treats truth as an inaccessible epistemic ideal, argues that we pursue truth by pursuing a variety of proxies: "security, explanatory and predictive power, scope and precision, and the like" (1995: 322, *et passim*). I include these among the means by which we pursue justification.

22 Crispin Wright (1999: 224–38) also arrives at a pluralist account of truth, although by a different path. I will not discuss the details of Wright's account, but I note an important point of general agreement: it is not enough to declare

that truth is a family resemblance concept; we must work out in detail the overlapping strands and differences among the various truth concepts.

23 The solar-neutrino problem provides a recent example (for reviews see Bahcall 1989, Ch. 10; Franklin 2001, Chs 8–9). The problem emerged in 1968 from a new test of an established theory, and led to more than three decades of further experimentation and theoretical analysis. The original experiment was developed to test the theory of how stars generate energy – a theory that was not actually doubted by the scientists involved. The motivation for the experiment derived from the newly developed ability to detect neutrinos, since the theory makes predictions about neutrinos that had never been tested. The conflict seems to have been resolved with a significant revision in the previously accepted theory of neutrinos – a revision whose further consequences are currently being tested.

24 I am building on an argument due to Crispin Wright (1999: 209–13).

## 9 Historical Studies I: Seventeenth-Century Physics

1 We saw in Sec. 2.1 that properly speaking each object we encounter is a mixture of elements; I will pick up this idea shortly. For the moment, attributions of an element to an object refer to the dominant element.

2 The Ptolemaic approach is a later development that arose in response to predictive failures of the Aristotelian model.

3 Galileo notes that at his latitude a point on a spinning earth moves through at least 16,000 miles in twenty-four hours (1967: 132). Later he gives the time of fall from a height of 100 yards as five seconds (1967: 223). These values imply that the stone would land a little more than 9/10 mile from the foot of the tower. Drake notes that the acceleration of gravity implied by Galileo's calculation is too low (1967: 484–85), although presumably the time of fall could be arrived at by means other than calculation. Galileo's value for the speed of rotation is also a bit low; a modern value for Pisa would be about 18,000 miles in 24 hours. A calculation using modern values (including a modern value for a yard, and ignoring air resistance) gives about 4.3 seconds for the time of fall. So the stone would land a little less than 9/10 mile from the foot of the tower.

4 As the passage continues, Galileo extends this conclusion to "impressed force."

5 There are places in *Dialogue* where Galileo expresses doubts about whether there is a single primary element constituting our planet (e.g., 400–3, 412–13). Nevertheless, we will see, the doctrine of elements plays a central role in his mechanics.

6 There is an additional argument for the motion of the earth from observations of sunspots (347–55).

7 As our discussion of the winds indicates, the claim that water is the *only* element that provides evidence of the earth's motion is a passing rhetorical exaggeration.

8 Galileo's account seems to yield incorrect results for a number of tidal phenomena, and much of the "Fourth Day" is devoted to resolving these anomalies. I will not discuss these details here, but I note that Galileo explicitly rejects any attempt to attribute the tides to action by the sun or moon (420, 445), and offers an account of why the motions of the moon are correlated with the tides (452–54). That Galileo's theory of the tides is wrong was recognized even by his favorite disciples.

9 See especially 31–32 where he also maintains that only circular motion can continue indefinitely. It should be clear that when Galileo writes about acceleration he is considering only changes in speed; he does not have the vector-concepts of velocity or acceleration.

10 This account of the elements improves on my 1976 discussion, although not in ways that affect my interpretation of the role of the elements in generating the winds and the tides.
11 The theory was discussed at least since the sixth century, and was developed in detail by Buridan in the fourteenth century (cf., Drake 1978: 9).
12 The two natural motions that a falling rock shares with the earth are neither inertial nor sustained by an impetus. I am concerned now only with projectiles – which include a stone falling on a moving ship.
13 Throughout this chapter I use "approximation" to include considerations that are also discussed under the rubrics "idealization" and "simplification." Some draw distinctions among these that do not concern us here.
14 The experiment involves rolling a bronze ball in a groove on an inclined plane; the groove has been prepared so as to minimize friction. The experimental arrangement is described in detail, and leads to an extended discussion of the relation between vertical fall and fall on an inclined plane.
15 The translations of PP in Descartes 1972 and Descartes 1985 are incomplete, omitting the content of articles the translators consider science, rather than philosophy. A complete English translation has been provided by Miller and Miller (Descartes 1991, henceforth MM); I rely primarily on this translation. Reynolds (Descartes 1988) translates those sections omitted by Haldane and Ross. In referring to PP I will give the number of the book, followed by the article; usually I include page numbers only when quoting a particular translation. Page references are to MM unless otherwise stated. A full citation will read, for example, (III 52: 110).
16 Tycho Brahe held that the traditional planets move around the sun and that the sun, carrying the planets, moves around the earth.
17 In *The World* Descartes began with a somewhat more diverse imaginary world: particles exist in "as many parts and shapes as we can imagine, and . . . each of its parts can take on as many motions as we can conceive" (1998: 23).
18 See Garber (1992a) for comparison of the laws of nature in *The World* and in the PP, along with an account of the development of Descartes' theory of motion.
19 See also Descartes' use of these terms in discussing reflection of light (2001: 75–77).
20 This radial determination plays a central role in Descartes' cosmology and, we will see, his account of weight.
21 Recall Galileo's argument that tangential motion is motion away from the center.
22 Figures 9.2 and 9.3 are simplified versions of Figures 8 and 6, respectively, of Descartes' *Optics* (2001: 75 and 78). I have left out some artistic features of Descartes' diagrams, as well as some parts that are irrelevant to the argument. I have also modified Figure 9.3 to include an item from Descartes' Figure 10. In the case of reflection, I have simplified Descartes' argument somewhat.
23 Descartes discusses decomposition of motion at II 32: 55. In note 23 MM point out that this is an ancient technique.
24 Gabbey (1980: 256) notes that Descartes has no mechanical account of how this rebound occurs.
25 In the second model the cloth is replaced by water. The third model concerns light passing into a medium in which its speed increases: Descartes implements this condition by having the ball hit again by a racket at the surface of this medium.
26 Figure 9.3 assumes that BE is shorter than the radius of the circle. Descartes recognizes that this need not be the case and maintains that if FEI falls outside the circle, we get total reflection of the light as it hits the boundary between the media (2001: 79). In the present case this implies that if light moves from air into

a material with a refractive index of two, total reflection occurs for any angle of incidence greater than sixty degrees.

27 It is this account of motion that allows Descartes to adopt the main features of Copernican cosmology while holding that the earth does not move: the earth is embedded in matter that engages in the daily and annual motion, but the earth is stationary with respect to this surrounding matter (III 16–19). See Garber 1992a, Ch. 6 for a detailed analysis of Descartes' account of motion and Slowik 1998, 1999b for a contrasting view.

28 The passage in square brackets indicates a modification of MM's translation on the basis of my reading of the French text. MM translate the last clause of this sentence as "it is turned aside in another direction, retaining its quantity of motion and changing only the direction of that motion." We will see that the shift between change of determination and change of direction makes no real difference in the present context, and MM's translation gives a more coherent reading of the entire paragraph. Still, it is not an accurate translation of the French text.

29 This is a different circular flow than the one that turns the earth.

30 Descartes introduces further GAs that cover these cases in a letter to Clerselier (17 February 1645, translated in Garber 1992a: 260–62). First he holds that:

> when two bodies having incompatible modes collide there must really be some change in these modes, in order to render them compatible, but that this change is always the least possible, that is, if they can become compatible by changing a certain quantity of these modes, a greater quantity of them will not be changed.

He then applies this principle to R4–R6, where the relevant incompatibility is between an object in motion and one at rest. Descartes claims that there are two means of eliminating this incompatibility: either "*B changes its entire determination*" or B moves C in such a way that they end up with the same speed. In R4, to move C, B would have to transfer "more than half its speed, and at the same time more than half of its determination to go from left to right, insofar as this determination is joined to speed." Instead, B "changes only its entire determination" and retains its speed. In R5, C is moved because B transfers "less than half of its own speed and less than half of the determination which is joined to it." In R6 "the change is made half in one way and half in the other." Descartes does not include these additional principles in the French edition of PP (1647). For discussion see Gabbey 1980: 263–65 and Garber 1992a: 246–48.

31 Garber reads the letter to Clerselier as suggesting a somewhat different conception of the force to resist being moved and notes that if we adopt his view, along with the remaining rules of impact, "the force for proceeding a moving body has is measured differently in colliding with a resting body than in colliding with another body in motion" (1992a: 244, n. 25).

32 Garber suggests that C's resistance to being moved is equal to the product of C's size and B's speed (1992a: 240). Gabbey considers two other proposals: the force with which C resists is equal to its size times the speed it would have if it were to move; and C's force to resist is equal to B's speed (1980: 269). But no measure that ignores the relative size of B and C will work.

33 Only the first two cases occur in the Latin version. See Garber 1992a: 234–53, for a comparison of the two versions and discussion of developments in the interim.

34 The passage in square brackets was added by MM; see Garber 1992a: 213 for discussion.

35 If one of a set of contraries must apply to an object, and the set contains only two members, then these are also contradictories, so Descartes' terminology is

not wrong. But we will see that a slightly different terminology is more illuminating.

36 Descartes also says that there is opposition between "rapidity of movement and slowness of movement (i.e., to the extent that this slowness partakes of the nature of rest) . . . " (II 44: 63). Gaukroger (2002: 106) says that this does not make sense; I agree.

37 Slowik (1999a: 187–92) offers a defense of Descartes' idealizations as not different in kind from those of later physicists. I will return to this topic in Sec. 9.4.

38 See Westfall 1971: 78–82 for discussion of a circular tendency. Gaukroger (1995: 246) and Shea (1991: 218–19) find evidence of circular inertia in Descartes' writings, and circular inertia implies circular determination.

39 In references to *Principia*, Newton's "Propositions" will be cited as *book.proposition*, e.g., III.5; definitions, laws, and other types of statements will be labeled accordingly. Page numbers will usually be given only for quotations and refer to Newton 1999. I will focus on the third edition.

40 Newton 1962: this is an unfinished manuscript generally referred to as *De Gravitatione*. It was long considered an early piece written in the 1660s or early 1670s, but Dobbs (1991: 139–46) challenges this dating, arguing that the piece was written in 1684 or early 1685, just before Newton began work on *Principia*. There is still controversy on this proposal. For example, Cohen (1999: 47) accepts the new dating but Stein disagrees (2002: 263 and n. 27, 272 and n. 39). Dobbs also argues that Newton's break with Descartes comes only with this manuscript (148, cf. 185–86).

41 Cf. Cohen 1983: 182–93, 1999: 43–49. Brackenridge (1995: 17–24) provides a useful discussion of Descartes' impact on Newton.

42 In his calculations Newton assumes uniform density for the earth; greater density at the equator will counterbalance the effect of greater distance from the center.

43 The two kinds of mass came to be known as "gravitational mass" and "inertial mass." Einstein identified them as a matter of principle in his new theory of gravitation – but this involved significant conceptual change from Newton's physics.

44 "With the air removed, as it is in Boyle's vacuum, resistance ceases, since a tenuous feather and solid gold fall with equal velocity in such a vacuum" (939). At the end of Book II, Sec. 7 Newton reports several experiments (mostly his own) on falling bodies: twelve in water, two in air (750–61). See Smith (2001: 272–82) for discussion.

45 Much later in *Principia*, when Newton is discussing essential properties of bodies, he writes: "That all bodies are movable and persevere in motion or in rest by means of certain forces (which we call forces of inertia) we infer from finding these properties in the bodies that we have seen" (795). This passage is also compatible with the view that the force of inertia appears only in response to an impressed force.

46 In the first edition of *Principia* Newton did not provide a proof; a minimal sketch of a proof was added in the second edition. There is still some controversy about whether Newton ever provided a formally correct proof of this result; see Cohen (1999: 135–36) for discussion.

47 Newton sometimes call satellites "secondary planets" and sometimes uses "planet" to encompass both sets.

48 Newton's resolution of the question – given in III.12 and its corollary – is not quite what any of his predecessors expected. Newton introduces the hypothesis (816) that the center of the world is at rest. He notes that the common center of gravity of the earth, sun, and planets is at rest (III.11), and "*The sun is engaged in continual motion but never recedes far from the common center of gravity of all the planets*" (III.12: 816). This common center of gravity "is to be considered the

center of the universe" (817). Since the sun, unlike the planets, is always close to this center, treating the sun as stationary is often a good approximation. At the end of *Principia* Newton speaks of "six primary planets" (940).

49 The apogee is the point on the moon's orbit that is farthest from the earth. The apogee does not move if it is always at the same place on the orbit.

50 Newton adds a sentence that somewhat confuses the issue: "For if gravity were different from this force, then bodies making for the earth by both forces acting together would descend twice as fast, and in the space of one second would by falling describe 30 1/6 Paris feet, entirely contrary to experience" (804). This sentence suggests a somewhat different view of Newton's aim: He never doubts that gravity acts on the moon, and his aim is to show that gravity is *sufficient* to account for the motion of the moon. But on this reading it is unclear why NP1 and NP2 are needed. One possibility is that this is an additional argument, which would be clearer if we read the initial word of the sentence as meaning "moreover." Densmore (1996: 307) indicates some doubts about the point of the sentence since we do not have the experience that Newton cites. Stein (1991: 210–13) suggests a very different reading: Newton has already established that the moon is governed by an inverse-square force in the preceding proposition; the point of III.4 is to demonstrate that terrestrial gravity obeys an inverse-square law.

51 This discussion has benefited significantly from Cohen's account of "the Newtonian Style" (e.g., 1983: *xii-xiii*, 15–16, 62–64, 99–109; 1999: 148–55), although I do not follow Cohen in all details. I avoid Cohen's label because it has generated disputes about the role of Newton and others in the development of this style, disputes that do not concern us here.

52 This contrasts with later accounts of Newtonian physics which hold that absolute time is central, while absolute space is not required – e.g., DiSalle 2002: 35.

53 I am not saying that Newton has proved the existence of absolute motion. All I am saying is that he provides an IC for some cases – as Aristotle does for natural and violent motions. This is not sufficient to establish that a conceptual system is instantiated.

**10 Historical Studies II: Interactions**

1 Teller (1995, Ch. 2) argues against the particle interpretation, approaching the question by examining the conceptual jobs that the particle concept is intended to perform. The use of analogies in the construction of quantum field theory is a pervasive theme of Teller's book.

2 There are theories in which neutrinos have mass and change type; recently these theories have received considerable empirical support. I will not discuss these challenges to SM here; the present discussion concerns an historical stage of a continuing body of research.

3 They are also called *generations*, a term deriving from the historical sequence in which the particles were discovered.

4 I want to stress that this principle specifies a limit on the simultaneous determination of *certain pairs* of properties; there are also pairs for which no such limitation holds. For example, there is no such limitation on the properties that determine the state of an electron in an atom. There is controversy over the exact interpretation of this principle, in particular about whether it tells us something about the measuring process or about the actual location and momentum of these particles. Physics textbooks often mix the two perspectives, and I will not debate the issue here. When I speak of properties that cannot be *determined* simultaneously I am using language that is intentionally ambiguous on this

point; it is for this reason that I have followed several other writers in using the term *indeterminacy* rather than *uncertainty*.

5 The full story of virtual particles is much more complex. For example, a sufficiently energetic virtual photon may transform into an electron and a positron, which will interact, annihilate each other, and produce a photon that will then be absorbed. All of this must occur in too brief a time for detection. There are many other more complex virtual processes. In doing calculations, the process discussed in the main text provides the first approximation and the major contribution to the outcome. The additional processes provide relatively small corrections. Computations of corrections can be difficult and time consuming, but they have been done to the relevant order in cases where experimental technique is capable of measuring the difference. Such tests provide impressive support for QED. Correction terms occur in all cases I discuss here, but I consider only the major contribution.

6 Additional problems with spin and statistics were also resolved by introducing neutrinos.

7 Chadwick identified neutrons in 1932; this resolved a number of problems about radioactivity and the nature of the nucleus (see Franklin 2001 for this history).

8 This was actually a unified theory of WI and SI in which beta decay – the only WI process then known – occurs as a result of decay of the SI carrier.

9 It is a pion, one of the many mesons built out of two quarks. While pions are exchanged between nuclear constituents, they are not the SI carriers.

10 The average kinetic energy of a molecule at room temperature is about .03 eV. The coupling constants are also equal at short distances.

11 In quantum mechanics $a$ and $b$ are complex numbers; $a^2$ stands for multiplication of $a$ by its complex conjugate, which yields a real number (recall Sec. 2.2).

12 The term "gauge" provides one example of a disconnect between later uses of a term and its original sense. "Gauge" was introduced by Weyl in 1918 to refer to a change in the scale used for measuring the magnitude of a vector. (See O'Raifeartaigh 1997 for Weyl's paper and discussion; O'Raifeartaigh notes the common use of "gauge" for the distance between the two tracks of a railroad.) That use has long passed but the term continues to be used, although with a new sense. Another example is "lepton," from the Greek word for "small" (Pais 1986: 450). The term originally referred to the lighter particles, but now refers to fundamental particles that are not affected by the strong interaction; some are more massive than some hadrons. A similar story can be told for the term "quantum theory" – see the final paragraph of *A2*. Those who infer the content of a physical theory from some everyday use of a term do so at their own intellectual peril.

13 I follow the discussion in R 129–34 which is illustrative and will serve well for present purposes. See R 135 for a description of how one might proceed in actual research. I mostly use R's notation in this section.

14 More precisely, $\psi$ is complex; to calculate a measurable quantity it must be multiplied by its complex conjugate. To introduce a phase change we multiply $\psi$ by $\exp(ik\varphi)$, where $\varphi$ is a number. When this modified $\psi$ is multiplied by its complex conjugate the phase terms cancel, so there is no change to the measurable quantity.

15 See Earman 2002 for a dissenting view.

16 Recall the discussion in Sec. 2.2 of the relation between factorials and the gamma function.

17 "WI" stands for both "the weak interaction," and "the theory of the weak interaction;" context will clarify which is relevant. "EW" stands for the electroweak theory that unifies WI and QED.

18 More precisely, rotations in real space are represented by orthogonal matrices. The defining feature of an orthogonal matrix is that its transpose – the result of

interchanging rows and columns – is its inverse. Unitary matrices are a generalization of orthogonal matrices in which matrix elements may be complex numbers. We construct the inverse of a unitary matrix by taking the transpose of the matrix *and* replacing each number by its complex conjugate.

19 "SU" stands for "special unitary." SU(2) has an important relation to another group, SO(3), which is the group of all rotations in ordinary 3D space. Consider two members of SU(2), which we may designate U and –U. Each element of –U is –1 times the corresponding element of U; for 2 × 2 matrices this will not change the value of the determinant. Each pair, {U, –U}, corresponds to the same member of SO(3). This relation is worth noting because it provides a visual analogy for the requirement of SU(2) symmetry: it is something like requiring that the interaction remain the same when we do something like rotating the interacting system. These remarks are part of a Sellarsian commentary.

20 We will see in our discussion of SI that it is possible to limit the range of an interaction while maintaining massless bosons.

21 See Cao 1997: 281–83 for the early history of this concept. It is more accurate to describe the symmetry as *hidden* since the symmetry of $\mathscr{L}_w$ is not eliminated. Rather, it is hidden by the symmetry-breaking mechanism.

22 The problem is that calculations yield infinite results. The same problem arose in the development of QED and was solved when Feynman, Schwinger, and Tomonaga independently developed the technique known as "renormalization" (see Schweber 1994; Teller 1995). When EW was proposed in 1967 it was not known to be renormalizable. This was established by 't Hooft in 1971.

23 Introduction of mass by means of the Higgs mechanism is required for 't Hooft's proof of renormalizability. The Higgs field is also responsible for all of the quark and lepton masses. Use of the Higgs mechanism is another analogical adaptation, drawing on ideas already present in other fields of physics.

24 Morrison (2000: 126) also concludes, independently of TC, that this mixing of the original $\gamma$ and Z results in a reconceptualization of each.

25 The familiar electric charge $Q = Y/2 + I_z$. $I_z$ is an isospin component (Sec. 10.3.5).

26 More precisely, parity is reflection through the origin: each spatial parameter is multiplied by –1. It seems that physicists had assumed conservation of parity in all interactions without any empirical evidence; cf. Pais 1986: 532–33 and R 212–13.

27 A singlet is analogous to a scalar; it is invariant with respect to SU(2) operations. The appropriate IRR is the trivial representation (*A6*) so that an SU(2) operation is equivalent to multiplication by one.

28 There are additional complications involving weak interactions of quarks that I will not consider here. In textbooks these will be found under the rubrics "Cabibbo angle," "GIM mechanism," and "CKM matrix" (e.g., R 195–202).

29 $\bar{\mathbf{3}}$ stands for the conjugate representation of **3**. For our purposes it is sufficient to note that certain characteristic quantum numbers of Equation 6 are negatives of those in **3**, so we have two distinct IRRs of the same size.

30 Six of the eight gluons change quark color in the way described. These are represented by matrices that have all zeros along the main diagonal. The remaining gluons are linear combinations of gluons consisting of a color and its anti-color. They are represented by matrices that have non-zero entries only on the main diagonal. Each of these matrices must have at least two non-zero entries on the diagonal since the trace must be zero (*A7*). But the linear combination consisting of have $r\bar{r} + g\bar{g} + b\bar{b}$ cannot represent a gluon since it has no color charge. (This is a consequence of the postulate that r + g + b has no color charge.) The upshot is that there are only two distinct gluons of this type.

31 In general, the magnitude of **p** X **r** equals the product of the magnitudes of **p**, **r**, and the sine of the angle between them. In our example **p** and **r** are perpendicular, so the sine is one.
32 Since Bohr's theory is set in the context of classical electromagnetic theory, it is inconsistent. This was a well-known problem that was resolved with the advent of quantum theory some twelve years later. Work in the period from Planck's original quantum hypothesis in 1900 until the development of modern quantum theory by Heisenberg (1925) and Schrödinger (1926) is generally referred to as "old quantum theory"; later developments are referred to as "quantum theory."
33 Since the electron is not literally moving, the rate of change of $\varphi$ will have to be given a different interpretation than it receives in classical physics. This suggests further conceptual differences that I will not pursue here.
34 In a spherically symmetrical case, such as the hydrogen atom, the choice of this axis is arbitrary. In more complex problems the choice will be suggested by physical features of the problem, such as the direction of a magnetic field.
35 There is no indeterminacy relation between these operators and the square of the angular momentum operator.
36 "The discovery of spin, though occurring between the beginnings of matrix mechanics and wave mechanics, was nevertheless an advance made entirely independent of quantum mechanics" (Pais 1986: 267).
37 The spinning-electron hypothesis was considered by Kronig a little before Goudsmit and Uhlenbeck, but Kronig explained his idea to Pauli, Heisenberg, and others who were skeptical, so he did not publish (Pais 1986: 280; Tomonaga 1997: 32–35). Ehrenfest, the teacher of Goudsmit and Uhlenbeck, was more encouraging.
38 We now know that Dirac's equation applies to objects with half-integral spin while the Klein-Gordon equation applies to objects with integral spin. This was not immediately clear in 1928. Indeed, Dirac maintained that objects with integral spin do not exist in nature. In 1934 Pauli and Weisskopf argued that nature has no reason to reject spin-zero particles, although at the time no candidates had been discovered (Tomonoga 1997: 108).
39 See R 39–42 for details. This result applies only to objects of half-integral spin; objects of integral spin yield the same wave function after a 360° rotation.
40 See Carson 1996a,b for a history of this concept, including changes as theory developed.
41 On Tomonaga's reading (1997: 169) only a charge is exchanged, not a particle. Carson gives a somewhat different reading (1996b: 103–6), as does Kragh (1999: 185–86).
42 Maxwell's work also resulted in an integration of optics into the new framework, although this was not part of the original project.
43 I discuss a modern version of Maxwell's equations. The exact formulation of the equations depends on the system of units used. I adopt a version that is common in particle physics (cf. Aitchison and Hey 1996: 42–43).
44 Bold-face indicates vectors; div and curl are vector operators.
45 One reason for the long-term interest in the existence of magnetic monopoles – isolated north or south poles – is that if they exist the formal similarity in these two equations would be complete.
46 The term "displacement current" survives, as does a successor of Maxwell's concept, although it is used only in certain special situations.
47 More precisely, the coordinate on each rotated axis is a linear combination of the original x, y, and z values; a parallel description holds in the converse direction.
48 There is a complication since "rotations" that mix space and time terms take place on a hyperbola, not on a circle. Although I will leave this feature of SR in

the background, it is central to the reasons (given in the main text) for the non-Euclidean nature of the SR spacetime.

49 Whether we treat the space terms or the time term as negative is of no significance as long as the assignment is consistent. Some prefer to multiply the time term by *i*. This results in all positive signs in the formula, but does not change the fact that the space and time terms are treated differently.

50 Linear operators must not be confused with linear equations. The linear equation $f(x) = ax + b$ can be thought of as describing a compound operation on $x$: first multiply by $a$ then add $b$. This is not a linear operation because adding $b$ is not linear; both criteria fail in this case.

51 All the expressions mentioned in this discussion can serve as state descriptions. Integration (with an additional complication) will allow us to move from, say, the velocity-time relation to the relation between location and time.

52 The nature of this change of eigenstate is a central problem in interpreting quantum theory; it is known as the *measurement problem*.

53 The point of the discussion remains if we also shift the origin, but the description is more complex.

54 Derivatives express rates of change, so the result follows from the linearity of the operation plus the point that the rate of change of a constant is zero. In the present case we need not specify what we are differentiating with respect to.

55 Recall Sec. 4.3. I repeat the definition using slightly different notation. Other items besides operations can form a group, but they do not concern me here.

56 The elements of G and H need not commute among themselves.

57 This section assumes more linear algebra than I have discussed in this appendix; those who have some experience with matrices and vectors should have encountered enough linear algebra for my purposes, even if they have not encountered the label.

58 For example, in the case of rotations each matrix represents rotation around a particular axis and the parameter gives the angle of a specific rotation.

59 More precisely, each transformation has the form $U = \exp(iH \cdot \alpha)$, where U is a transformation in the group U(n), H is vector consisting of the group generators, and $\alpha$ is a vector consisting of values of the group parameters. This can be approximated (via Taylor series) by $U = 1 + iH \cdot \alpha$.

### 11 Conceptual Change, Incommensurability, and Progress

1 I refer to *The Structure of Scientific Revolutions* as SSR; quotations are from the third edition (1996). Kuhn's later papers are collected in *The Road Since Structure* (2000); I cite papers by their date of initial publication, but give page references to the collection.

2 I am attempting to extract a more systematic account from Kuhn's assorted late papers than he provides – although there are some clear themes. However, detailed discussion of the crucial notion of a lexicon's *structure* (e.g., 1993: 239) is missing. We are told that this, and other topics, are discussed in the book Kuhn was writing at his death. Presumably the manuscript will be edited and published at some point.

3 His actual description is in terms of two ways of learning the concepts.

4 This form of incommensurability was stressed in Doppelt 1978. Recent discussions will be found in Bird 2002 and Brown 1996.

5 This point was especially emphasized by Popper in his *Logik der Forschung* (1934); cf. Popper (1992).

6 We might have to choose between continuing our study of the standard model and playing baseball, but that is a different issue. Although this is not a case of scientific theory choice, it does meet the constraint of involving genuine competitors since the issue is how to apportion time – a limited resource.

7 I urge that this remark is confused. If *no sense* can be made of this notion of reality, then we would be unable to understand what Kuhn is rejecting. TC provides an account of the content of this concept.

8 Kuhn discusses two situations in which logic and observation are not sufficient for theory choice. One occurs in normal science, where accepted GAs close the gap. A more severe case occurs in revolutionary situations where these GAs are among the items being challenged.

9 For example: "On my first reading of Thomas S. Kuhn's *The Structure of Scientific Revolutions* (1962) I was so deeply shocked at his repudiation of the distinction between the context of discovery and the context of justification that I put the book down without finishing it" (Salmon 1991: 325).

10 He approaches the point when he writes that a scientist's world is "determined jointly by the environment and the particular normal-scientific tradition that the student has been trained to pursue" (SSR 112), and by his frequent descriptions of the role of nature in producing anomalies. But the emphasis in these passages is usually on the role of the tradition, and Kuhn does not seriously pursue the role of items that are independent of our beliefs in scientific research.

11 As one indicator of this explosion consider that we no longer just have telescopes that gather light. In addition to these optical telescopes, we have radio, infra-red, ultraviolet, X-ray, and neutrino telescopes. I discuss this view of evidence at length elsewhere; see Brown 1987, 1995, 2001, and 2005.

12 There is a tension between this account of improving grounds for accepting theories and the pessimistic induction from the failures of previously well-supported theories. In Brown 1990 I argue that there are major defects in the pessimistic induction, and that – on inductive grounds – several contemporary scientific theories are better supported than the pessimistic conclusion. Aspects of this argument are further elaborated in Brown 2001.

# References


Abachi, S. *et al*. (1995) "Observation of the Top Quark," *Physical Review Letters*, 74: 2632–637.

Abe F. *et al*. (1995), "Observation of Top Quark Production in $\bar{p}p$ Collisions with the Collider Detector at Fermilab," *Physical Review Letters*, 74: 2626w–31.

Achinstein, P. (1979) "The Causal Relation," in P. French, T. Uehling, and H. Wettstein (eds) *Midwest Studies in Philosophy IV: studies in metaphysics*, Minneapolis, MN: University of Minnesota Press.

Aitchison, I. and Hey, A. (1996), *Gauge Theories in Particle Physics*, 2nd edition, Bristol: Institute of Physics Publishing.

Andersen, H. (1996) "Categorization, Anomalies, and the Discovery of Nuclear Fission," *Studies in the History and Philosophy of Modern Physics*, 27B: 463–92.

Andersen, H. and Nersessian, N. (2000) "Nomic Concepts, Frames, and Conceptual Change," in D. Howard (ed.) *PSA 98: Proceedings of the 1998 Biennial Meeting of the Philosophy of Science Association Part II*, *Supplement to Philosophy of Science*, 67, no. 3: S224–41.

Annis, D. (1978) "A Contextualist Theory of Epistemic Justification," *American Philosophical Quarterly*, 15: 213–19.

Antony, L. (1993) "Quine as Feminist: The radical import of naturalized epistemology," in L. Antony and C. Witt (eds) *A Mind of One's Own*, San Francisco, CA: Westview Press.

Aristotle (1995a) "Meteorology," trans. E. Webster, in J. Barnes (ed.) *The Complete Works of Aristotle*, Princeton, NJ: Princeton University Press.

——(1995b) "On Generation and Corruption," trans. H. Joachim, in J. Barnes (ed.) *The Complete Works of Aristotle*, Princeton, NJ: Princeton University Press.

——(1995c) "On the Heavens," trans. J. Stocks, in J. Barnes (ed.) *The Complete Works of Aristotle*, Princeton, NJ: Princeton University Press.

——(1995d) "Physics," trans. R. Hardie and R. Gaye, in J. Barnes (ed.) *The Complete Works of Aristotle*, Princeton, NJ: Princeton University Press.

Armstrong, D. (1999) "The Open Door: counterfactual versus singularist theories of causation," in H. Sankey (ed.) *Causation and Laws of Nature*, Dordrecht: Kluwer.

Armstrong, S., Gleitman, L., and Gleitman, H. (1999) "What Some Concepts Might Not Be," in E. Margolis and S. Laurence (eds) *Concepts: core readings*, Cambridge, MA: MIT Press.

Ayer, A. (1936) *Language, Truth and Logic*, New York: Dover.

——(1961) *The Foundations of Empirical Knowledge*, London: Macmillan.

——(1965) "Phenomenalism," in *Philosophical Essays*, London: Macmillan.

Bahcall, J. (1989) *Neutrino Astrophysics*, Cambridge: Cambridge University Press.
Barker, P. (2001) "Incommensurability and Conceptual Change During the Copernican Revolution," in P. Hoyningen-Huene and H. Sankey (eds) *Incommensurability and Related Matters*, Dordrecht: Kluwer.
Baron, M. (1969) *The Origins of the Infinitesimal Calculus*, New York: Dover.
Barsalou, L. (1992) "Frames, Concepts, and Conceptual Fields," in A. Lehrer and E. Kittay (eds) *Frames, Fields, and Contrasts*, Hillsdale, NJ: Earlbaum.
——(1999) "Perceptual Symbol Systems," *Behavioral and Brain Sciences*, 22: 577–609.
Barsalou, L. and Hale, C. (1993) "Components of Conceptual Representation: from feature lists to recursive frames," in I. Mechelen, J. Hampton, R. Michalski, and P. Theuns (eds) *Categories and Concepts*, New York: Academic Press.
Beck, L. (1968) "The Kantianism of Lewis," in P. Schilpp (ed.) *The Philosophy of C. I. Lewis*, La Salle, IL: Open Court.
Berkeley, G. (1948) *The Works of George Berkeley Bishop of Cloyne*, A. Luce and T. Jessop (eds) London: Thomas Nelson.
Berlin, I. (1965) "Empirical Propositions and Hypothetical Statements," in R. Swartz (ed.) *Perceiving, Sensing, and Knowing*, Garden City, NY: Anchor Books.
Bernstein, R. (1966–67) "Sellars' Vision of Man in the Universe," *Review of Metaphysics*, 20: 113–43 and 290–316.
Bird, A. (2002) "Kuhn's Wrong Turning," *Studies in History and Philosophy of Science*, 33A: 443–63.
Birkhoff, G. and MacLane, S. (1953) *A Survey of Modern Algebra*, revised edition, New York: Macmillan.
Bishop, M. (1992) "The Possibility of Conceptual Clarity in Philosophy," *American Philosophical Quarterly*, 29: 267–77.
Blanshard, B. (1939) *The Nature of Thought*, New York: Humanities Press.
——(1962) *Reason and Analysis*, LaSalle, IL: Open Court.
Bogen, J. and Woodward, J. (1988) "Saving the Phenomena," *Philosophical Review,* 97: 303–52.
Bonevac, D. (2002) "Sellars *vs*. the Given," *Philosophy and Phenomenological Research*, 64: 1–30.
Boole, G. (1958) *The Laws of Thought*, New York: Dover.
Boring, E. (1950) *A History of Experimental Psychology*, 2nd edition, New York: Appleton-Century-Crofts.
Boyer, C. (1959) *The History of the Calculus and its Conceptual Development*, New York: Dover.
——(1991) *A History of Mathematics*, 2nd edition, revised by U. Merzbach, New York: Wiley.
Boysen, S. (1993) "Counting in Chimpanzees: nonhuman principles and emergent properties of number," in S. Boysen, and E. Capaldi (eds) *The Development of Numerical Competence: animal and human models*, Hillsdale, NJ: Earlbaum.
Brackenridge, J. (1995) *The Key to Newton's Dynamics*, Berkeley, CA: University of California Press.
Braithwaite, R. (1953) *Scientific Explanation*, New York: Harper Torchbooks.
Brand, M. (1980) "Simultaneous Causation," in P. van Inwagen (ed.) *Time and Cause*, Dordrecht: Reidel.
Brandom, R. (1994) *Making it Explicit*, Cambridge, MA: Harvard University Press.
Brock, W. (1993) *The Norton History of Chemistry*, New York: Norton.

Brown, H. (1975) "Paradigmatic Propositions," *American Philosophical Quarterly*, 12: 85–90.
——(1976) "Galileo, the Elements, and the Tides," *Studies in History and Philosophy of Science*, 7: 337–51.
——(1978) "On Being Rational," *American Philosophical Quarterly*, 15: 241–48.
——(1979) *Perception, Theory and Commitment: the new philosophy of science*, Chicago, IL: University of Chicago Press.
——(1985) "Galileo on the Telescope and the Eye," *Journal of the History of Ideas*, 46: 487–501.
——(1986) "Sellars, Concepts and Conceptual Change," *Synthese*, 68: 275–307.
——(1987) *Observation and Objectivity*, New York: Oxford University Press.
——(1988) *Rationality*, London: Routledge.
——(1990) "Prospective Realism," *Studies in History and Philosophy of Science*, 21: 211–42.
——(1991) "Epistemic Concepts," *Inquiry*, 34: 323–51.
——(1992a) "Direct Realism, Indirect Realism, and Epistemology," *Philosophy and Phenomenological Research*, 52: 341–63.
——(1992b) "Response to Matheson's review of *Rationality*," *Social Epistemology*, 6: 45–55.
——(1993) "A Theory-Laden Observation *Can* Test the Theory," *British Journal for the Philosophy of Science*, 44: 555–59.
——(1994a) "Circular Justifications," in D. Hull, M. Forbes and R. Burian (eds) *PSA 1994* vol. 1, East Lansing, MI: The Philosophy of Science Association.
——(1994b) "Judgment and Reason: responses to Healy and Reiner and beyond," *Electronic Journal of Analytic Philosophy*, 2: 5.
——(1995) "Empirical Testing," *Inquiry*, 38: 353–99.
——(1996) "The Methodological Roles of Theory in Science," in B. Rhoads and C. Thorn (eds) *The Scientific Nature of Geomorphology*, Chichester: Wiley.
——(1999) "Why do Conceptual Analysts Disagree?" *Metaphilosophy*, 30: 33–59.
——(2000a) "Berkeley on the Conceivability of Qualities and Material Objects," in M. Gedney (ed.) *Proceedings of the Twentieth World Congress of Philosophy*, vol. 7, Bowling Green, OH: Philosophy Documentation Center.
——(2000b) "Review of A. Goldman, *Knowledge in a Social World*," *Philosophy of Science*, 67: 348–52.
——(2000c) "The Role of Judgment in Science," in W. Newton-Smith (ed.) *A Companion to the Philosophy of Science*, Oxford: Blackwell.
——(2001) "Incommensurability and Reality," in P. Hoyningen-Huene and H. Sankey (eds) *Incommensurability and Related Matters*, Dordrecht: Kluwer.
——(2005) "On the Epistemology of Theory-Dependent Evidence," in A. Raftopoulos (ed.) *Cognitive Penetrability of Perception*, Hauppauge, NY: Nova Science Publishers.
Bruzzaniti, G. and Robotti, N. (1989) "The Affirmation of The Concept of Isotopy and the Birth of Mass Spectrography," *Archives Internationales D'Histoire des Sciences*, 39: 309–34.
Buchdahl, G. (1969) *Metaphysics and the Philosophy of Science, the classical origins: Descartes to Kant*, Cambridge, MA: MIT Press.
Burge, T. (1979) "Individualism and the Mental," in P. French, T. Uehling, and H. Wettstein (eds) *Midwest Studies in Philosophy IV: studies in metaphysics*, Minneapolis, MN: University of Minnesota Press.

——(1982) "Other Bodies," in A. Woodfield (ed.) *Thought and Object*, Oxford: Clarendon Press.
——(1986) "Intellectual Norms and Foundations of Mind," *Journal of Philosophy*, 83: 697–720.
——(1993) "Concepts, Definitions, and Meaning," *Metaphilosophy*, 24: 309–25.
Burian, R. (1979) "Sellarsian Realism and Conceptual Change in Science," in P. Beri, R. Horstmann, and L. Krüger (eds) *Transcendental Arguments and Science*, Dordrecht: Reidel.
Burks, A. (1951) "The Logic of Causal Propositions," *Mind*, 60: 263–82.
Campbell, N. (1957) *Foundations of Science*, New York: Dover.
Cao, Tian Yu (1997) *Conceptual Developments of 20th Century Field Theories*, Cambridge: Cambridge University Press.
Carnap, R. (1950) *Logical Foundations of Probability*, Chicago, IL: University of Chicago Press.
——(1952) "Meaning Postulates," *Philosophical Studies*, 5: 65–73.
——(1956a) "Empiricism, Semantics, and Ontology," in *Meaning and Necessity*, Chicago, IL: University of Chicago Press.
——(1956b) "The Methodological Character of Theoretical Terms," in H. Feigl and M. Scriven (eds) *Minnesota Studies in the Philosophy of Science*, vol. I, Minneapolis, MN: University of Minnesota Press.
——(1959) *The Logical Syntax of Language*, trans. A. Smeaton, Chicago, IL: University of Chicago Press.
——(1963a) "Intellectual Autobiography," in P. Schilpp (ed.) *The Philosophy of Rudolph Carnap*, La Salle, IL: Open Court.
——(1963b) "Replies and Systematic Expositions," in P. Schilpp (ed.) *The Philosophy of Rudolph Carnap*, La Salle, IL: Open Court.
——(1996) "Testability and Meaning," in S. Sarkar (ed.) *Science and Philosophy in the Twentieth Century*, vol. 2, New York: Garland.
Carroll, L. (1895) "What the Tortoise Said to Achilles," *Mind*, 4: 278–80.
Carson, C. (1996a) "The Peculiar Notion of Exchange Forces – I: Origins in Quantum Mechanics, 1926–1928," *Studies in History and Philosophy of Science*, 27B: 23–45.
——(1996b) "The Peculiar Notion of Exchange Forces II – From Nuclear Forces to QED, 1929–1950," *Studies in History and Philosophy of Science*, 27B: 99–131.
Cartwright, N. (1983) *How the Laws of Physics Lie*, Oxford: Clarendon Press.
Chen, X. (2001) "Perceptual Symbols and Taxonomy Comparison," in J. Barrett and J. Alexander (eds) *PSA00: Proceedings of the 2000 Biennial Meeting of the Philosophy of Science Association Part II*, *Supplement to Philosophy of Science*, 68, no. 3: S200–212.
Chen, X. and Barker, P. (2000) "Continuity Through Revolutions: A Frame-Based account of Conceptual Change During Scientific Revolutions," in D. Howard (ed.) *PSA 98 Proceedings of the 1998 Biennial Meeting of the Philosophy of Science Association Part II*, *Supplement to Philosophy of Science*, 67, no. 3: S208–23.
Cheney, D. and Seyfarth, R. (1992) *How Monkeys see the World*, Chicago, IL: University of Chicago Press.
Child, J., Bertholle, L. and Beck, S. (1983) *Mastering the Art of French Cooking*, New York: Knopf.
Churchill, F. (1979) "Sex and the Single Organism: biological theories of sexuality in the mid-nineteenth century," *Studies in the History of Biology*, 3: 139–77.

Churchland, P. (1979) *Scientific Realism and the Plasticity of Mind*, Cambridge: Cambridge University Press.

——(1985) "Conceptual Progress and Word/World Relations: in search of the essence of natural kinds," *Canadian Journal of Philosophy*, 15: 1–17.

Clapham, C. (1996) *Oxford Concise Dictionary of Mathematic*, 2nd edition, Oxford: Oxford University Press.

Clark, A. (1997) *Being There*, Cambridge, MA: MIT Press.

Clavelin, M. (1974) *The Natural Philosophy of Galileo*, trans. A. Pomerans, Cambridge, MA: MIT Press.

Clendinnen, F. (1999) "Causal Dependence and Laws," in H. Sankey (ed.) *Causation and Laws of Nature*, Dordrecht: Kluwer.

Code, L. (1991) *What Can She Know?* Ithaca, NY: Cornell University Press.

Cohen, I. (1983) *The Newtonian Revolution*, Cambridge: Cambridge University Press.

——(1999) "A Guide to Newton's *Principia*," in I. Newton, *The Principia: mathematical principles of natural philosophy*, trans. I. Cohen and A. Whitman, Berkeley, CA: University of California Press.

Cohen, S. (1987) "Context and Social Standards," *Synthese*, 73: 3–26.

Colburn, T. (1999) "Software, Abstraction, and Ontology," *The Monist*, 82: 3–19.

Cottingham, W. and Greenwood, D. (1998) *An Introduction to the Standard Model of Particle Physics*, Cambridge: Cambridge University Press.

Crane, T. (1991) "All the Difference in the World," *Philosophical Quarterly*, 41: 1–25.

Cushing, J. (1990) "Foundational Problems in and Methodological Lessons from Quantum Field Theory," in H. R. Brown and R. Harré (eds) *Philosophical Foundations of Quantum Field Theory*, Oxford: Clarendon Press.

——(1991) "Quantum Theory and Explanatory Discourse: end game for understanding?" *Philosophy of Science*, 58: 337–58.

——(1994) *Quantum Mechanics*, Chicago, IL: University of Chicago Press.

Damasio, A. (1994) *Descartes' Error*, New York: Avon.

Davidson, D. (1974) "Causal Relations," in T. Beauchamp (ed.) *Philosophical Problems of Causation*, Encino, CA: Dickenson.

——(1984) "On The Very Idea of a Conceptual Scheme," in *Inquiries into Truth and Interpretation*, Oxford: Clarendon Press.

Davis, W. (1988) "Probabilistic Theories of Causation," in J. Fetzer (ed.) *Probability and Causality*, Dordrecht: Reidel.

Davis, M. and Hersh, R. (1972) "Nonstandard Analysis," *Scientific American*, 226, no. 6: 78–85.

de Gandt, F. (1995) *Force and Geometry in Newton's Principia*, trans. W. Curtis, Princeton, NJ: Princeton University Press.

Deetz, J. and Deetz, P. (2000) *The Times of their Lives: life, love, and death in Plymouth Colony*, New York: Freeman.

Delaney, C. *et al.* (1977) *The Synoptic Vision: essays on the philosophy of Wilfrid Sellars*, Notre Dame, IN: University of Notre Dame Press.

Densmore, D (1996) *Newton's Principia: the central argument*, Santa Fe, NM: Green Lion Press.

Descartes, R. (1972) *The Philosophical Works of Descartes*, trans. E. Haldane and G. Ross, Cambridge: Cambridge University Press.

——(1985) *The Philosophical Writings of Descartes*, trans. J. Cottingham, R. Stoothoff and D. Murdoch, Cambridge: Cambridge University Press.

——(1988) *Principles of Philosophy*, trans. B. Reynolds, Lewiston, NY: Edwin Mellen.

——(1991) *Principles of Philosophy*, trans. V. Miller and R. Miller, Dordrecht: Kluwer.
——(1998) *The World and Other Writings*, trans. S. Gaukroger, Cambridge: Cambridge University Press.
——(2001) *Discourse on Method, Optics, Geometry, and Meteorology*, revised edition, trans. P. Olscamp, Indianapolis, IN: Hackett.
de Sousa, R. (1984) "The Natural Shiftiness of Natural Kinds," *Canadian Journal of Philosophy*, 14: 561–80.
DiSalle, (2002) "Newton's Philosophical Analysis of Space and Time," in I. Cohen and G. Smith (eds) *The Cambridge Companion to Newton*, Cambridge: Cambridge University Press.
Dobbs, B. (1991) *The Janus Faces of Genius: the role of alchemy in Newton's thought*, Cambridge: Cambridge University Press.
Donnellan, K. (1966) "Reference and Definite Descriptions," *Philosophical Review,* 75: 281–304.
Doppelt, G. (1978) "Kuhn's Epistemological Relativism: an interpretation and defense," *Inquiry*, 27: 33–86.
Dowe, P. (1992) "Wesley Salmon's Process Theory of Causality and the Conserved Quantity Theory," *Philosophy of Science*, 59: 195–216.
——(1995) "Causality and Conserved Quantities: a reply to Salmon," *Philosophy of Science*, 62: 321–33.
——(2000) *Physical Causation*, Cambridge: Cambridge University Press.
Drake, S. (1978) *Galileo at Work*, Chicago, IL: University of Chicago Press.
Dretske, F. (1981) *Knowledge and the Flow of Information*, Cambridge, MA: MIT Press.
Ducasse, C. (1926) "On the Nature and Observability of the Causal Relation," *Journal of Philosophy*, 23: 57–68.
——(1951) "Analysis of the Causal Relation," in T. Beauchamp (ed.) *Philosophical Problems of Causation*, Encino, CA: Dickenson.
Dummett, M. (1954) "Can an Effect Precede its Cause?" *Aristotelian Society Supplementary Volume*, 28: 27–44.
——(1964) "Bringing about the Past," *Philosophical Review*, 73: 338–59.
Dunham, W. (1999) *Euler: the Master of us all*, Washington, DC: Mathematical Association of America.
Dupré, J. (1993) *The Disorder of Things*, Cambridge, MA: Harvard University Press.
Earman, J. (2002) "Gauge Matters," in J. Barrett and J. Alexander (eds) *PSA00: Proceedings of the 2000 Biennial Meeting of the Philosophy of Science Association Part I*, *Supplement to Philosophy of Science*, 68, no. 3: S209–20.
Eells, E. (1991) *Probabilistic Causality*, Cambridge: Cambridge University Press.
Eiseley, L. (1961) *Darwin's Century*, Garden City, NY: Anchor Books.
Emsley, J. (2001) *Nature's Building Blocks*, Oxford: Oxford University Press.
Etchemendy, J. (1990) *The Concept of Logical Consequence*, Cambridge, MA: Harvard University Press.
Farley, J. (1981a) "Generation-Reproduction," in W. Bynum, E. Browne, and R. Porter (eds) *Dictionary of the History of Science*, Princeton, NJ: Princeton University Press.
——(1981b) "Sperm," in W. Bynum, E. Browne, and R. Porter (eds) *Dictionary of the History of Science*, Princeton, NJ: Princeton University Press.
——(1982) *Gametes and Spores: ideas about sexual reproduction, 1750–1914*, Baltimore, MD: Johns Hopkins Press.

Feigl, H. (1956) "Some Major Issues and Developments in the Philosophy of Science of Logical Empiricism," in H. Feigl and M. Scriven (eds) *Minnesota Studies in the Philosophy of Science*, vol. IV, Minneapolis, MN: University of Minnesota Press.

——(1970) "The 'Orthodox' View of Theories," in M. Radner and S. Winokur (eds) *Minnesota Studies in the Philosophy of Science*, vol. IV, Minneapolis, MN: University of Minnesota Press.

Fenwick, L. (1998) *Private Choices, Public Consequences: Reproductive Technology and the New Ethics of Conception Pregnancy, and Family*, New York: Dutton Books.

Feyerabend, P. (1962) "Explanation, Reduction, and Empiricism," in H. Feigl and G. Maxwell (eds) *Minnesota Studies in the Philosophy of Science*, vol. III, Minneapolis,MN: University of Minnesota Press.

——(1975) *Against Method*, London: New Left Books.

Field, H. (1973) "Theory Change and the Indeterminacy of Reference," *Journal of Philosophy*, 70: 462–81.

Finocchiaro, M. (1980) *Galileo and the Art of Reasoning*, Dordrecht: Reidel.

Fitch, F. (1946) "Self-Reference in Philosophy," *Mind*, 65: 64–73.

Fodor, J. (1975) *The Language of Thought*, Cambridge, MA: Harvard University Press.

——(1988) *Psychosemantics*, Cambridge, MA: MIT Press.

——(1995) *The Elm and the Expert*, Cambridge, MA: MIT Press.

——(1998) *Concepts: where cognitive science went wrong*, Oxford: Clarendon Press.

Fodor, J. and Lepore, E. (1992) *Holism: a shopper's guide*, Oxford: Blackwell.

Foley, R. (1987) *The Theory of Epistemic Rationality*, Cambridge, MA: Harvard University Press.

——(1995) "Analysis," in R. Audi (ed.) *The Cambridge Dictionary of Philosophy*, Cambridge: Cambridge University Press.

Foster, J. and Sen, A. (1997) "On Economic Inequality After a Quarter Century," in A. Sen *On Economic Inequality, Expanded Edition*, Oxford: Clarendon Press.

Franklin, A. (1986) *The Neglect of Experiment*, Cambridge: Cambridge University Press.

——(1993) *The Rise and Fall of the Fifth Force*, New York: American Institute of Physics.

——(2001) *Are There Really Neutrinos? An evidential history*, Cambridge, MA: Perseus.

Frege, G. (1997) "On Concept and Object," trans. P. Geach, in M. Beaney (ed.) *The Frege Reader*, Oxford: Blackwell Publishers.

Friedman, M. (1999) *Reconsidering Logical Positivism*, Cambridge: Cambridge University Press.

Gabbey, A. (1971) "Force and Inertia in Seventeenth-Century Dynamics," *Studies in History and Philosophy of Science*, 2: 1–67.

—(1980) "Force and Inertia in the Seventeenth Century: Descartes and Newton," in S. Gaukroger (ed.) *Descartes: Philosophy, Mathematics, and Physics*, Brighton: Harvester Press.

Galileo (1960) *On Motion and On Mechanics*, trans. I. Drabkin and S. Drake, Madison, WI: University of Wisconsin Press.

——(1967) *Dialogue Concerning the Two Chief World Systems*, trans. S. Drake, Berkeley, CA: University of California Press.

——(1974) *Two New Sciences*, trans. S. Drake, Madison, WI: University of Wisconsin Press.

Galison, P. (1987) *How Experiments End*, Chicago: University of Chicago Press.
——(1997) *Image and Logic*, Chicago, IL: University of Chicago Press.
Garber, D. (1992a) *Descartes' Metaphysical Physics*, Chicago, IL: University of Chicago Press.
——(1992b) "Descartes' Physics," in J. Cottingham (ed.) *The Cambridge Companion to Descartes*, Cambridge: Cambridge University Press.
Gasking, D. (1955) "Causation and Recipes," *Mind*, 64: 479–87.
Gasking, E. (1967) *Investigations into Generation 1651–1828*, London: Hutchinson.
Gaukroger, S. (1995) *Descartes: An Intellectual Biography*, Oxford: Clarendon Press.
——(2002) *Descartes' System of Natural Philosophy*, Cambridge: Cambridge University Press.
Gettier, E. (1963) "Is Knowledge Justified True Belief?" *Analysis*, 23: 121–23.
Giere, R. (1988) *Explaining Science*, Chicago, IL: University of Chicago Press.
——(2000) "Naturalism," in W. Newton-Smith (ed.) *A Companion to the Philosophy of Science*, Oxford: Blackwell.
Giere, R. and Richardson, A. (eds) (1996) *Origins of Logical Empiricism*, Minneapolis, MN: University of Minnesota Press.
Glock, H. (2000) "Animals, Thoughts and Concepts," *Synthese*, 123: 35–64.
Goble, L. (ed.) (2001) *The Blackwell Guide to Philosophical Logic*, Oxford: Blackwell Publishers.
Goldman, A. (1986) *Epistemology and Cognition*, Cambridge, MA: Harvard University Press.
——(1992) *Liaisons*, Cambridge, MA: MIT Press.
——(1999) *Knowledge in a Social World*, Oxford: Clarendon Press.
Goldman, A. and Pust, J. (1998) "Philosophical Theory and Intuitional Evidence," in M. DePaul and W. Ramsey (eds) *Rethinking Intuition*, New York: Rowman and Littlefield.
Gould, S. (1996) "The Model Batter: extinction of the .400 hitter and the improvement of baseball," in *Full House*, New York: Three Rivers Press.
Graham, G. and Horgan, T. (1998) "Southern Fundamentalism and the End of Philosophy," in M. DePaul and W. Ramsey (eds) *Rethinking Intuition*, New York: Rowman and Littlefield.
Griffin, D. (1994) *Animal Minds*, Chicago, IL: University of Chicago Press.
Guiot, J. (1985) "Zur Entdeckung Der Ultravioletten Strahlen Durch Johann Wilhelm Ritter," *Archives Internationale d'Historie des Sciences*, 35: 346–56.
Hacking, I. (1983) *Representing and Intervening*, Cambridge: Cambridge University Press.
Hall, N. (2000) "Causation and the Price of Transitivity," *Journal of Philosophy*, 97: 198–222.
Hanson, N. (1958) *Patterns of Discovery*, Cambridge: Cambridge University Press.
Harman, G. (1982) "Conceptual Role Semantics," *Notre Dame Journal of Formal Logic*, 23: 242–56.
——(1999) "(Nonsolipsistic) Conceptual Role Semantics," in *Reasoning, Meaning, and Mind*, Oxford: Clarendon Press.
Harper, W. (2002) "Newton's Argument for Universal Gravitation," in I. Cohen and G. Smith (eds) *The Cambridge Companion to Newton*, Cambridge: Cambridge University Press.
Hassani, S. (1999) *Mathematical Physics*, New York: Springer.

Hausman, D. (1988) "Can Hume's use of a Simple/Complex Distinction be made Consistent?" *Hume Studies*, 14: 424–28.

Heathcote, A. (1989) "A Theory of Causality: causality = interaction (as defined by a suitable quantum field theory)," *Erkenntnis*, 31: 77–108.

Hebb, D. (1949) *Organization of Behavior*, New York: Wiley.

Hempel, C. (1952) *Fundamentals of Concept Formation in Empirical Science*, Chicago, IL: University of Chicago Press.

——(1963) "Implications of Carnap's Work for the Philosophy of Science," in P. Schilpp (ed.) *The Philosophy of Rudolph Carnap*, La Salle, IL: Open Court.

——(1965) *Aspects of Scientific Explanation*, New York: Free Press.

——(1970) "On the Standard Conception of Scientific Theories," in M. Radner and S. Winokur (eds) *Minnesota Studies in the Philosophy of Science*, vol. IV, Minneapolis, MN: University of Minnesota Press.

Henderson, D. (1994) "Epistemic Competence and Contextualist Epistemology," *Journal of Philosophy*, 91: 627–49.

Herschel, W. (1800) "Experiments on the Refrangibility of the invisible Rays of the Sun," *Philosophical Transactions of the Royal Society of London*: 284–92.

Hesse, M. (1966) *Models and Analogies in Science*, Notre Dame, IN: University of Notre Dame Press.

——(1970a) "An Inductive Logic of Theories," in M. Radner and S. Winokur (eds) *Minnesota Studies in the Philosophy of Science*, vol. IV, Minneapolis, MN: University of Minnesota Press.

——(1970b) "Is There and Independent Observation Language?" in R. Colodny (ed.) *The Nature and Function of Scientific Theories*, Pittsburgh, PA: University of Pittsburgh Press.

Holmes, H. (1992) "To Freeze or not to Freeze: Is that an option," in H. Holmes (ed.) *Issues in Reproductive Technology I*, New York: Garland.

Home, H. (1996) "Preliminary Discourse, Concerning the Origin of Men and of Languages," in H. Augstein (ed.) *Race: The Origins of an Idea, 1760–1850*, Bristol: Thoemmes Press.

Hooker, C. (1985) "Surface Dazzle, Ghostly Depths," in Churchland, P. and Hooker, C. (eds) *Images of Science*, Chicago, IL: University of Chicago Press.

——(1987) *A Realistic Theory of Science*, Albany, NY: State University of New York Press.

——(1995) *Reason, Regulation, and Realism*, Albany, NY: State University of New York Press.

Horwich, P. (1998) *Truth*, Oxford: Clarendon Press.

Hoyningen-Huene, P. (1993) *Reconstructing Scientific Revolutions*, trans. A. Levine, Chicago, IL: University of Chicago Press.

Hoyningen-Huene, P. and Sankey, H. (eds) (2001) *Incommensurability and Related Matters*, Dordrecht: Kluwer.

Hull, D. (1988) *Science as a Process*, Chicago, IL: University of Chicago Press.

Hume, D. (1975) *Enquiries concerning Human Understanding and Concerning the Principles of Morals*, 3rd edition, L. A. Selby-Bigge (ed.), revised by P. H. Nidditch, Oxford: Clarendon Press.

——(2001) *A Treatise of Human Nature*, D. Norton and M. Norton (eds) Oxford: Oxford University Press.

Humphreys, P. (2000) "Causation," in W. Newton-Smith (ed.) *A Companion to the Philosophy of Science*, Oxford: Blackwell.

Jackman, H, (1999) "Moderate Holism and the Instability Thesis," *American Philosophical Quarterly*, 36: 361–69.
Jackson, F. (1998) *From Metaphysics to Ethics: a defence of conceptual analysis*, Oxford: Clarendon Press.
Jammer, M. (1999) *Concepts of Force*, New York: Dover.
Judson, O. (2002) *Dr. Tatiana's Sex Advice to All Creation*, New York: Metropolitan.
Kant, I. (1963) *Critique of Pure Reason*, trans. N. Smith, London: Macmillan.
Kearney, B. (1998) *High-Tech Conception*, New York: Bantam.
Keil, F. and Wilson, R. (2000) "The Concept Concept: the wayward path of cognitive science," *Mind and Language*, 15: 308–18.
Kellert, S. (1994) *In the Wake of Chaos*, Chicago, IL: University of Chicago Press.
Kim, J. (1995) "Causation," in R. Audi (ed.) *The Cambridge Dictionary of Philosophy*, Cambridge: Cambridge University Press.
Kitcher, P. (1978) "Theories, Theorists, and Theoretical Change," *Philosophical Review*, 87: 519–47.
——(1983) *The Nature of Mathematical Knowledge*, New York: Oxford University Press.
——(1992) "The Naturalists Return," *Philosophical Review*, 101: 53–114.
——(1993) *The Advancement of Science*, New York: Oxford University Press.
Kline, A. (1980) "Are there Cases of Simultaneous Causation?" in P. Asquith and R. Giere (eds.) *PSA 1980*, East Lansing, MI: The Philosophy of Science Association.
Kline, M. (1972) *Mathematical Thought from Ancient to Modern Times*, New York: Oxford University Press.
——(1980) *Mathematics: The Loss of Certainty*, New York: Oxford University Press.
Konopinski, E. (1981) *Electromagnetic Fields and Relativistic Particles*, New York: McGraw-Hill.
Kornblith, H. (1995) *Inductive Inference and Its Natural Ground*, Cambridge, MA: MIT Press.
Kosso, P. (1989) *Observability and Observation in Physical Science*, Dordrecht: Kluwer.
Kostro, L. (2000) *Einstein and the Ether*, Montreal: Apeiron.
Kragh, H. (1999) *Quantum Generations*, Princeton, NJ: Princeton University Press.
——(2000) "Conceptual Changes in Chemistry: the notion of a chemical element, ca. 1900–925," *Studies in History and Philosophy of Science*, 31B: 435–50.
Kripke, S. (1980) *Naming and Necessity*, Cambridge, MA: Harvard University Press.
——(1982) *Wittgenstein on Rules and Private Language*, Cambridge, MA: Harvard University Press.
Kuhn, T. (1962) *The Structure of Scientific Revolutions*, Chicago, IL: University of Chicago Press.
——(1983) "Commensurability, Comparability, Communicability," in *The Road Since Structure* (2000), Chicago, IL: University of Chicago Press.
——(1987) "What Are Scientific Revolutions?" in *The Road Since Structure* (2000), Chicago, IL: University of Chicago Press.
——(1989) "Possible Worlds in History of Science," in *The Road Since Structure* (2000), Chicago, IL: University of Chicago Press.
——(1991a) "The Natural and the Human Sciences," in *The Road Since Structure* (2000), Chicago, IL: University of Chicago Press.
——(1991b) "The Road since *Structure*," in *The Road Since Structure* (2000), Chicago, IL: University of Chicago Press.

——(1991c) "The Trouble with the Historical Philosophy of Science," in *The Road Since Structure* (2000), Chicago, IL: University of Chicago Press.
——(1993) "Afterwords," in *The Road Since Structure* (2000), Chicago, IL: University of Chicago Press.
——(1996) *The Structure of Scientific Revolutions*, 3rd edition, Chicago, IL: University of Chicago Press.
Kyburg, H. (1977) "All Acceptable Generalizations are Analytic," *American Philosophical Quarterly*, 14: 201–10.
Lakatos, I. (1970) "Falsification and the Methodology of Scientific Research Programmes," in I. Lakatos and A. Musgrave (eds) *Criticism and the Growth of Knowledge*, Cambridge: Cambridge University Press.
Lakoff, G. (1987) *Women, Fire, and Dangerous Things*, Chicago, IL: University of Chicago Press.
Larson, D. (ed.) (1996) *Mayo Clinic Family Health Book*, 2nd edition, New York: William Morrow and Co.
Lau, J. (2003) "Externalism About Mental Content", in E. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*. Online. Available HTTP: http://plato.stanford.edu/entries/content-externalism/ (accessed 05/01/06).
Laudan, L. (1984) *Science and Values*, Berkeley, CA: University of California Press.
Laudan, L. *et al*. (1986) "Scientific Change: Philosophical Models and Historical Research," *Synthese*, 69: 141–223.
Lee, J. (1988) "The Nontransitivity of Causation," *American Philosophical Quarterly,* 25: 87–94.
Leicester, H. (1971) *The Historical Background of Chemistry*, New York: Dover.
Levi, E. (1949) *An Introduction to Legal Reasoning*, Chicago, IL: University of Chicago Press.
Lewis, C. I. (1946) *An Analysis of Knowledge and Valuation*, La Salle, IL: Open Court.
——(1956) *Mind and the World Order*, New York: Dover.
——(1968a) "Autobiography," in P. Schilpp (ed.) *The Philosophy of C. I. Lewis*, La Salle, IL: Open Court.
——(1968b) "Replies to my Critics," in P. Schilpp (ed.) *The Philosophy of C. I. Lewis*, La Salle, IL: Open Court.
——(1970) *Collected Papers*, J. Goheen and J. Mothershead (eds) Stanford, CA: Stanford University Press.
Lewis, D. (1973) "Causation," *Journal of Philosophy*, 70: 556–67.
Lightstone, A. (1978) *Mathematical Logic: an introduction to model theory*, H. Enderton (ed.), New York: Plenum.
Lloyd, E. (1994) *The Structure and Confirmation of Evolutionary Theory*, Princeton, NJ: Princeton University Press.
Locke, J. (1984) *An Essay Concerning Human Understanding*, P. H. Nidditch (ed.) Oxford: Clarendon Press.
Longino, H. (1990) *Science as Social Knowledge*, Princeton, NJ: Princeton University Press.
Mackie, J. (1980) *The Cement of the Universe: a study of causation*, Oxford: Clarendon Press.
Maclaurin, C. (1968) *An Account of Sir Isaac Newton's Philosophical Discoveries*, New York: Johnson Reprint Corporation.
Maienschein, J. (1981a) "Development," in W. Bynum, E. Browne, and R. Porter (eds) *Dictionary of the History of Science*, Princeton, NJ: Princeton University Press.

——(1981b) "Encapsulation," in W. Bynum, E. Browne, and R. Porter (eds) *Dictionary of the History of Science*, Princeton, NJ: Princeton University Press.
——(1981c) "Epigenesis/preformation," in W. Bynum, E. Browne, and R. Porter (eds) *Dictionary of the History of Science*, Princeton, NJ: Princeton University Press.
——(1981d) "Germ," in W. Bynum, E. Browne, and R. Porter (eds) *Dictionary of the History of Science*, Princeton, NJ: Princeton University Press.
——(1981e) "Ovism/animalculism," in W. Bynum, E. Browne, and R. Porter (eds) *Dictionary of the History of Science*, Princeton, NJ: Princeton University Press.
Maor, E. (1995) *e: the story of a number*, Princeton, NJ: Princeton University Press.
Margolis, E. and Laurence, S. (1999) "Concepts and Cognitive Science," in E. Margolis and S. Laurence (eds) *Concepts: core readings*, Cambridge, MA: MIT Press.
Mayr, E. (1982) *The Growth of Biological Thought*, Cambridge, MA: Harvard University Press.
——(1991) *One Long Argument: Charles Darwin and the genesis of modern evolutionary thought*, Cambridge, MA: Harvard University Press.
McMullin, E. (1967) "Introduction," in *Galileo: Man of Science*, E. McMullin (ed.) New York: Basic Books.
Medin, D. (1989) "Concepts and Conceptual Structure," *American Psychologist*, 44: 1469–481.
Mellor, H. (1977) "Natural Kinds," *British Journal for the Philosophy of Science*, 28: 299–312.
——(1995) *The Facts of Causation*, London: Routledge.
Menzies, P. (1989) "A Unified Account of Causal Relata," *Australasian Journal of Philosophy*, 67: 59–62.
Mill, J. (1868) *A System of Logic*, 7th edition, London: Longmans, Green, Reader, and Dyer.
Miller, R. (1987) *Fact and Method*, Princeton, NJ: Princeton University Press.
Moore, G. (1962) *Some Main Problems of Philosophy*, New York: Collier.
Morreall, J. (1982) "Hume's Missing Shade of Blue," *Philosophy and Phenomenological Research*, 42: 407–15.
Morrison, M. (2000) *Unifying Scientific Theories: physical concepts and mathematical structures*, Cambridge: Cambridge University Press.
Murphy, G. (2002) *The Big Book of Concepts*, Cambridge, MA: MIT Press.
Nagel, E. (1961) *The Structure of Science*, New York: Harcourt, Brace, & World.
Nahin, P. (1998) *An Imaginary Tale: the story of* $\sqrt{-1}$, Princeton, NJ: Princeton University Press.
Nauenberg (2001) "Newton's Perturbation Methods For the Three-Body Problem and Their Application to Lunar Motion," in J. Buchwald and I. Cohen (eds) *Isaac Newton's Natural Philosophy*, Cambridge, MA: MIT Press.
Nelson, H. (1992) "Scrutinizing Surrogacy," in H. Holmes (ed.) *Issues in Reproductive Technology I*, New York: Garland.
Nersessian, N. (1984) *Faraday to Einstein: constructing meaning in scientific theories*, Dordrecht: Nijhoff.
——(1986) "A Cognitive-Historical Approach to Meaning in Scientific Theories," in N. Nersessian (ed.) *The Process of Science*, Dordrecht: Nijhoff.
——(1992) "How do Scientists Think? Capturing the dynamics of conceptual change in science," in R. Giere (ed.) *Cognitive Models of Science*, Minneapolis, MN: University of Minnesota Press.

——(2001) "Concept Formation and Commensurability," in P. Hoyningen-Huene and H. Sankey (eds) *Incommensurability and Related Matters*, Dordrecht: Kluwer.
Nersessian, N. and Andersen, H. (1997) "Conceptual Change and Incommensurability: a cognitive-historical view," *Danish Yearbook of Philosophy*, 32: 111–52.
Newton, I. (1952) *Opticks*, New York: Dover.
——(1962) "On the Gravity and Equilibrium of Fluids," in A. Hall and M. Hall (eds and trans) *Unpublished Scientific Papers of Isaac Newton*, Cambridge: Cambridge University Press.
——(1999) *The Principia: mathematical principles of natural philosophy*, trans. I. Cohen and A. Whitman, Berkeley, CA: University of California Press.
Olby, R. (1981) "Heredity and Variation," in W. Bynum, E. Browne, and R. Porter (eds) *Dictionary of the History of Science*, Princeton, NJ: Princeton University Press.
Oliver, K. (1992) "The Matter of Baby M: Surrogacy and The Courts," in H. Holmes (ed.) *Issues in Reproductive Technology I*, New York: Garland.
O'Raifeartaigh, L. (1997) *The Dawning of Gauge Theory*, Princeton, NJ: Princeton University Press.
Overall, C. (1992) "Selective Termination in Pregnancy and Women's Reproductive Autonomy," in H. Holmes (ed.) *Issues in Reproductive Technology I*, New York: Garland.
Pais, A. (1986) *Inward Bound*, New York: Oxford University Press.
Papineau, D. (1985) "Causal Asymmetry," *British Journal for the Philosophy of Science*, 36: 273–289.
Peacocke, C. (1992) *A Study of Concepts*, Cambridge, MA: MIT Press.
Pearl, J. (2001) *Causality*, Cambridge: Cambridge University Press.
Pepperberg, I. (1991) "A Communicative Approach to Animal Cognition: a study of conceptual abilities of an African grey parrot," in C. Ristau (ed.) *Cognitive Ethology: the minds of other animals*, Hillsdale, NJ: Earlbaum.
——(1999) *The Alex Studies: cognitive and communicative abilities of grey parrots*, Cambridge, MA: Harvard University Press.
Perkins, D. (2000) *Introduction to High Energy Physics*, 4th edition, Cambridge: Cambridge University Press.
Perlman, M. (2000) *Conceptual Flux*, Dordrecht: Kluwer.
Pitt, J. (1981) *Pictures, Images and Conceptual Change: an analysis of Wilfrid Sellars' philosophy of science*, Dordrecht: Reidel.
Plato (1961) "Phaedo," trans. H. Tredennick, in E. Hamilton and H. Cairns (eds) *Plato: collected dialogues*, New York: Pantheon.
Polanyi, M. (1958) *Personal Knowledge*, Chicago, IL: University of Chicago Press.
——(1969) "The Potential Theory of Adsorption," in M. Grene (ed.) *Knowing and Being*, Chicago, IL: University of Chicago Press.
Pollock, J. (1986) *Contemporary Theories of Knowledge*, Totowa, NJ: Rowman & Littlefield.
Popper, K. (1992) *The Logic of Scientific Discovery*, New York: Routledge.
Price, H. H. (1940) "The Permanent Significance of Hume's Philosophy," *Philosophy*, 15: 7–37.
——(1964) *Perception*, London: Methuen.
Price, Huw (2003) "Truth as Convenient Friction," *Journal of Philosophy*, 100: 167–90.
Prinz, J. (2002) *Furnishing the Mind*, Cambridge, MA: MIT Press.
Purdy, L. (1992) "Another Look at Contract Pregnancy," in H. Holmes (ed.) *Issues in Reproductive Technology I*, New York: Garland.

Putnam, H. (1962) "The Analytic and the Synthetic," in H. Feigl and G. Maxwell (eds) *Minnesota Studies in the Philosophy of Science* III, Minneapolis, MN: University of Minnesota Press.

——(1975) "The Meaning of 'Meaning'," in *Mind, Language and Reality: philosophical papers*, vol. 2, Cambridge: Cambridge University Press.

——(1978) *Meaning and the Moral Sciences*, London: Routledge and Kegan Paul.

——(1981) *Reason, Truth and History*, Cambridge: Cambridge University Press.

Quine, W. (1953) "Two Dogmas of Empiricism," in *From a Logical Point of View*, New York: Harper Torchbooks.

——(1969) "Natural Kinds," in *Ontological Relativity*, New York: Columbia University Press.

Ramsey, W. (1998) "Prototypes and Conceptual Analysis," in M. DePaul and W. Ramsey (eds) *Rethinking Intuition*, New York: Rowman and Littlefield.

Rescher, N. (1988) *Rationality*, New York: Oxford University Press.

Rey, G. (1999) "Concepts and Stereotypes,"in E. Margolis and S. Laurence (eds) *Concepts: core readings*, Cambridge, MA: MIT Press.

Rigden, J. (2002) *Hydrogen: the essential element*, Cambridge, MA: Harvard University Press.

Ritvo, H. (1997) *The Platypus and the Mermaid and other Figments of the Classifying Imagination*, Cambridge, MA: Harvard University Press.

Rohault, J. (1969) *A System of Natural Philosophy*, trans. J. Clarke, L. Laudan (ed.), New York: Johnson Reprint Corporation.

Rolnick, W. (1994) *The Fundamental Particles and Their Interactions*, Reading, MA: Addison Wesley.

Romer, A. (ed.) (1964) *The Discovery of Radioactivity and Transmutation*, New York: Dover.

——(ed.) (1970) *Radiochemistry and the Discovery of Isotopes*, New York, Dover.

Rosch, E. (1973a) "Natural Categories," *Cognitive Psychology*, 4: 328–50.

——(1973b) "On the Internal Structure of Perceptual and Semantic Categories," in T. Moore (ed.) *Cognitive Development and the Acquisition of Language*, New York: Academic Press.

——(1978) "Principles of Categorization," in E. Rosch and B. Lloyd (eds) *Cognition and Categorization*, Hillsdale, NJ: Earlbaum.

Rosch, E. and Mervis, C. (1975) "Family Resemblance: studies in the internal structure of categories," *Cognitive Psychology*, 7: 573–605.

Rowland, R. (1992) *Living Laboratories*, Sydney: Sun Australia.

Rueger, A. and Sharp, W. (1996) "Simple Theories of a Messy World: truth and explanatory power in nonlinear dynamics," *British Journal for the Philosophy of Science*, 47: 93–112.

Russell, B. (1921) *The Analysis of Mind*, London: George Allen and Unwin.

——(1948) *Human Knowledge: its scope and its limits*, New York: Simon and Schuster.

——(1957) *Mysticism and Logic*, Garden City, NY: Anchor.

——(1959) *The Problems of Philosophy*, New York: Oxford University Press.

——(1960) *Our Knowledge of the External World*, New York: Mentor.

Russow, L. (1980) "Simple Ideas and Resemblance," *Philosophical Quarterly*, 30: 342–50.

Ryle, G. (1949) *The Concept of Mind*, London: Hutchinson.

Sabra, A. (1981) *Theories of Light from Descartes to Newton*, 2nd edition, Cambridge, MA: Harvard University Press.

Salmon, W. (1984) *Scientific Explanation and the Causal Structure of the World*, Princeton, NJ: Princeton University Press.
——(1991) "The Appraisal of Theories: Kuhn Meets Bayes," in A. Fine, M. Forbes, and L. Wessels (eds) *PSA 1990*, vol. 2, East Lansing, MI: The Philosophy of Science Association.
——(1994) "Causality without Counterfactuals," *Philosophy of Science*, 61: 297–312.
Sanford, D. (1995) "Causation," in J. Kim and E. Sosa (eds) *A Companion to Metaphysics*, Oxford: Blackwell.
Sankey, H. (1994) *The Incommensurability Thesis*, Aldershot: Avebury.
——(1997) "Judgement and Rational Theory Choice," in *Rationality, Relativism, and Incommensurability*, Aldershot: Ashgate.
Sartwell, C. (1992) "Why Knowledge is Merely True Belief," *Journal of Philosophy*, 86: 167–80.
Scheffler, I. (1963) *The Anatomy of Inquiry*, Indianapolis, IN: Bobbs-Merrill.
——(1967) *Science and Subjectivity*, Indianapolis, IN: Bobbs-Merrill.
Schweber, S. (1994) *QED and the Men Who Made It*, Princeton, NJ: Princeton University Press.
Scriven, M. (1962) "Explanations, Predictions, and Laws," in H. Feigl and G. Maxwell (eds) *Minnesota Studies in the Philosophy of Science*, vol. III, Minneapolis, MN: University of Minnesota Press.
Sellars, W. (1947a) "Epistemology and the New Way of Words," *Journal of Philosophy*, 44: 645–60.
——(1947b) "Pure Pragmatics and Epistemology," *Philosophy of Science*, 47:181–202.
——(1948a) "Concepts as Involving Laws and Inconceivable Without Them," *Philosophy of Science*, 15: 287–315.
——(1948b) "Realism and the New Way of Words," *Philosophy and Phenomenological Research*, 8: 601–34.
——(1950) "Language, Rules and Behavior," in S. Hook (ed.) *John Dewey Philosopher of Science and Freedom*, New York: Dial.
——(1952) "Obligation and Motivation," in W. Sellars and J. Hospers (eds) *Readings in Ethical Theory*, New York: Appleton-Century Crofts.
——(1953) "Inference and Meaning," *Mind*, 62: 313–38.
——(1958) "Counterfactuals, Dispositions, and the Causal Modalities," in H. Feigl, M. Scriven and G. Maxwell (eds) *Minnesota Studies in the Philosophy of Science*, vol. II, Minneapolis, MN: University of Minnesota Press.
——(1962) "Time and the World Order," in H. Feigl and G. Maxwell (eds) *Minnesota Studies in the Philosophy of Science*, vol. III, Minneapolis, MN: University of Minnesota Press.
——(1963a) "Empiricism and Abstract Entities," in P. Schilpp (ed.) *The Philosophy of Rudolph Carnap*, La Salle, IL: Open Court.
——(1963b) "Imperatives, Intentions, and the Logic of 'Ought'," in H. Casteñeda (ed.) *Morality and the Language of Conduct*, Detroit, MI: Wayne State University Press.
——(1963c) *Science, Perception and Reality*, New York: Humanities Press.
——(1963d) "Theoretical Explanation," in B. Baumrin (ed.) *Philosophy of Science: the Delaware seminar*, vol. 2, New York: Wiley.
——(1964) "Induction as Vindication," *Philosophy of Science*, 31: 197–231.
——(1965) "Scientific Realism or Irenic Instrumentalism," in R. Cohen and M. Wartofsky (eds) *Boston Studies in the Philosophy of Science*, 2, Dordrecht: Reidel.

——(1966) "Thought and Action," in K. Lehrer (ed.) *Freedom and Determinism*, New York: Random House.
——(1967a) "Science and Ethics," in *Philosophical Perspectives*, Springfield, IL: Charles Thomas.
——(1967b) "Some Reflections on Thoughts and Things," *Nous*, 1: 97–121.
——(1968) *Science and Metaphysics*, New York: Humanities Press.
——(1969) "Language as Thought and Communication," *Philosophy and Phenomenological Research*, 29: 506–27.
——(1973) "Conceptual Change," in G. Pearce and M. Maynard (eds.) *Conceptual Change*, Dordrecht: Reidel.
——(1974a) "Meaning as Functional Classification," *Synthese*, 27: 417–37.
——(1974b) "Reply to Marras," in *Essays in Philosophy and its History*, Dordrecht: Reidel.
——(1975) "The Structure of Knowledge," in H. Castañeda (ed.) *Action, Knowledge, and Reality*, Indianapolis, IN: Bobbs-Merrill.
——(1979a) "More on Givenness and Explanatory Coherence," in G. Pappas (ed.) *Justification and Knowledge*, Dordrecht: Reidel.
——(1979b) *Naturalism and Ontology*, Reseda, CA: Ridgeway.
——(1981) "Mental Events," *Philosophical Studies*, 39: 325–45.
——(1982) "Sensa or Sensings: Reflections on the Ontology of Perception," *Philosophical Studies*, 41: 83–111.
Sen, A. (1997) *On Economic Inequality, Expanded Edition*, Oxford: Clarendon Press.
Shapere, D. (1967) "The Philosophical Significance of Newton's Science," *The Texas Quarterly*, 10: 201–15.
——(1982) "The Concept of Observation in Science and Philosophy," *Philosophy of Science*, 49: 485–525.
——(1984) "Reason, Reference, and the Quest for Knowledge," in *Reason and the Search for Knowledge*, Dordrecht: Reidel.
Shea, W. (1991) *The Magic of Numbers and Motion: the scientific career of René Descartes*, Canton, MA: Watson.
Shevory, T. (1992) "Thorough a Glass Darkly: Law, politics, and frozen human embryos," in H. Holmes (ed.) *Issues in Reproductive Technology I*, New York: Garland.
Shogenji, T. (2000) "Self-Dependent Justification Without Circularity," *British Journal for the Philosophy of Science*, 52: 287–98.
Siegel, H. (1989) "Philosophy of Science Naturalized? Some Problems with Giere's Naturalism," *Studies in History and Philosophy of Science*, 20: 365–75.
Singer, P., *et al.* (1993) *Embryo Experimentation: ethical, legal and social issues*, Cambridge: Cambridge University Press.
Slowik, E. (1998) "Cartesianism and the Kinematics of Mechanisms: or, how to find fixed reference frames in a Cartesian space-time," *Nous*, 32: 364–85.
——(1999a) "Descartes' Quantity of Motion: 'new age' holism meets the Cartesian conservation principle," *Pacific Philosophical Quarterly*, 80: 178–202.
——(1999b) "Descartes, Spacetime, and Relational Motion," *Philosophy of Science*, 66: 117–39.
Smith, G. (2001) "The Newtonian Style in Book II of the *Principia*," in J. Buchwald and I. Cohen I. (eds) *Isaac Newton's Natural Philosophy*, Cambridge, MA: MIT Press.
——(2002) "The Methodology of the *Principia*," in I. Cohen and G. Smith (eds) *The Cambridge Companion to Newton*, Cambridge: Cambridge University Press.

Smith, E. and Medin, D. (1981) *Categories and Concepts*, Cambridge, MA: Harvard University Press.
Soddy, F. (1913) "Intra-atomic Charge," *Nature*, 92: 399–400.
——(1932) *The Interpretation of the Atom*, London: John Murray.
Solomon, M. (1994) "Social Empiricism," *Nous*, 28: 325–43.
——(2001) *Social Empiricism*, Cambridge, MA: MIT Press.
Solomon, W. (1977) "Ethical Theory," in C. Delaney *et al*. *The Synoptic Vision: essays on the philosophy of Wilfrid Sellars*, Notre Dame, IN: University of Notre Dame Press.
Sosa, E. (1980) "The Raft and the Pyramid," in P. French, T. Uehling, and H. Wettstein (eds) *Midwest Studies in Philosophy V: studies in epistemology*, Minneapolis, MN: University of Minnesota Press.
Staley, K. (2004) *The Evidence for the Top Quark*, Cambridge: Cambridge University Press.
Stark, H. (1995) *Rationality Without Rules*, unpublished thesis, University of Memphis.
Stein, H. (1991) "From the Phenomena of Motions to the Forces of Nature: hypothesis or deduction," in A. Fine, M. Forbes, and L. Wessels (eds) *PSA 1990*, vol. 2, East Lansing, MI: The Philosophy of Science Association.
——(2002) "Newton's Metaphysics," in I. Cohen and G. Smith (eds) *The Cambridge Companion to Newton*, Cambridge: Cambridge University Press.
Stewart, I. (1992) *The Problems of Mathematics*, 2nd edition, Oxford: Oxford University Press.
——(2001) *Flatterland*, Cambridge, MA: Perseus.
Stroud, B. (1985) "The Significance of Naturalized Epistemology," in H. Kornblith (ed.) *Naturalizing Epistemology*, Cambridge, MA: MIT Press.
Suppe, F. (1989) *The Semantic Conception of Theories and Scientific Realism*, Urbana, IL: University of Illinois Press.
Suppes, P. (1970) *A Probabilistic Theory of Causality*, Amsterdam: North Holland.
Tarski, A. (1983) *Logic, Semantics, Metamathematics*, 2nd edition, trans. J. Woodger, John Corcoran (ed.), Indianapolis, IN: Hackett.
Taylor, E. and Wheeler, J. (1966) *Spacetime Physics*, San Francisco, CA: Freeman.
Taylor, R. (1963) "Causation," *Monist*, 47: 287–313.
——(1966) *Action and Purpose*, Englewood Cliffs, NJ: Prentice-Hall.
Teller, P. (1995) *An Interpretive Introduction to Quantum Field Theory*, Princeton, NJ: Princeton University Press.
——(2001) "Twilight of the Perfect Model," *Erkenntnis*, 55: 393–415.
Thagard, P. (1984) "Frames, Knowledge, and Inference," *Synthese*, 61: 233–59.
——(1988) *Computational Philosophy of Science*, Cambridge, MA: MIT Press.
——(1992) *Conceptual Revolutions*, Princeton, NJ: Princeton University Press.
Tomonaga, S. (1997) *The Story of Spin*, trans. Takeshi Oka, Chicago, IL: University of Chicago Press.
Tooley, M. (1987) *Causation: a realist approach*, Oxford: Clarendon Press.
——(1990) "The Nature of Causation: a singularist account," in D. Copp (ed.) *Canadian Philosophers: celebrating twenty years of the Canadian Journal of Philosophy*, Calgary: University of Calgary Press.
Topper, D. (1990) "Newton on the Number of Colours in the Spectrum," *Studies in History and Philosophy of Science*, 21: 269–79.
Torretti, R. (1999) *The Philosophy of Physics*, Cambridge: Cambridge University Press.

Toulmin, S. (1961) *Foresight and Understanding*, New York: Harper & Row.
Trenn, T. (1977) *The Self-Splitting Atom: the history of the Rutherford-Soddy collaboration*, London: Taylor & Francis.
Tsou, J. (2003) "The Justification of Concepts in Carnap's *Aufbau*," *Philosophy of Science*, 70: 671–89.
Unger, P. (1975) *Ignorance*, Oxford: Clarendon Press.
van den Broek, A. (1913) "Intra-atomic Charge," *Nature*, 92: 372–73.
van Fraassen, B. (1975) "Theories and Counterfactuals," in H. Castañeda (ed) *Action, Knowledge, and Reality*, Indianapolis, IN: Bobbs-Merrill.
——(1980) *The Scientific Image*, Oxford: Clarendon Press.
Veltman, M. (2003) *Facts and Mysteries in Elementary Particle Physics*, River Edge, NJ: World Scientific Publishing Co.
von Wright, G. (1971) *Explanation and Understanding*, Ithaca, NY: Cornell University Press.
——(1993) "On the Logic and Epistemology of the Causal Relation," in E. Sosa and M. Tooley (eds) *Causation*, Oxford: Oxford University Press.
Wang, X. (2002) "Taxonomy, Truth-Value Gaps and Incommensurability: a reconstruction of Kuhn's taxonomic interpretation of incommensurability," *Studies in History and Philosophy of Science*, 33A: 465–85.
Weitz, M. (1988) *Theories of Concepts*, London: Routledge.
Westfall, R. (1971) *Force in Newton's Physics*, New York: American Elsevier.
——(1983) *Never at Rest: A Biography of Isaac Newton*, Cambridge: Cambridge University Press.
Wetzels, W. (1990) "Johann Wihlelm Ritter: Romantic Physics in Germany," in A. Cunnigham and N. Jardine (eds) *Romanticism and the Sciences*, Cambridge: Cambridge University Press.
Wilczek, F. (1999) "The Persistence of Ether," *Physics Today*, 52, no. 1: 11–13.
Williams, W. (1992) "Is Hume's Missing Shade of Blue a Red Herring?" *Synthese*, 92: 83–99.
Winch, P. (1958) *The Idea of a Social Science*, London: Routledge and Kegan Paul.
Winkler, K. (1989) *Berkeley: An interpretation*, Oxford: Clarendon Press.
Wittgenstein, L. (1953) *Philosophical Investigations*, trans. G. Anscombe, New York: Macmillan.
Wright, C. (1999) "Truth: A Traditional Debate Reviewed," in S. Blackburn and K. Simmons (eds) *Truth*, Oxford: Oxford University Press.
Wright, E. (1977) "Perception: A New Theory," *American Philosophy Quarterly*, 14: 273–86.
——(1985) "A Defence of Sellars," *Philosophy and Phenomenological Research*, 46: 73–90.
——(1993) *New Representationalisms*, Brookfield, VT: Ashgate.
Zemach, E. (1976) "Putnam's Theory on the Reference of Substance Terms," *Journal of Philosophy*, 73: 116–27.

# Index