EXPLANATIONS IN *K*

An Analysis of Explanation as a Belief Revision Operation

by

Andrés Páez

A dissertation submitted to the Graduate Faculty in Philosophy in partial fulfillment of the requirements for the degree of Doctor in Philosophy.

The City University of New York

2002

This manuscript has been read and accepted for the Graduate Faculty in Philosophy in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy

1.28.02
Date

Chair of Examining Committee

1/28/02
Date

Executive Officer

Supervisory Committee:

Arnold Koslow (Advisor)

Jonathan Adler

Alberto Cordero

Michael Levin

Rohit Parikh

THE CITY UNIVERSITY OF NEW YORK

ABSTRACT

EXPLANATIONS IN *K*
An Analysis of Explanation as a Belief Revision Operation

by

Andrés Páez

Advisor: Professor Arnold Koslow

Explanation and understanding are intimately connected notions, but the nature of that connection has generally not been considered a topic worthy of serious philosophical investigation. Most authors have avoided making reference to the notion of understanding in their accounts of explanation because they fear that any mention of the epistemic states of the individuals involved compromises the objectivity of explanation. Understanding is a pragmatic notion, they argue, and pragmatics should be kept at a safe distance from the universal features of explanation. My main contention in this dissertation is that there is a sense in which it is meaningful and useful to talk about *objective* understanding, and that to characterize this notion it is necessary to formulate an account of objective explanation that makes reference to the beliefs and epistemic goals of the participants in a cognitive enterprise.

My account of explanation is based on the belief-doubt model of inquiry first proposed by Peirce. Using the tools provided by decision theory, and the models of belief revision developed by Isaac Levi and by Alchourrón, Gärdenfors, and Makinson, I analyze the objective and pragmatic conditions that a piece of information must fulfill to be both explanatory and epistemically useful to an inquiring agent. The objective basis for an explanation is provided by the relation of probabilistic relevance that the fact described by the explanans bares to the fact described by the explanandum. But an explanation is much more than its foundation. As an epistemological notion, a potential explanation can only be a bona fide explanation if it becomes part of inquiry, that is, if an agent or a group of agents can see any value in it for their cognitive purposes. I provide a way to evaluate an explanation in terms of its credibility, its content, and its explanatory value. I argue that although an objective measure of the credibility and content of an explanation can be established, in many cases a complete determination of its explanatory value will ultimately depend on the interests and goals of the individual inquirers.

To my parents,
Gonzalo and Stella,
who have never quite understood what I do,
but have supported me regardless.

PREFACE

In the early spring of 2001 I attended a talk by Isaac Levi in which he defended his well-known pragmatist thesis that truth ought to be the immediate aim of inquiry. During the talk, a professor sitting behind me whispered to one of his colleagues, loudly enough so that the people sitting in the vicinity could hear, "Isaac has become an intellectual terrorist!" At the time, I had been trying to come up with a way to integrate the strong intuition that pragmatics ought to be an essential part of an account of explanation with the competing intuition that explanations must be objective. The approach to inquiry that Levi defended during that talk, and that ultimately stems from Peirce's belief-doubt model, provided a way to integrate the two intuitions, and I decided to follow that path at the risk of being deemed an intellectual terrorist by some of my peers. The result is the account of explanation provided within these pages.

Many people have contributed to the completion of this project, but no one deserves more credit than Arnold Koslow, my advisor and philosophical mentor throughout my years in graduate school. Although Arnie does not agree with the philosophical route that I have decided to travel, his support and encouragement in dealing with philosophical and nonphilosophical problems has been fundamental. Michael Levin's suggestions and questions have helped me see my own work from

perspectives that I had not previously considered. Isaac Levi kindly answered my questions about his work, even when they were not the most insightful ones. Jonathan Adler, Alberto Cordero, and Rohit Parikh read the final draft and offered helpful suggestions. Samir Chopra provided me with a map to navigate the extensive literature on belief revision, and Michael Devitt helped me deal with all the practical obstacles that stood in the way of my dissertation defense.

David Rosenthal was not directly involved in this project, but the many conversations we have had over the years have helped me be a better philosopher. My friend Roblin Meeks has been my travel companion through the complicated maze that is graduate school. Our afternoons in Bryant Park discussing Kemp Smith's interpretation of Hume or Kant are among my most memorable moments as a graduate student. I profited enormously from the animated philosophical discussions with my friends Jared Blank, Jennifer Fisher, Eric Hetherington, Russell Marcus, and Mark McEvoy, the members of what Roblin and I quite pedantically called "The New York Circle."

My dear friend Andrea Knutson provided me with her love and companionship during most of my life in graduate school. Her kindness and sense of wonder were a constant source of inspiration.

I must thank not only those who helped me write my dissertation, but also those who were responsible for taking my mind away from the books: my friends Marcelo Bucheli and Carlos A. González.

The financial support of Colfuturo and Colciencias, two Colombian agencies that invest in the education of Colombian citizens abroad, enabled me to come to the Graduate Center and to have the peace of mind to pursue my studies without having to worry too much about my financial situation. During my final year, I was fortunate to have been awarded the Helain Newstead Dissertation Year Fellowship in the Humanities, which allowed me to give up my teaching responsibilities and finish writing my dissertation.

Gratitude of a kind that does not fit easily into a Preface is due to the woman whose love, friendship, and unconditional support inspired me to complete this work—*meine liebe Lebensgefährtin* Friederike Fleischer. I owe her more than I can possibly say.

TABLE OF CONTENTS

INTRODUCTION

Explanation and understanding are intimately connected notions, but with the exception of von Wright's (1971) and Friedman's (1974) writings on the subject, the nature of that connection has generally not been considered a topic worthy of serious philosophical investigation. Most authors have avoided making reference to the notion of understanding in their accounts of explanation because they fear that any mention of the epistemic states of the individuals involved compromises the objectivity of explanation. Understanding is a pragmatic notion, they argue, and although a subject worthy of curiosity, pragmatics should be kept at a safe distance from the universal features of explanation. My main contention in this dissertation is that there is a sense in which it is meaningful and useful to talk about *objective* understanding, and that to characterize this notion it is necessary to formulate an account of objective explanation that makes reference to the beliefs and epistemic goals of the participants in a cognitive enterprise.

Explanation is often said to be an interest-relative notion. Different agents impose different demands on the information that they regard as explanatorily valuable. Van Fraassen (1980) takes this to mean that there is no fundamental difference between descriptive information and explanatory information. Explanatory information is just descriptive information that, in a given context, answers a particular why-question of interest to the inquirer. I find this view unacceptable. To

deny that information must fulfill certain fixed requirements in order to have the potential to be explanatorily valuable is to transform explanation into a psychological notion. Social psychologists often refer to explanations as the "cognitive relief" that an agent seeks when confronted with a "cognitive dissonance" (Festinger, 1957). An explanation in this sense is a subjective mechanism of rationalization. I will argue that van Fraassen's pragmatic theory of explanation lacks the sort of objective grounding that separates the psychological from the epistemological notion of explanation.

In my view, the interest-relativity of explanation has a much deeper origin. It derives from the interest-relativity of inquiry in general. Different inquiring agents use information for different purposes, and their acceptance of new information is directed by their cognitive interests and goals. Far from being a superficial characteristic of the search for knowledge, I believe that this is a fundamental trait of the acquisition of knowledge in general. The cost and effort that goes into obtaining new information makes the beliefs that an inquiring agent has accepted a valuable asset that must be treated with care. Gratuitous losses must be prevented, and the agent's acceptance of new information always involves the risk of bringing error into his system of beliefs. The risk must be compensated by an epistemic incentive that outweighs the cost. And one of the biggest incentives of all is the obtainment of understanding of the matter at hand. If an explanation fulfills no purpose in the eyes of an agent, he will be more reluctant to incur the risks involved in

accepting it. And if the information explains too much, it might be too good to be true. The acceptance of new information requires a delicate balance between two conflicting cognitive goals: the obtainment of valuable information and the avoidance of error.

To say that the acquisition of valuable information is a desideratum of inquiry is, I hope, an uncontroversial claim. But the assertion that the avoidance of error is a desideratum of inquiry is based on the controversial assumption that an inquiring agent is always in a position to judge what is true and what is false. This assumption has been rejected by many prominent thinkers. The truth about extralogical matters cannot be had, they say. We might approach the truth with better theories and in that sense truth is indeed the *telos* of inquiry, but we cannot be concerned with truth as an immediate goal because any general empirical claim will always be open to doubt.

Following the steps of Peirce, Dewey, Levi, and others, I will adopt the belief-doubt model of inquiry in providing my account of explanation. According to the belief-doubt model, an inquiring agent presupposes that everything he is currently committed to fully believing is true. This does *not* mean that truth or falsity is relative to what the agent believes. But the agent's *judgments* of what is true and what is false are relative to what he currently believes. If the agent is concerned with the acquisition of new error-free information, his assessment of the risk of

error incurred and of the epistemic value obtained in accepting a piece of information can only be made relative to the judgments of truth available to him.

To claim that an inquiring agent presupposes that everything he is currently committed to fully believing is true is not to say that he cannot change his mind. Certainty or full belief does not entail incorrigibility. Levi explains the claim thus: "To regard some proposition as certainly true and as settled is to rule out its falsity as a serious possibility for the time being. ... But from this it does not follow that good reasons will not become available in the future for a change of mind and for calling into question what is currently considered to be true" (1991, p. 3). Peirce puts it more graphically: "The scientific spirit requires a man to be at all times ready to dump his whole cartload of beliefs, the moment experience is against them" (1932, pp. 46-47).

The claim that certainty or full belief does not entail incorrigibility is a rejection of a central tenant of both foundationalism and skepticism. But the alternative is not an unbridled relativism. Levi, for example, tries to establish a delicate connection between pragmatism and realism by contrasting "secular" with "messianic" realism. The hallmark of the latter is an "inexplicable albeit touching faith" in convergence to the truth "at the End of Days." Levi sides with a full-blooded secular realism and its "myopic" concern with seeking truth and avoiding error in the very next step of inquiry (1991, p. 163).

The main reason why I have adopted the belief-doubt model is that an account of explanation that takes into consideration the epistemic value of the information that we acquire through inquiry leads to a natural resolution of the conflict between the purely pragmatic approach to explanation defended by van Fraassen, for example, and the more common approach in which pragmatic considerations are not assigned any serious role. By taking into account the shared commitments and the cognitive interests and goals of the individuals engaged in a cognitive enterprise, we obtain a notion of explanation that is objective by any reasonable standard of objectivity, and that clarifies the connection between explanation and understanding.

Some might think that adopting the belief-doubt model is too high a price to pay for a satisfactory account of explanation. I will not attempt a defense of the model here, among other things because I have little to add to what others have said in its favor. But any virtue of the account of explanation presented here will ultimately stem from my adoption of the belief-doubt approach to inquiry. In that sense, the entire dissertation is an argument for the belief-doubt model.

The rest of this introduction presents the general structure of the dissertation. The work is divided into four chapters as follows.

The purpose of Chapter 1 is to defend the claim that neither a purely pragmatic nor a purely nonpragmatic analysis of explanation is correct. The former approach is adopted by Achinstein and van Fraassen, and it focuses almost exclu-

sively on the illocutionary aspects of explanatory acts. The latter is the more traditional view that eschews all pragmatic considerations and concentrates on the syntactic and semantic conditions that must be imposed on the content of explanatory acts. My contention is that explanation is *essentially* a mixed notion. Any attempt to transform an explanation into a logical relation between statements or into a linguistic transaction that fulfills certain contextual requirements will only offer a distorted picture.

In the last section of the chapter, I introduce a new way of understanding the pragmatics of explanation. As I said in the beginning, the rejection of pragmatics in the analysis of explanation stems from the fear that that the objectivity of explanation will be jeopardized by any reference to the belief systems and the cognitive interests of different individuals. Using the tools provided by the theories of belief revision developed during the last two decades, I will argue that it is possible to characterize, in precise terms, a notion of explanation that is both objective and pragmatic, that does not depend on the idiosyncrasies of the individuals involved but that takes their epistemic commitments and goals into consideration. The rest of the dissertation is devoted to that project.

Chapter 2 is a basic introduction to the relevant aspects of the logic of belief revision. I begin by discussing the three types of belief revision operations defined in the pioneering work of Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson (1985). The AGM model, as it is known, provides a simple way of un-

derstanding the changes that an inquiring agent makes to her evolving doctrine. But the AGM model is an incomplete structure. It is necessary to fill in the details of the decision process involved in the acceptance or the rejection of a piece of information. Levi's account of belief revision, which was developed independently of the work of AGM, provides the missing pieces. Although the relevance of some of the notions that I discuss in Chapter 2 will not be immediately obvious, virtually every single notion discussed there is used in one way or another in the remaining chapters of the dissertation.

The third chapter is a first approximation to the problems involved in characterizing explanation as a belief revision operation. I begin the chapter by offering an analysis of the different epistemic contexts in which an inquiring agent might consider the adoption of an explanation. This analysis will allow us to see the complexities involved in characterizing the changes that the agent must make to his epistemic state in order to understand a given phenomenon.

There has only been one philosophically informed attempt to characterize explanation as a belief revision operation—Gärdenfors's account in *Knowledge in Flux* (1988). Other accounts provided by researchers in artificial intelligence and computer science use an idiosyncratic notion of explanation that bears little resemblance to the philosophical notion. I analyze the best-known account in the AI literature, the one offered by Maurice Pagnucco (1996), to illustrate the differences.

The chapter concludes with a brief, informal presentation of the account of explanation that I develop in more detail in the following chapter.

Chapter 4 contains a detailed presentation of my account. Unlike prior pragmatic accounts of explanation, I believe that we should be able to impose objective conditions on the information that has the potential to explain a phenomenon. The main requirement will be that a potential explanation must exhibit the relation of probabilistic relevance that the state of affairs described by the explanans sentence bears to the fact described by the explanandum sentence.[1] Statistical relevance has been used before in the study of explanation, but unlike other authors, I take the discovery of relevance relations as an end in itself, and not as a means to determine the exact value of the probability that the explanans confers upon the explanandum.

An explanation is much more than its foundation. As an epistemological notion, a potential explanation can only be a bona fide explanation if it becomes part of inquiry, that is, if an agent or a group of agents can see any value in it for their cognitive purposes. The second section of the chapter indicates how an explanation can be evaluated both in terms of its credibility, its content, and its explanatory value. I will argue that although an objective measure of the credibility and content of an explanation can be established, in most cases a complete deter-

1. Typically, the explanation of a singular fact will require the inclusion of several positively and negatively relevant factors, and thus of several probability sentences.

mination of its epistemic value will ultimately depend on the interests and goals of the individual inquirers. With this analysis in hand, it will be possible to offer a definition of an explanation of a given fact relative to K, the corpus of beliefs of an individual inquirer, and of the type of objective understanding thereby obtained.

CHAPTER 1

TWO APPROACHES TO THE PRAGMATICS OF EXPLANATION

In the opening chapter of *Logical Foundations of Probability* (1950), Carnap went to great lengths to convince us of the importance of having a prior informal grasp of any concept for which we want to offer a philosophical analysis. If we do not possess a sound informal understanding of the concept we want to explicate, Carnap argued, any attempt to find a precise philosophical explication will egregiously miss the mark. Wesley Salmon (1984) has justly complained that most attempts to find an adequate characterization of explanation are guilty of this original sin. As the discussion in this chapter will make evident, many of the disagreements in the philosophy of explanation are the result of taking different notions of explanation as starting points for the analysis, and many of the weaknesses in the resulting theories a consequence of using superficial characteristics of explanation as sufficient to ground an analysis of this notion.

The argumentative methods followed by many philosophers are partly to blame for this situation. Many writers have the tendency to base their analyses of explanation on the intuitions elicited by certain compelling examples. The result is usually a logical model that has been carefully crafted to account for the examples, but that has nothing illuminating to say about why the examples are so compelling in the first place. What is missing in most cases is a prior understanding of the role

of explanation within our cognitive practices. It is pointless, in my view, to try to say in precise terms what an explanation is if we do not have a clear idea of what explanations are supposed to accomplish in the course of inquiry. Our first task will thus be to offer a "clarification of the explicandum," to use Carnap's terminology, that brings out the epistemological aspects of explanation.

In this chapter I examine the notions of explanation that different authors have regarded as the legitimate explicandum of a theory of explanation. In particular, I want to focus on the two endpoints of the spectrum of possible approaches to the pragmatics of explanation. On one end we find the purely pragmatic approach adopted by Achinstein and van Fraassen, which focuses almost exclusively on the illocutionary aspects of explanatory acts. On the other end we find the more traditional view that eschews all pragmatic considerations and concentrates on the syntactic and semantic conditions that must be imposed on the content of explanatory acts. Instead of arguing, *à la* Carnap, for the existence of two different concepts of explanation, I will show that each approach, individually considered, is inevitably incomplete. Both approaches have missed a crucial epistemological element that unifies them and at the same time preserves the insights that make each of them compelling individually. The missing pieces of the puzzle will be provided at the end of this chapter, and they will become the basis for the account of explanation that I develop in Chapter 4.

1. Scientific and Everyday Explanations

I begin with a question that every approach to explanation ought to confront at the outset: Is there any philosophically significant difference between scientific explanations and the ordinary explanations offered in the course of everyday life? Since most philosophical discussions of explanation are presented under the heading "scientific explanation," we must consider whether there is a special concept of scientific explanation, discontinuous from its everyday counterpart. If there is one, we will end up with two different explicanda after all. If there is no special concept, we have to explain how the epithet 'scientific' is to be construed as a qualification of 'explanation'.

1.1 *The Institutional Theory*

There is an undeniable *prima facie* distinction between ordinary and scientific explanations, if only because some explanations are given in the course of everyday life and others are given by scientists in the course of their work. Some philosophers seem to believe that this fact alone can be the basis for establishing a genuine distinction between two types of explanation. By assuming that "scientific explanation" *means* "an explanation actually offered in science," they find a fast and ready criterion to determine which explanations count as scientific and which do

not.[1] In contemporary aesthetics, where one encounters the analogous problem of establishing what makes an everyday object a work of art, there is a similar approach: Danto's institutional theory of art. In Danto's theory, the question 'What makes something a work of art?' is answered by saying that a work of art is whatever the artworld determines to be art. The artworld, in turn, is defined as "an atmosphere of artistic theory, a knowledge of the history of art" (1981, p. 135). Taking this cue from aesthetics, I will call the approach we are presently discussing "the institutional theory of scientific explanation." The institutional theory holds that whether an explanation is scientific or not is determined by the science-world, which is defined as an atmosphere of scientific theories and institutions informed by knowledge of the history of science. The theory rests on the questionable assumption that there is a unified science-world, and that it has clearly marked boundaries. The success of the theory will thus depend, in part, on the possibility of clearly identifying the practices that have historically counted as scientific.

Most philosophers who write about scientific explanation have rejected the institutional approach. The theory contradicts a widely held belief about explanation, namely, that explanations must be factive, i.e., that both the explanans and the

---

1. This view is explicitly adopted by Van Fraassen: "To call an explanation scientific, is to say nothing about its form or the sort of information adduced, but only that the explanation draws on science to get this information (at least to some extent) and, more importantly, that the criteria of evaluation of how good an explanation it is, are being applied using a scientific theory" (1980, pp. 155-156).

explanandum must be true.[2] If one denies the factivity of explanation, they argue, one cannot avoid the conclusion that the Ptolemaic theory, the phlogiston theory, or the caloric theory, provided bona fide scientific explanations. According to its critics, the institutional theory transforms the explication of the concept into a question of "intellectual fashion" (Friedman, 1974, p. 11), and one of its consequences is that it reduces theory choice to "a matter for mob psychology" (Lakatos, 1978, p. 90). Underlying these criticisms is the idea that the boundaries of science ought not to be the consequence of historical trends. Instead, it ought to be possible to find rational, unanimous criteria for the acceptance of scientific theories, and a unique set of necessary and sufficient conditions for scientific explanation.

The main problem with these demands is that they do not confront the fact that virtually every theory used in science has eventually turned out to be false, or at least only approximately true. Thus we face the following dilemma: we must either conclude that there has never been a genuine scientific explanation, or deny that explanations should meet the factivity condition. The latter seems to be van Fraassen's position:

> There are many examples, taken from actual usage, which show that truth is not presupposed by the assertion that a theory explains something. ... Newton's theory explained the tides, Huygen's theory ex-

---

2. More precisely, the explanans-statements and the explanandum-statement must be true. If one holds, following Lewis (1986) and Woodward (1984a, 1986), that the relata of the explanation relation are particulars, i.e., things or events, the claim amounts to saying that the things or events occurring in both the explanans and the explanandum position exist or occur.

> plained the diffraction of light, Rutherford's theory of the atom explained the scattering of alpha particles... Hence, to say that a theory explains some fact or other, is to assert a relationship between this theory and that fact, which is independent of the question whether the real world, as a whole, fits that theory (1980, p. 98).[3]

At the heart of the conflict regarding the factivity condition lies the question whether the philosophy of science ought to be normative or purely descriptive. Should we take scientific practice at face value, as the institutional theory suggests, or should we tell scientists what conditions must be imposed on their explanations to make them truly scientific? My intention here is not to answer this question, but rather to show that in either case the distinction between scientific and ordinary explanations is philosophically irrelevant. I will begin with the descriptive approach.

The institutional theory offers a taxonomy of explanation, but in order to be an adequate taxonomy, it must provide the criteria used to classify different types of explanation. The only criterion that the theory provides to determine which explanations are entitled to be called 'scientific' is a sociological one—the entitlement derives solely from the institutional status of the persons from whom the

---

3. We can add to this analysis what is usually known as the "pessimistic induction," an argument to the effect that the only reasonable conclusion we can have about the present claims of science is that they are false. The argument is controversial and its plausibility depends on rejecting the standard conception of theories and invoking some version of the Duhem-Quine thesis. I mention it here simply to indicate that any argument in support of the pessimistic induction will only reinforce the aforementioned dilemma.

explanations originate. The question is whether this criterion has genuine metaphysical or epistemological consequences. To answer this question affirmatively we must find a set of unique, metaphysically or epistemologically significant properties that all the institutional practices that have served as a basis for the distinction have in common, and that all other practices lack. But the properties that characterize scientific practices and set the boundaries of science change from time to time, and even from discipline to discipline. Even if we single out some very weak criteria to identify the whole range of practices that have been historically regarded as scientific, I suspect that some very questionable candidates will sneak in, and some obvious candidates will be left out. This is not to say that there are no demarcation criteria for science. The demarcation problem, as I understand it, is a normative one. I am simply denying that there is a nonempty set of philosophically significant criteria that have *historically* determined the boundaries of science. In consequence, it will be impossible to identify one philosophically significant difference between scientific and nonscientific explanations if the distinction in question is drawn in purely sociological terms.

1.2 *Are There Scientific Explananda?*

Some philosophers have claimed that a philosophically significant difference between actual scientific explanations and their everyday counterparts lies in the type of explanandum that each of them is concerned with. It is remarkable that many

theorists of explanation, and especially of causal explanation, seldom use examples from real scientific contexts in the exposition of their accounts. A brief glance at the literature on explanation reveals a macabre landscape filled with horrible accidents and terrible tragedies: car crashes, buildings on fire, smokers dying of lung cancer, climbers falling to their deaths, and so on.[4] The reason why everyday explanations of particular events are preferred as examples is twofold. In the first place, it is generally acknowledged that it is easier to provide an account of the explanation of particulars than it is to provide one of regularities or laws. The reason lies in the recalcitrant difficulties associated with the notion of lawlikeness. Secondly, it is claimed that the explanation of particular occurrences is something "comparatively rare" in scientific textbooks and journals—"found only perhaps in geology and astronomy" (Friedman, 1974, p. 5). Scientists are generally not interested in a phenomenon as a particular, but rather as an instance of a type. They want to know why a certain phenomenon happens in general, not why it happened on this or that particular occasion.

It is tempting to conclude from this that the difference between scientific and everyday explanations lies in the type of explanandum involved in each case: scientists explain regularities, the rest of us explain why this or that particular thing or event happened. But such a criterion would be clearly inadequate. To begin

---

4. These examples can be found in Lewis (1986a), Woodward (1984a), Hitchcock (1995), and Mellor (1995).

with, there are many clear instances of scientific explanations of nonrecurring phenomena—and not just in fields such as astronomy and geology, as Friedman claims. Likewise, the everyday explanation of regularities is the basis of the folk theories with which we make sense of our more prosaic endeavors. Finally, there is no clear reason why the obvious differences between the explanation of a particular and that of a regularity should prevent us from establishing a more general account in which both types of explanation turn out to be different sides of the same coin. Although the account of explanation that I provide in Chapter 4 is limited to the explanation of singular facts, it can be adapted as an account of the explanation of laws. But that will have to wait until we have a better idea of what a law of nature is.

1.3 *What Are Hempel's Models of Scientific Explanation About?*

Although most philosophers do not define scientific explanation in terms of current scientific practices, they obviously think that the models proposed as explications of explanation bear some relation to the explanations provided in the empirical sciences. It is not an easy task, however, to establish precisely what that relation is. Consider Hempel's exposition of the purpose of his models of explanation:

> [T]hese models are not meant to describe how working scientists actually formulate their explanatory accounts. Their purpose is rather to

> indicate in reasonably precise terms the logical structure and the rationale of various ways in which empirical science answers explanation-seeking why-questions. The construction of our models therefore involves some measure of abstraction and of logical schematization (1965, p. 412).

Despite Hempel's reference to empirical science in this passage, at the end of his discussion of the general character and intent of his models of explanation he states that the models are also intended to capture the logic of everyday explanations: "I think that all adequate scientific explanations and their everyday counterparts claim or presuppose at least implicitly the deductive or inductive subsumability of whatever is to be explained under general laws or theoretical principles" (p. 424-5). But if that is the case, why did Hempel bother to make a distinction between two types of explanation? Why did he not offer an account of explanation *simpliciter*?

The issue is complicated by the fact that Hempel claims that the explanations formulated both in science and in everyday contexts diverge more or less markedly from the idealized and schematized covering-law models *in exactly the same ways*. In other words, Hempel does not claim that the explanations actually offered in science are generally closer to his idealized models, or that they are less prone to fall short of the standards incorporated into the models. But if the models are not intended exclusively as an idealization of the explanations offered in science or by scientists, and if pragmatic considerations affect the explanations of-

fered in science and in everyday life in exactly the same way, what does Hempel mean by 'scientific explanation'?

The most likely answer is that Hempel wanted to establish a distinction between 'explanation' understood as a pragmatic term that requires reference to the persons involved in the process of explaining, and 'explanation' understood as an inferential relation between true sentences that does not need to be relativized to an individual. This distinction is of central importance in the study of explanation, as we will see throughout this dissertation. However, as Hempel has already mentioned, the inference relation between sentences described in his models can be instantiated in any area of human activity, not only in scientific research. Thus to equate the distinction between the pragmatic and the nonpragmatic sense of 'explanation' to the distinction between everyday and scientific explanation would rest on a purely idiosyncratic understanding of the term 'scientific'.

David-Hillel Ruben argues that "the term 'scientific explanation' is meant to conjure up the fact that there is a goal or ideal of precision and completeness, explicated by Hempel's models, which explanations in science can aspire to and can actually meet if so required" (1990, p. 17). If that was Hempel's purpose, then he should have used the epithet 'scientific' in the same way that Carnap used it in his analysis of the concept of probability, namely, to qualify the new concept of probability that would replace the inexact, prescientific one. Hempel's models provide a scientific concept of explanation in the sense that they offer the exact-

ness, fruitfulness, simplicity, and similarity to the explicandum that Carnap required of an adequate explicatum. Thus Hempel's models are better understood, not as models of *scientific explanation*, but rather as *scientific models* of explanation. This is, I believe, the only relevant meaning of the term 'scientific' in Hempel's theory of explanation.

1.4 *Normativity and Scientific Explanation*

We must now consider the relevance of the distinction between scientific and non-scientific explanations for a normative account of scientific explanation.[5] By a normative account I mean one that provides a set of necessary and sufficient conditions that determine whether an explanans correctly explains an explanandum, regardless of whether any actual explanation satisfies these conditions. Since the proponent of a normative account is not bound by our current scientific practices, the only restriction she encounters in providing a set of necessary and sufficient conditions originates in her desire to honor some methodological rule or some epistemological or metaphysical principle.

The only way in which a distinction between scientific and nonscientific explanations can be relevant in a normative account is if any of the rules or

---

5. I devoted a separate section to Hempel's theory because it cannot be considered a purely normative theory. In several passages Hempel claims or implies that he wants to do justice to generally agreed instances of successful explanations, and his analysis is careful to fulfill the first of Carnap's requirements for an adequate explicatum, namely, that it resembles the explicandum.

principles that the philosopher wants to honor is such that it sanctions the imposition of a condition only on certain types of explanation, but not on others, and the epithet 'scientific' is used to qualify one of the resulting types of explanation. Most methodological or philosophical principles will call for the same necessary and sufficient conditions to be imposed on all explanations. Presumably, it is desirable that *all* explanations be objective, accurate, refutable, simple, fruitful, and so on. There are certain cases, however, in which a distinction between types of explanations may be called for: causal and noncausal, deductive and inductive, potential and actual, and so on. The problem is that in a purely normative approach there is no nonarbitrary way in which the word 'scientific' can be used to qualify one of the resulting types of explanation. In Popper's words, "what is to be called a 'science' and who is to be called a 'scientist' must always remain a matter of convention or decision" (1959, p. 52). It would be a mistake for someone proposing a purely normative account to say, for example, that only causal explanations can be called scientific because these are the only explanations currently or commonly accepted in science. If the philosopher cannot appeal to the actual use of the term 'scientific', his own use is therefore absolutely idiosyncratic. He can *stipulate*, for example, that only causal explanations should be considered scientific, but in that case the word 'scientific' would be used simply as a commendatory or honorific title. The epithet would add nothing to our understanding of that type of explanation.

1.5 *Salmon on Scientific Understanding*

A further attempt to establish a distinction between scientific and nonscientific explanations is based on the intuition that the kind of understanding of the world that scientific explanations provide is of a different nature than the understanding provided by other types of explanation. In order to examine the plausibility of this distinction, we must consider what is meant by the term 'understanding'. In a recent book, Salmon offers a thorough analysis of the notion of understanding. In his view, the notion corresponds to "a cluster of concepts" (1998, p. 8), and he offers the following classification of the different *types* of understanding:

(i) *symbolic understanding*: Its objects are the *meanings* of linguistic and nonlinguistic symbols—natural and artificial languages, rules and instructions, concepts, works of art, pictographs, religious icons, rituals and ceremonies, etc.

(ii) *empathic understanding*: The kind of understanding involved in sharing the feelings and emotions of others, a notion better captured by the word 'empathy'. It is based on emotive factors, on feelings and values.

(iii) *goal-oriented understanding*: The type of understanding obtained by invoking purposes, reasons, aims or goals. Salmon divides this type of understanding into two sub-types. We can understand the deliberate actions of human beings by means of the motives, purposes, and reasons behind them. This is the kind of un-

derstanding provided by teleological explanations. "This kind of understanding tends to blend into empathic understanding, for knowledge of the desires and values of others enables us to know, if not share, their feelings" (pp. 8-9). On the other hand, we can understand a given phenomenon by invoking its *function*. "We understand why our blood contains hemoglobin: its function is to transport oxygen from the lungs to other parts of the body" (p. 9). This is the kind of understanding provided by functional explanations. It is essential, Salmon argues, to separate function from conscious purpose in order to avoid the invocation of supernatural purposes in the explanation of nature.

(iv) *scientific understanding*: "The fourth major type of understanding is linked to scientific explanations in the physical, biological, behavioral, and social sciences" (p. 9). Salmon divides this type of understanding into two sub-types. "We can say that we have *scientific understanding* of phenomena when we can fit them into the general scheme of things, that is, into the *scientific world-picture*." The second sub-type of scientific understanding "is especially prominent in the curiosity of children. We want to know *how things work* and, it should be added, *what they are made of*. This may be characterized as causal-mechanical understanding" (p. 87, emphasis kept). The two sub-types of scientific understanding "complement rather than conflict with each other" (p. 90).

This classification is supposed to identify different *types* of understanding. But this claim is extremely problematic. In the first place, to offer a classification of the types of things that can be understood is not to offer a classification of different concepts of understanding. In each case, Salmon is simply singling out a feature in virtue of which a certain use of the term 'understanding' differs from other uses of the same term. But the fact that a concept has different uses does not mean that each usage defines or determines a new concept. By attempting to fragment 'understanding' and accepting a cluster of special concepts we will only succeed in making them all unintelligible together.

Secondly, and more importantly, Salmon's classification is extremely superficial and can be challenged in a number of ways. Let us begin with the second type of understanding: empathic understanding. To use the term "empathy" in this context is misleading. In order to *feel* sympathy or empathy one has to *understand* the predicament of another human being, but the understanding and the feeling are two different things, regardless of how the term 'understanding' is actually used. For example, I can tell my neighbor, "I understand what you're going through" because I know that he needs my support, or because I am really in grief because of his personal tragedy. But even though my feelings are different in each case, in both cases I understand the same thing, namely, that a personal tragedy is causing him a great amount of grief. Now such understanding comes from recognizing the connection between a given event and the emotions and feelings it generates, a

connection that can only be established in the light of our knowledge of the desires and values of human beings within a culture. That knowledge can only be acquired through experience, when we learn—to paraphrase Salmon—how to fit these connections into the general scheme of things, that is, into our cultural and psychological world-picture. In consequence, understanding emotions and feelings does not seem to differ in that respect from Salmon's characterization of scientific understanding. In both cases we are looking for, in Toulmin's words, "rational patterns of connections in terms of which we can make sense of the flux of events" (1961, p. 99).

The same conclusion will be reached if we look at the third type of understanding postulated by Salmon: goal-oriented understanding. As we have seen, Salmon divides this kind of understanding into teleological and functional understanding. Conscious purposes and goals help us understand purposive human behavior, but the same teleological explanatory principles cannot be extended to natural phenomena. The consequence would be the anthropomorphization of nature—Aristotle's claims that nature *abhors* a vacuum and that bodies *seek* their natural place are often cited as examples. On the other hand, natural, cultural, and social functions, it is claimed, provide a more objective type of understanding of natural and cultural phenomena. Functional analysis is used to explain some recurrent activity or some behavior pattern in an individual or a group, such as a phobia, a neurosis, a cultural pattern, or a social structure or norm. The main goal is to re-

veal the contribution that the pattern makes to the preservation or the development of the individual or the group to which he or she belongs. But independently of the differences in their field of application or their scientific status, both teleology and functional analysis have something in common. Any teleological or functional claim is based on our ability to identify the connection between a phenomenon and its goal or purpose, a connection that, once again, can only be established in the light of the laws and regularities that form our theoretical world-picture.

Finally, it would seem that the only distinction that might yield an independent type of concept of understanding is what Salmon calls "symbolic understanding." But here, again, we find the same situation. Suppose an art historian is studying the symbolic meaning of certain objects present in the paintings of a little-known Renaissance artist from a small Tuscan town. The meaning of such objects cannot be understood in a vacuum. It is only when placed within a background of cultural and social theory and history that the objects yield their secrets. The apple in the picture will only be an apple until the painting is placed within a theoretical context in which it becomes a symbol for lust or sin. We cannot say we understand the "message" in the work of this particular painter until we see how the images he uses fit within that background.

Or consider an entirely different example taken from mathematics. Philip Kitcher offers the following case:

> When mathematicians found that there are methods for solving the general quadratic equation, the general cubic equation, and the general biquadratic equation in terms of the roots of the coefficients, it was rational for them to inquire whether the general polynomial equation of degree *n* admits of a similar method. Discovering that the general quintic is not solvable in radicals, but that some general classes of equations of high degree (for example, the *cyclotomic* equation $x^p - 1 = 0$, where p is a prime) do admit of a method of solution in radicals, it is rational to ask, as Galois did, under what conditions equations can be solved in radicals (1984, pp. 204-205).

After Galois's development of the theory that bears his name, the scattered methods for solving equations were incorporated within a general pattern or conceptual framework that made them intelligible.

Clearly, the four characterizations offered by Salmon do not define or determine four different concepts of understanding, and thus cannot serve as the basis for identifying different types of explanations. Furthermore, the main characteristic that Salmon attributes to scientific understanding, namely, being the type of knowledge that allows us to fit phenomena into "the general scheme of things," turns out to be a feature that it shares with all the other types of understanding that he identifies. Adding the second characteristic that Salmon attributes to scientific understanding—being the type of knowledge that tells us "how things work" and "what they are made of"—will not change our conclusion. Our interest in the material constitution and the internal workings of objects is exclusive to this type of

understanding simply because it is the only type of understanding in Salmon's typology that refers to the material aspect of things. But just as the difference between desiring a new car and desiring a sharper wit does not determine two different concepts of desire, the difference between understanding the structure of material things, and understanding the structure of a language, an artistic movement, a social institution, or a cultural pattern does not determine two different concepts of understanding. Whether the structure we are trying to understand is material or not is irrelevant. The same argument applies, *mutatis mutandis*, to our understanding of the causal processes that bring about changes both in material objects, and in social and cultural structures and institutions.

The conclusion of the foregoing analysis is that the *prima facie* difference between the explanations offered in scientific contexts and the ones offered in the course of everyday life is not a philosophically relevant one. Scientific explanations certainly have a higher degree of precision, detail, and complexity than their everyday counterparts. They normally involve the use of mathematics and of highly regimented languages. But a difference of degree is not a difference of kind. Whatever makes the former an explanation is what makes the latter an explanation. In brief, adding the word 'scientific' to the explicandum of a theory of explanation has no real philosophical significance, and for the most part I will ignore this distinction in what follows.

But to disregard the distinction between scientific and nonscientific explanations is not to open the door to an unrestricted use of the term 'explanation'. In particular, to ignore this distinction is not to erase the boundary between those explanations that simply provide psychological comfort to an individual, and the explanations that form an essential part of inquiry, broadly conceived as the search for new, error-free, valuable information. The type of explanations that a philosophical theory ought to be concerned with are the latter ones. Science, of course, is a party in the quest for true, valuable information, but so are many other human activities of lesser pedigree.

2. The Ambiguity of Explanation

We have completed the first step in our efforts to clarify the explicandum of a theory of explanation. But there is another central issue that must be clarified at the outset. As is well known, 'explanation' is an ambiguous term. It may refer either to the act or process by means of which we make something intelligible, understandable, or clear to someone, but it may also refer to the product or content of such an act. The process is a linguistic performance; the product is the content of the linguistic performance. Depending on which sense of the term one considers to be the primary one, a theory of explanation will focus either on the pragmatic aspects of the linguistic performance, or on the formal and material conditions that must be imposed on the content of that performance. There seems to be, once again, a dis-

agreement about what constitutes the legitimate explicandum of a theory of explanation.

Until the 1970s, most accounts of explanation had been constructed entirely in syntactical or semantical terms, disregarding the pragmatic aspects of explanation. But for someone who considers explanation to be essentially a linguistic transaction, the pragmatic aspects are most conspicuous. The development of formal pragmatics and the influence of empirical linguistics, especially through the work of Bromberger (1966), and Belnap and Steel (1976), allowed philosophers to build a strong case for the approach that takes individual speech acts as basic. The result has been an increasing tension between the pragmatic and the nonpragmatic approach. In order to see the source of this tension, and further clarify our *explicandum*, we must take a closer look at the claims and assumptions involved in these two approaches.[6]

2.1 *The Nonpragmatic Approach*

The defenders of the nonpragmatic approach argue that the idea of an explanatory act is logically dependent on the idea of an explanatory product, but not vice versa. More precisely, in their view a speech act is explanatory if and only if the information it contains is intrinsically explanatory. The explanatory character of such information must be defined independently of any contextual factor. There are dif-

---

6. The following classification owes much to Ruben's (1990) discussion of these issues.

ferent versions of this approach, but for convenience I will distinguish between two extreme versions, to be called *methodological rationalism* and *methodological naturalism*. The views held by different philosophers differ from these extremes in various ways, but they provide a convenient framework to understand the nonpragmatic approach.

According to the rationalist version of the nonpragmatic approach, there are certain general norms to which all explanations must conform. These are established largely on *a priori* grounds, in the light of some methodological or philosophical principle. It will be irrelevant whether any of the cases that we have believed to be genuine explanations actually falls under the defined concept. The concept of explanation thus characterized is introduced in order to improve or correct our actual cognitive practices, or to provide an ideal that serves as a guide for further inquiries.

In its naturalistic version, the nonpragmatic approach seeks to describe the essential logical features of actual explanations. The approach thus presupposes the existence of genuinely explanatory cases. The goal is to identify the methodological standards of procedure for the acceptance of explanations adopted by a learning community, which are a reflection of the epistemic values or norms to which they adhere.

Neither version of the nonpragmatic approach denies the importance of pragmatic considerations. Context always determines whether an explanation is

appropriate, suitable, or relevant. Nonetheless, the advocates of the nonpragmatic approach want to offer an account that does not need to be relativized to a certain individual or specific context. The fact that pragmatic conditions allow us to qualify an explanation in any of these respects, they maintain, does not mean that there are no audience-invariant conditions that ought to be imposed on the content of explanations, in the case of rationalism, or that accurately describe our cognitive practices, in the case of naturalism.

2.2 *The Pragmatic Approach*

According to the defenders of the pragmatic approach, pragmatics is all there is to explanation. In their view, the idea of an explanatory product is logically dependent on the idea of an explanatory act, but not vice versa. In other words, a speech act is explanatory if and only if it is performed with the intention of explaining something, and certain linguistic and contextual conditions based on the actual usage of the concept are met.

Van Fraassen and Achinstein are the best-known representatives of this approach. The main argument in support of this view is based on the fact that one and the same information can be used to achieve different linguistic goals. I can use a statement to describe something, to explain it, to criticize it, or to complain about it. The only way to differentiate an explanation from other linguistic performances is by arguing that an explanation cannot be reduced to the information

contained in it. "Explaining is what Austin calls an illocutionary act. Like warning and promising, it is typically performed by uttering words in certain contexts with appropriate intentions" (Achinstein, 1983, p. 16). The speech act depends primarily on the purposes and presuppositions of the speaker, on the context of utterance, and on any other factor that affects the success of the intended speech act.

2.3 *Explanation$_1$ and Explanation$_2$?*

It is tempting, in view of the sharp differences between the two approaches, to argue that we should deal with the notion of explanation in the same way that Carnap dealt with the notion of probability, namely, by arguing for the existence of two different explicanda—explanation$_1$ and explanation$_2$. I believe, however, that such an approach would be a mistake. My purpose, in the next two sections, is to show that a purely pragmatic or a purely nonpragmatic approach to explanation is irremediably incomplete. In my view, the tension between the pragmatic and the nonpragmatic approach can only be eliminated if we incorporate them into a broader analysis.

The plan for the rest of this chapter is as follows. In Section 3, I examine the insurmountable obstacles faced by a purely pragmatic approach. Most of that section will be devoted to the attempt to characterize explanation using why-questions. Section 4 deals with the reasons why pragmatic considerations are shunned from most theories of explanation. A closer look at the pragmatics of explanation

will reveal that the main argument for a purely nonpragmatic approach is simply a piece of methodological prejudice. I conclude the chapter by introducing a new approach to the pragmatics of explanation, which is the basis for the theory of explanation that I develop in the rest of the dissertation.

3. Why-Questions and the Pragmatics of Explanation

The starting point for a purely pragmatic approach to explanation is the characterization of an explanatory context. Since an explanation is viewed in this approach essentially as a type of linguistic transaction, it is necessary to identify the defining characteristics that distinguish an explanation from other types of linguistic transactions. And since we are considering a *purely* pragmatic approach, these characteristics cannot be based on the content of the speech act, only on the context of utterance, which includes the intentions of the individuals involved.

The easiest and most direct way to identify an explanatory context is by means of a why-question, that is, a question of the form 'Why (is it the case that) *p*?', where *p* is a statement. The why-question format has been adopted by virtually every philosopher as the standard for formulating explanatory requests. Simply being an answer to a why-question is obviously not enough to classify something as an explanation, so further conditions are necessary. In this section I examine some of the conditions that have been adopted by different authors. My purpose is to show that it is hopeless to attempt to characterize explanatory contexts by means

of why-questions, and more importantly, that the view that regards explanations essentially as answers to such questions is irremediably incomplete.

The only other way in which an explanatory context can be characterized without making reference to the content of the speech act involved is by considering the intentions of the speaker. This is the strategy followed by Achinstein. At the end of this section I examine his illocutionary theory of explanation and conclude that it also fails to provide a satisfactory account of explanation.

3.1 *Hempel's Use of Why-Questions*

In "Aspects of Scientific Explanation" (1965), Hempel introduces why-questions as part of his attempt to clarify the explicandum of his theory of explanation. In the beginning of that work, Hempel provides several guidelines to delimit the notion of scientific explanation. In the first place, he argues, "a scientific explanation may be regarded as an answer to a why-question" (p. 334). Secondly, why-questions whose explanandum is a nondescriptive singular referring term such as a noun make sense only when reconstructed as questions about whatever is properly characterized by means of a sentence. And thirdly, every adequate response to an explanation-seeking why-question should be able to serve as an adequate response to a reason-seeking why-question.

The first claim is a fairly weak one. It simply says that we can use the why-question format, not that we must. One of the purposes of the claim is to make ex-

plicit the fact that we are not interested in explanations that can be formulated as answers to how- or what-questions. Such cases would include "explaining the rules of a contest, explaining the meaning of a cuneiform inscription or of a complex legal clause or of a passage in a symbolist poem, explaining how to bake Sacher torte or how to repair a radio" (pp. 412-413). Hempel recognizes that the why-question format is not the only way to formulate explanation-seeking questions, although he claims that why-questions always provide an adequate standard phrase in which the request can be framed. So, for example, the questions '*How* did the crash happen' and '*What* is wrong with your TV?' can be rephrased, respectively, as 'Why was there an accident? and 'Why doesn't your TV work properly?'. Salmon makes the same claim: "A request for explanation can always be reasonably posed by means of a why-question. If the request is not originally formulated in such terms, it can, I believe, be recast as a why-question without distortion of meaning" (1984, p. 10).[7]

Neither Hempel nor Salmon makes the explicit claim that why-questions are sufficient to identify an explanatory context, as Hempel acknowledges in the following passage:

> Not all why-questions call for explanation, however. Some of them solicit reasons in support of an assertion. Thus statements such as 'Hurricane Delila will veer into the Atlantic', 'He must have died of

---

7. In later writings, Salmon has expressed some misgivings about this claim. See (1989, p. 137).

> a heart attack', 'Plato would have disliked Stravinsky's music' might be met with the question 'Why should this be so?', which seeks to elicit, not an explanation, but evidence or grounds or reasons in support of the given assertion. Questions of this kind will be called *reason-seeking* or *epistemic*. To put them in the form 'Why should it be the case that *p*?' is misleading; their intent is more adequately conveyed by a phrasing such as 'Why should it be believed that *p*?' or 'What reasons are there for believing that *p*?' (pp. 334-35).

Notice that Hempel's examples correspond, respectively, to a prediction, a hypothesis, and a counterfactual statement. Hempel argues that a why-question can be a disguised epistemic or reason-seeking question only in those cases in which *p* is an explanandum statement whose truth-value has not been established. Indeed, one of the conditions that Hempel imposes on explanation is that the explanandum statement must be true. Should we assume, then, that if *p* is known to be true, the question 'Why *p*?' will determine an explanatory context?

Unfortunately, this is not sufficient to distinguish the two types of questions. There are why-questions about true statements that seem best interpreted as requests for moral justification. Consider the following example provided by Salmon: "The question has been raised in courts of law as to why a member of a minority group was admitted to medical school to the exclusion of some non-minority candidate whose qualifications were somewhat better. The point at issue is the ethical basis for that decision" (1989, p. 136). Furthermore, there are why-

questions about true statements that are ambiguous; they can be interpreted either as requests for explanation or for justification. Consider the question 'Why did the French introduce *parité*, a constitutional amendment acknowledging the right of French women to equal access to elected office, and a subsequent law obliging the country's political parties to fill fifty per cent of the candidacies in virtually any race with women?'. An acceptable answer to this question will provide the ethical, political, and historical reasons the French adduced for implementing such a change. At the same time, however, these reasons, together with other circumstances that the French might not have explicitly considered, can also be used to provide a sociological or psychological explanation of their decision.

Hempel does not provide any further criteria to distinguish explanation-seeking from epistemic why-questions. His failure to provide any further criteria does not tell against his theory of explanation because his purpose was simply to use why-questions as a way of introducing the concept of explanation in an informal way. However, the difficulties described in this section show that the why-question format is not sufficient to determine an explanatory context. As we will see in the next section, the attempt to characterize explanation using the why-question format can also distort the nature of explanation.

3.2 *Why-Questions and Contrastive Explanations*

Dretske (1972), Garfinkel (1981), van Fraassen (1980), and many others, have used why-questions and the existence of sentential allomorphs to argue for the contrastive nature of explanation. For example, if one asks "Why did Adam eat the apple?" by stressing, in turn, "Adam," "eat," and "the apple," one can apparently generate three distinct requests for explanation. If the stress is on "the apple," for example, the question refers to the choice of fruit, and it can be rephrased as 'Why did Adam eat the apple, instead of another fruit?'. Thus, van Fraassen concludes, "The correct general, underlying structure of a why-question is therefore 'Why (is it the case that) *P in contrast to* (other members of) *X*?' where *X, the contrast class*, is a set of alternatives" (1980, p. 127).[8] If this analysis is correct, it introduces an irreducible element of pragmatic dependence into explanation because the choice of a contrast class depends on the interests of the questioner, not on any factive criteria.

There is no doubt that some explanations are contrastive. But from the fact that many why-questions are potentially ambiguous it follows neither that all explananda are contrastive nor that noncomparative why-questions are illegitimate. More importantly, there is a more natural way to account for the ambiguity of why-questions. Paul Humphreys argues that "it is at least as plausible to suggest that

---

8. It should be noted that not every erotetic theory argues for the contrastive nature of why-questions. Achinstein's (1983) illocutionary theory is a case in point.

this issue is just another facet of the deficiencies of propositional representations of explananda and that the ambiguity of many of these examples is a result of different aspects of the spatiotemporal events being selected as the subject of explanation" (1989, p. 137).

For instance, if I observe the flame of a Bunsen burner turning purple in the presence of a potassium salt, I can focus on the purpleness of the event, thereby picking out as my explanandum a single property of the complex physical event that I am observing. Furthermore, what I want explained is the purpleness of the event, not why it was purple rather than any other color.[9] The ambiguity only results when I try to fix the format of the explanandum description through the use of some particular why-question. But the fact that the question might be ambiguous does not entail that the property or event that I want to explain is comparative in form.

Since the identification of the explanandum is almost always a process mediated by natural language, the process is bound to inherit the vagueness and ambiguity of natural language. If we choose to use the why-question format to identify the explanandum, we should always be prepared to ask and answer a few questions before it becomes clear which is the property, object, fact, or event that

---

9. Ruben (1990) uses much stronger language to make the same point: "The fact that *p* rather than ~*p* is just a tedious pleonasm for the fact that *p*. ...if it is a fact that *p*, it follows, by double negation, which only those bordering on idiocy could fail to appreciate, that it is not a fact that ~*p*" (p. 41).

we are referring to.[10] This process is part of the pragmatics of explanation-giving, but it does not determine what counts as an explanation. The ambiguity of why-questions is thus purely derivative.

3.3 *Van Fraassen's Pragmatic Theory of Explanation*

The problems we have encountered so far with the why-question format are the result of giving too much importance to the linguistic form in which a request for explanation is formulated. The natural way of avoiding these difficulties is to focus on the underlying structure that a request for explanation ought to have, independently of what kind of interrogative sentence is used to express it. This is what van Fraassen does in *The Scientific Image* (1980).

Van Fraassen characterizes explanations *essentially* as answers to why-questions, where a why-question is an abstract entity "expressed by an interrogative (a piece of language) in the same sense that a proposition is expressed by a declarative sentence" (p. 138). He identifies the abstract why-question with the ordered triple $<P_K, X, R>$, where P_K is the topic of the question, X is the contrast class, which includes the topic P_K, and R is the relevance relation. Why-questions are raised within a context of background knowledge K. Why-questions have

---

10. This becomes more evident when we are dealing with the explanation of unfamiliar phenomena. In such cases, the initial description of the explanandum will be crude and misleading, not only because of the limitations in our knowledge, but also because we lack the appropriate conceptual and linguistic tools to refer to it.

presuppositions: each why-question presupposes that its topic is the only true member of the contrast class, and that there is at least one true proposition A that stands in the relation R to $\langle P_K, X \rangle$. The proposition A is the direct answer to the why-question.

An explanation of why P_K is the case thus takes the form:

(1) P_K in contrast to the rest of X because A.

The concept of explanation thus defined is a pragmatic one because the choice of the topic P_K and of the contrast class X depends entirely on the speaker. In a given context, two persons using the same interrogative sentence may express questions that have the same topic but different contrast classes, or vice versa. In the same fashion, the direct answers given to the same interrogative sentence can be different depending on context.

Since some explanations are clearly better than others, the theory also offers grounds for evaluating answers to why-questions. Van Fraassen offers three criteria: First, we must ask how probable the answer is in light of our background knowledge K. Second, we must ask to what extent the answer favors the topic vis-à-vis the other members of the contrast class X. Third, an answer must be compared to other possible answers in three respects: Is A (relative to K) more probable than the alternatives? Does it favor the topic to a greater extent? And, is it made wholly or partially irrelevant by other answers?

The main problem with van Fraassen's account lies in the relation R between why-questions and their answers. More precisely, van Fraassen does not impose any constraints on the relations that serve as relevance relations in why-questions. A number of his informal remarks give the impression that there are restrictions. For example, at the outset of his exposition he says that the "evaluation [of answers] proceeds with reference to the part of science accepted as 'background theory' in that context" (p. 141). And earlier on he says: "No factor is explanatory relevant unless it is scientifically relevant; and among the scientifically relevant factors, context determines explanatorily relevant ones" (p. 126). However, in the formal account there is no restriction on the relevance relation.

Kitcher and Salmon (1987) have shown that this lacuna makes the account vulnerable to trivialization. If no restriction is placed on R, it will be possible to prove that any true proposition can explain any other true proposition. More precisely, if A and B are any two true propositions, A will explain B in context K provided there is a question "Why B?" that arises in K for which A is a direct answer. Such a question is easily construed. Let $X = \{B, \sim B\}$ and $R = \{<A, <B, X>>\}$. Provided that K entails the truth of B and does not entail that the question has no answer, i.e., that there is no truth bearing R to $<B, X>$, then the direct answer to question $<B, X, R>$ in K is A. "If explanations are answers to why-questions," Kitcher and Salmon conclude, "then it follows that, for any pair of true propositions, there

is a context in which the first is the (core of the) only explanation of the second" (p. 319).[11]

The only way to avoid this trivialization is by imposing conditions on *R* in order to make it a *genuine* relevance relation. The problem is, as Kitcher points out, that this is no small deed: "A large part of the task of a theory of explanation is to characterize the notion of a genuine relevance relation" (1989, p. 415). Unless it is supplemented with such an account, van Fraassen's theory will remain incomplete. But in trying to provide an account of genuine relevance relations, van Fraassen would apparently be faced with the same problems confronted by all the nonpragmatic accounts of explanation that his pragmatic theory was supposed to replace.

3.4 *Achinstein's Illocutionary Theory*

Finally, we must consider Peter Achinstein's (1983) illocutionary theory of explanation. In an illocutionary theory, the intentions of the speaker are essential in determining the nature of explanation. Since the same proposition can be used to explain or to perform an entirely different speech act, the explanatory character of the speech act depends on the intentions of the speaker. In the case of why-questions, it could likewise be argued that their illocutionary aspect, not their content, is

---

11. Van Fraassen's criteria to evaluate possible answers to why-questions would not save the theory from trivialization because *A* is the *only* answer to the question <*B*, *X*, *R*>.

primary. As we have seen, why-questions can be raised with the intention of seeking an explanation, or with the intention of seeking epistemic or moral justification.[12] The fact that a why-question (or any other kind of question, for Achinstein does not limit his account to why-questions) was raised with the intention of finding an explanation would be sufficient to determine the existence of an explanatory context.

Implicit in the argument is the claim that the illocutionary product of an explanation-seeking question cannot be identical with the illocutionary product of a reason-seeking question because they were produced in speech acts of different kinds. But why should we assume that a necessary condition for a speech act to be an illocutionary product of a certain kind is that the person has actually produced it in an act of that kind? An assassin can lie shamelessly with the intention of misguiding an interrogator, but in the process he might contradict himself or he might infuriate the detective questioning him. Is the product of his speech act not a contradiction or a provocation because he did not intend it to be so? In Austin's theory, this phenomenon is handled with the notion of a perlocutionary act, but it is

---

12. Why-questions have many other uses. Consider the following examples: 'Why didn't you share your cookie with your sister, Johnny?' is more likely to be used as a reprimand than as a real question; 'Why can't you listen to your music at a decent volume, Kevin?' is probably a command; and 'Why did God do this to us?' is most likely an expression of grief and despair. Of course, there are contexts in which these same why-questions are used as requests for explanation.

generally agreed that Austin never fully succeeded in establishing a clear distinction between illocutionary and perlocutionary acts.

The claim that the product of a speech act depends on the kind of speech act in which it is produced simply assumes what it is used to prove. It seems to me, therefore, that the illocutionary theory does not provide a valid criterion to distinguish between explanation-seeking why-questions and other types of why-questions. For the same reason, the theory does not provide a valid criterion to distinguish between explanations and other types of speech acts.

We have seen in this section that all of the attempts to characterize an explanation as an answer to a particular why-question in a given context, or as a speech act performed with certain intentions, have failed to provide a satisfactory theory of explanation. This does not prove that a purely pragmatic approach is impossible, but it seems very unlikely that the connection between why-questions and their answers can be established without the aid of an audience-invariant relevance relation.

In the next section I will consider the other side of the issue by introducing some pragmatic considerations that will show that a purely nonpragmatic approach to explanation is equally misguided. Understanding the nature of explanation will require a mixed approach and a new way of conceiving of the pragmatics of explanation.

4. A New Approach to the Pragmatics of Explanation

The main reason why it is hopeless to characterize explanation using the why-question format, or any other canonical linguistic format, is that a why-question is simply one of the many possible ways in which a speaker can manifest her state of epistemic uncertainty regarding a particular phenomenon. The linguistic representation of an individual's epistemic situation is just that—a linguistic representation. To define an explanation as an answer to a question, even an abstract question, is to make explanations contingent on the way in which the speaker chooses to represent her epistemic situation. If a theory of explanation requires any reference to the individuals involved in the process of explaining, the focus ought to be directly on the epistemic state of someone searching for an explanation, and on how the person's epistemic state changes once an explanation is provided, not on the linguistic representation of these states. Isaac Levi summarizes the general point thus:

> One of the aims of inquiry is the acquisition of valuable new information. Although the way such information is represented linguistically may or may not facilitate such acquisition, in seeking valuable information, the inquiring agent is seeking to shift to belief states satisfying his curiosity. Linguistic representations of such belief states ought not to be confused with the states themselves (1991, 10-11).

The question we must now consider is whether it is really necessary to make any reference at all to the epistemic states of individuals in order to provide an

adequate account of explanation. Such reference seems granted by some *prima facie* characteristics of explanation. It is clear that the point and purpose of an explanation is to provide understanding. A request for explanation serves both to declare one's desire to understand something, and to specify what one wants to understand; the function of explanation is to provide the desired understanding. In order to account for these features of explanation, reference to the epistemic states of individuals is required.

Many philosophers, however, have claimed that the notion of understanding has no place in the study of explanation. Hempel, who takes a purely nonpragmatic approach to explanation, argues that "such expressions as 'realm of understanding' and 'comprehensible' do not belong to the vocabulary of logic, for they refer to the psychological or pragmatic aspects of explanation" (p. 413). The problem for Hempel is not that there is no connection between explanation and understanding: "Very broadly speaking, to explain something to a person is ... to make him understand it." The problem is that the use of terms like 'understanding' and 'comprehensible' compromise the *objectivity* of an explanation: "Thus construed, the word 'explanation' and its cognates are *pragmatic* terms: their use requires reference to the persons involved in the process of explaining" (p. 425).

Hempel recognizes that there are interesting issues associated with the process of providing an explanation in an actual context, and his intention is not to belittle their importance. But the concept of explanation that Hempel is characteriz-

ing is "a concept which is abstracted, as it were, from the pragmatic one, and which does not require relativization with respect to questioning individuals any more than does the concept of mathematical proof" (p. 426). The same general idea is defended by Rescher (1970), among many others:

> When we consider the topic of scientific explanation, we largely abstract from this pragmatic aspect of the interlocutor-involving setting within which an explanatory question arises: we imagine (or postulate) an abstract, impersonal framework, rather than a concrete dialectical setting. And we assume a range of questions marked off by abstracted conceptual boundaries of a *discipline* rather than by the personalized range of interests of an individual inquirer (p. 6).

Michael Friedman has pointed out that there is a certain equivocation about the term 'pragmatic' in Hempel's view. 'Pragmatic' can mean roughly the same as 'psychological', i.e., having to do with the thoughts, beliefs, attitudes, etc. of individuals. But 'pragmatic' can also be synonymous with 'subjective'. In the latter sense, a pragmatic notion must always be relativized to a particular individual. Friedman's claim is that "a concept can be pragmatic in the first sense without being pragmatic in the second." Further on he explains: "I don't see why there can't be an objective or rational sense of 'scientific understanding', a sense on which what is scientifically comprehensible is constant for a relatively large class of people" (1974, p. 8).

Hempel's avoidance of any pragmatic element in a theory of explanation can thus be evaluated in two different ways. If one takes 'pragmatic' to mean the same as 'subjective', Hempel's insistence in providing a nonpragmatic analysis of explanation, i.e., an analysis that does not depend on the idiosyncrasies of the individuals involved, is perfectly justified. But if 'pragmatic' is interpreted in Friedman's first sense, there is no reason why an analysis of the concept of explanation should not make reference to the epistemic states of the individuals involved in a cognitive project. Hempel rejects any appeal to the notion of understanding perhaps because he fears that the variations in the epistemic states and the cognitive interests of different individuals might not be resolvable. However, to preclude Friedman's strategy at the outset by denying that it is possible to construct a theory of objective explanation that takes into account the beliefs and aims of the participants in a cognitive enterprise, is to prejudge the issue.

I believe that we should take Friedman's suggestion seriously and explore the possibility of characterizing, in logically precise terms, a notion of explanation that is both objective and pragmatic, that does not depend on the idiosyncrasies of the individuals involved but that regards their epistemic states as a fundamental part of the analysis. The concept of explanation will still be an "abstraction," in Hempel's sense, but an abstraction based on the decisions that take place when an inquiring agent rationally accepts explanatory information, and not on the concrete dialectical settings in which specific explanations are provided. The resulting

concept will be a hybrid, a combination of the logical and the pragmatic dimensions of explanation.

The account I will provide is a normative or prescriptive one, not a descriptive one. My purpose is not to "naturalize" explanation by modeling the changes in an individual's epistemic state in the light of our current psychological knowledge. Rather, my purpose is to set conditions of rationality for the revisions that we perform on our belief system in our efforts to understand the world. On the other hand, I do not want to provide implausible conditions that no rational agent can fulfill. As Brian Ellis puts it, "strictly rational belief systems are only for the gods, and they have no need for logic anyway" (1979, p. 32). Instead, the desired account of explanation will take into account the role of explanation in the process of inquiry in which us mortals are usually engaged.

An account of explanation in terms of the rational transformations that take place in the epistemic states of individuals avoids most of the difficulties encountered by a purely pragmatic approach. In the first place, we will not have to deal with the performance conditions of explanatory speech acts, and thus with the problem of defining explanatory contexts. For the same reason, there will be no need to make reference to the intentions of the speaker. In fact, there will be no "speaker" to speak of. To be sure, language has to be used to represent epistemic states, but instead of natural language we will use highly regimented linguistic rep-

resentations of epistemic states in pretty much the same way that physicists use highly regimented language to represent physical states.

Secondly, the account offers a simple way of dealing with contrastive explanations. As we saw in Section 3.2, most of the confusion that surrounds this issue is generated by the vagueness and ambiguity of natural language. An account of explanation in terms of epistemic states avoids the problem of sentential allomorphs by providing direct access to the explanandum via a regimented language.

Finally, my account of explanation is intended to succeed where van Fraassen's failed, namely, in providing objective criteria to determine when an explanation is genuinely relevant to its topic. The account that I develop in Chapters 3 and 4 is more closely related to van Fraassen's work on rational belief (1980a) than to his account of explanation in *The Scientific Image*.

In the next chapter I introduce the technical tools that will allow us to characterize the notion of explanation in logically precise terms. These tools are then used in chapters 3 and 4 to offer the pragmatic account of explanation outlined in this section.

CHAPTER 2
THE LOGIC OF EPISTEMIC CHANGE

In order to develop an account of explanation that takes into consideration the epistemic states of the individuals involved, we need to introduce a formal representation of these states, and of the way they change as a result of the individual's interaction with the world. In recent years, the model of belief revision developed by Peter Gärdenfors in *Knowledge in Flux* (1988) has become the guiding light in the study of epistemic states. The model is based on his joint work with Carlos Alchourrón and David Makinson (1985), and is known as the AGM model. The best critical commentary of the AGM model can be found in Isaac Levi's *The Fixation of Belief and Its Undoing* (1991), which is also an elaboration of the account of belief change contained in his earlier book *The Enterprise of Knowledge* (1980).

Levi's theory of belief revision and the pertinent contributions from the AGM model will be the basis on which I will construct the account of explanation that I defend in Chapter 4. In the first two sections of this chapter I present the central elements of each theory, beginning with the AGM model. There is a vast literature on the model and I do not pretend to cover every single aspect of it. The discussion in this chapter focuses on the way that epistemic states and their changes are logically represented in each of the two theories, and on specific details that will prove relevant later on. In the last section I discuss two alternative ways of logically representing epistemic states, the Bayesian models used in de-

cision theory and game theory, and the modal approach introduced by Hintikka in *Knowledge and Belief* (1962). I will argue that the AGM-Levi approach to belief revision is preferable to either of these options.

1. The AGM Model

The following presentation of the AGM model is largely based on Gärdenfors's book *Knowledge in Flux*. There are some differences between the original AGM model and the version developed by Gärdenfors, but unless otherwise indicated, these differences are largely irrelevant for our purposes.

1.1. *Belief Sets*

The starting point of the AGM model is the linguistic representation of an epistemic state by a *belief set*. A belief set is the set of sentences in an object language L that a person accepts at time t. Accepting a sentence ϕ in a belief set K entails full belief, in the sense that in K there is no doubt about the truth of ϕ at t.[1]

1. Like Levi, Gärdenfors claims that certainty or full belief does not entail incorrigibility. Unlike Levi, however, Gärdenfors does not develop the epistemological implications of the belief-doubt model: He says, "On the theory presented here a belief system can almost be regarded as a closed system; the only external factor it presumes is the class of epistemic inputs. And ... we need not even know what the inputs really are; what matters are the effects they have on a belief system. This means that it does not matter much for the internal *theory* of belief systems what the factual connections with the external world are" (1988, p. 19). Gärdenfors says that the beliefs in K are to a large extent the result of sensorily mediated contacts between the belief system and the world, but nothing in the model reflects this assertion.

Since the main use of belief sets is to represent *rational* epistemic states, two conditions of rationality are imposed on belief sets: the set should be consistent, and it should be closed under logical implication.[2] Both of these criteria presuppose a logic for the language L. Besides containing expressions for the standard truth-functional connectives, the language L is governed by a consequence relation $\vdash$ between a set of sentences in L and a sentence in L. A sentence ϕ is logically valid iff it is a consequence of the empty set. The relation is assumed to satisfy the following conditions (p. 24):

($\vdash$ 1) If ϕ is a truth-functional tautology, then $\vdash \phi$.

($\vdash$ 2) Modus ponens. That is, if $\vdash \phi \rightarrow \psi$ and $\vdash \phi$, then $\vdash \psi$.

($\vdash$ 3) Not $\vdash \perp$. That is, $\vdash$ is consistent.

It follows that $\vdash$ contains classical propositional logic. It is assumed that $\vdash$ satisfies the deduction theorem and that it is compact. Gärdenfors (p. 24) defines a belief set thus:

(Def BS) A set K of sentences is a (nonabsurd) belief set iff (i) $\perp$ is not a logical consequence of the sentences in K and (ii) if $K \vdash \phi$, then $\phi \in K$.

---

2. These are the only two rationality criteria imposed on epistemic states by both Gärdenfors and Levi. Their theories correspond to what Harman (1986, ch. 4) has called "coherence theories of belief revision," that is, theories in which no belief has a special justificatory status and in which any change is made so as to increase and preserve overall coherence.

Gärdenfors also defines an *absurd* belief set, denoted $K_{\perp}$, as the set L of all sentences. A *maximal* belief set K is a belief set such that, for every sentence ϕ in the language, either $\phi \in K$ or $\sim\phi \in K$.

The set of all logical consequences of a set K is denoted $Cn(K)$. From (Def BS) it follows that all belief sets satisfy the following condition:

(Cn) $K = Cn(K)$.

Belief sets thus include both the beliefs that a person explicitly accepts at a given time, and all of the consequences of those beliefs.

The requirements imposed on belief sets clearly make them unsuitable as realistic descriptions of an individual's actual epistemic situation. Nonetheless, Gärdenfors argues that the criteria are useful as ideals of rationality. Belief states should be thought of as "equilibrium states," that is, as rational states of belief that are "in equilibrium under all forces of internal criticism" (p. 9-10).

1.2 *Belief Change Operators*

There are only three epistemic attitudes associated with any sentence ϕ and any belief set K:

(i) ϕ is accepted: $\phi \in K$.

(ii) ϕ is rejected: $\sim\phi \in K$.

(iii) ϕ is indetermined: $\sim\phi \notin K$ and $\phi \notin K$.

A change of belief concerning ϕ consists in changing one of these epistemic attitudes into one of the others. Gärdenfors groups the six possible change operations into three groups: expansions, revisions, and contractions.

In an *expansion*, the epistemic attitude 'ϕ is indetermined' is changed into either 'ϕ is accepted' or '$\sim\phi$ is accepted'. The expansion of K by ϕ is denoted K^{+}_{ϕ}. Such changes are the result of epistemic inputs—information added to an epistemic state, either through observation or reliable testimony. None of the beliefs accepted in an expansion can contradict any belief in the epistemic state, i.e., expansion into inconsistency is forbidden. Gärdenfors does not provide any criteria for the acceptance of ϕ. It is assumed in AGM that the epistemic inputs have been filtered in some way. As we will see in Section 2, Levi fills in this lacuna in the AGM model by providing conditions for the rational acceptance of new information.

Revisions occur when either 'ϕ is accepted' is changed to '$\sim\phi$ is accepted', or '$\sim\phi$ is accepted' is changed to 'ϕ is accepted'. A revision of K by ϕ is denoted K^{*}_{ϕ}. A revision is the result of an observation or testimony that contradicts one's current epistemic state. As in expansion, legitimate revisions of our epistemic states can only be represented as changes from one nonabsurd belief set to another. The central rationality criterion on revisions is that the revision of K by ϕ should be the minimal change of K that is consistent and includes ϕ. The idea that a revision

should only produce a minimal change in K is a reflection of a methodological principle that Gärdenfors uses in several places—the criterion of informational economy: "Information is in general not gratuitous, and unnecessary losses of information are therefore to be avoided" (p. 49).

Finally, in a *contraction*, the epistemic attitude 'ϕ is accepted' or '$\sim\phi$ is accepted' is changed into 'ϕ is indetermined' In other words, in a contraction the belief that ϕ (or that $\sim\phi$) is given up and is not replaced by its negation. The contraction of K by ϕ is denoted K^{-}_{ϕ}. The main problem concerning contractions is that, when retracting a belief ϕ from K, there may be other beliefs in K that jointly entail ϕ. In order to maintain closure, we must therefore give up some of those beliefs as well. The problem is then to determine which beliefs should be given up since we do not want to give up beliefs unnecessarily—the criterion of informational economy is also operative in contractions.

A further question regards the epistemological motivation for having a separate contraction operator. Why would anyone simply give up a belief that he fully accepts—as opposed to giving it up because of the acceptance of new information that contradicts it? Gärdenfors argues that one might hypothetically contract one's belief in ϕ "in order to give the negation of ϕ a hearing without begging the question" (p. 47). Levi offers the same reason. The main use of contraction is thus in hypothetical and counterfactual reasoning. Contractions also play a central role in Gärdenfors's account of explanation, which I will discuss in the next chapter.

These three belief change operators are essentially functions taking a belief set K and an epistemic input ϕ to a new belief set K^+_ϕ, K^*_ϕ, or K^-_ϕ. Gärdenfors provides rationality postulates that specify the requirements that the respective operators should satisfy. The postulates are then related to the belief change operations via representation theorems. I begin with the rationality postulates for expansions.

1.2.1 Expansions

(K^+1)	For any sentence ϕ and any belief set K, K^+_ϕ is a belief set.	(closure)
(K^+2)	$\phi \in K^+_\phi$.	(success)
(K^+3)	$K \subseteq K^+_\phi$.	(inclusion)
(K^+4)	If $\phi \in K$, then $K^+_\phi = K$.	(vacuity)
(K^+5)	If $K \subseteq H$, then $K^+_\phi \subseteq H^+_\phi$.	(monotonicity)
(K^+6)	For all belief sets K and all sentences ϕ, K^+_ϕ is the smallest belief set that satisfies (K^+1) - (K^+5).	(minimality)

The first postulate expresses the fact that + is a function from $K \times L$ to K, where K is the class of all belief sets and L the class of all sentences. (K^+2) states the acceptance of ϕ in K^+_ϕ. (K^+3) says that no beliefs are retracted in an expansion. In the abnormal case in which $\sim\phi \in K$, adding ϕ produces an inconsistency. In that case K^+_ϕ is $K_\perp$, the absurd belief set, which is a superset of all belief sets. (K^+4) states that no expansion is necessary if the epistemic input is already believed.

Monotonicity says that if one belief set contains at least the same information as another belief set, then the expansion of the former will contain at least the same information as the expansion of the latter with respect to the same sentence. (K^+6) is an expression of the criterion of informational economy. These postulates lead to the following representation theorem:

THEOREM 2.1 The expansion function + satisfies (K^+1) - (K^+6) iff $K^+_\phi = Cn(K \cup \{\phi\})$.

The postulates for expansion uniquely determine the expansion of K by ϕ as the set of all logical consequences of K together with ϕ, and thus lead to an explicit definition of the expansion process. The postulates for revision and contraction will not allow such definitions. The main reason is that logic alone is not sufficient to determine which beliefs should be given up in a revision or in a contraction of K. Gärdenfors argues, however, that the postulates, together with certain nonlogical factors, uniquely determine the content of K^*_ϕ and K^-_ϕ. We will examine these nonlogical factors in Section 1.3. The following postulates thus only circumscribe the set of rational revisions and contractions, but they do not uniquely determine the respective change in belief.

1.2.2 Revisions

(K*1) For any sentence ϕ and any belief set K, K^*_ϕ is a belief set. (closure)

(K*2) $\phi \in K^*_\phi$. (success)

(K*3) $K^*_\phi \subseteq K^+_\phi$. (inclusion)

(K*4) If $\sim\phi \notin K$, then $K^+_\phi \subseteq K^*_\phi$. (preservation)

(K*5) $K^*_\phi = K_\perp$ iff $\vdash \sim\phi$. (vacuity)

(K*6) If $\vdash \phi \leftrightarrow \psi$, then $K^*_\phi = K^*_\psi$. (substitution)

(K*7) $K^*_{\phi\&\psi} \subseteq (K^*_\phi)^+_\psi$. (superexpansion)

(K*8) If $\sim\psi \notin K^*_\phi$, then $(K^*_\phi)^+_\psi \subseteq K^*_{\phi\&\psi}$. (subexpansion)

The first two postulates are straightforward. (K*3) says that an expansion always yields a larger belief set than a revision unless, (K*4), the negation of the epistemic input is not accepted in K. Gärdenfors argues that (K*3) and (K*4) entail that expansion is a special case of revision.[3] Vacuity says that K^*_ϕ is always consistent, except when the negation of ϕ is a tautology. (K*6) states that logically equivalent sentences yield the same revision of a belief set. Gärdenfors says that (K*1)-(K*6) are elementary requirements, and that (K*7) and (K*8) are introduced in order to deal with iterated changes. They can be regarded as generalizations of inclusion and preservation. "The idea is that, if K^*_ϕ is a revision of K and K^*_ϕ is to be changed by adding further sentences, such a change should be

3. Michael Levin has pointed out to me that if expansion is a case of revision, then Gärdenfors's initial definition of the revision operator has to be modified to include the case in which no substitution takes place.

made by using expansions of K^*_ϕ whenever possible" (p. 55). The reason is, once again, the criterion of informational economy.

A postulate that Gärdenfors briefly considers and then rejects is a generalization of (K^-5), the postulate of monotonicity:

(K*M) If $K \subseteq H$, then $K^*_\phi \subseteq H^*_\phi$.

A counterexample can be easily constructed. Consider a belief set K such that $\sim\phi \notin K$ and $\psi \in K$, and let H be $K^*_{\psi \to \sim\phi}$. It follows that $\sim\phi \in H$ because ψ and $\psi \to \sim\phi$ are both elements of H. Now let us assume that $\psi \in K^*_\phi$. If we want to revise H by ϕ, then either ψ or $\psi \to \sim\phi$ has to be retracted in order to avoid inconsistency. But in many cases there will be powerful nonlogical reasons to retract ψ rather than $\psi \to \sim\phi$. The result will be that $\psi \in K^*_\phi$ but $\psi \notin H^*_\phi$, thus contradicting (K*M).

1.2.3 Contractions

(K^-1)	For any sentence ϕ and any belief set K, K^-_ϕ is a belief set.	(closure)
(K^-2)	$K^-_\phi \subseteq K$.	(inclusion)
(K^-3)	If $\phi \notin K$, then $K^-_\phi = K$.	(vacuity)
(K^-4)	If not $\vdash \phi$, then $\phi \notin K^-_\phi$.	(success)
(K^-5)	If $\phi \in K$, then $K \subseteq (K^-_\phi)^*_\phi$.	(recovery)
(K^-6)	If $\vdash \phi \leftrightarrow \psi$, then $K^-_\phi = K^-_\psi$.	(substitution)

(K⁻7) $K^-_\phi \cap K^-_\psi \subseteq K^-_{\phi\&\psi}$. (intersection)

(K⁻8) If $\phi \notin K^-_{\phi\&\psi}$, then $K^-_{\phi\&\psi} \subseteq K^-_\phi$. (conjunction)

Inclusion requires that no new beliefs occur in K^-_ϕ. Vacuity is motivated by the criterion of informational economy. Success states that the only kind of epistemic input that cannot be retracted is a logical truth. Recovery guarantees that all the beliefs in K will be contained in the belief set that results from contracting and expanding K by the same sentence. This postulate is false in probabilistic contexts, and it is rejected by Levi and many others. I return to it in Section 2.2.2 where I discuss Levi's account of contraction. (K⁻6) is analogous to (K*6). As in the case of revisions, postulates (K⁻7) and (K⁻8) are supplementary; they deal specifically with the retraction of conjunctions. Intersection states that the beliefs that are in both K^-_ϕ and K^-_ψ are also in $K^-_{\phi\&\psi}$. Conjunction states that, when contracting K with respect to ϕ & ψ, either ϕ or ψ or both must be given up. If ϕ is rejected (for nonlogical reasons), then the minimal change in K necessary to give up ϕ & ψ is the same as required to reject ϕ itself. Finally, a monotonicity postulate is not adopted for reasons analogous to the ones presented against (K*M).

1.2.4 The Levi and Harper Identities

Levi (1977) has argued that the only two legitimate forms of change of belief are expansions and contractions. Revisions, he claims, should be analyzed as se-

quences of contractions and expansions. In Gärdenfors's interpretation of this claim, a revision of K by ϕ should be analyzed as an initial rejection of $\sim\phi$ and the beliefs that entail it, and the posterior addition of ϕ. Gärdenfors calls this description of revision the *Levi identity*:

(Def *) $K^*_\phi = (K^-_{\sim\phi})^+_\phi$.

The question is whether the Levi identity satisfies the postulates for revision. Gärdenfors (p. 69) establishes the following two representation theorems:

THEOREM 2.2 If the contraction function – satisfies (K^-1)-(K^-4) and (K^-6), and the expansions satisfy (K^-1)-(K^-6), then the revision function * obtained from (Def *) satisfies (K^*1)-(K^*6).

THEOREM 2.3 Suppose the assumptions of theorem 2.2 are fulfilled. Then (a) if (K^-7) is satisfied, ($K*7$) is satisfied, and (b) if (K^-8) is satisfied, ($K*8$) is satisfied for the function generated by (Def *).

Harper (1977), on the other hand, has shown that contractions can be defined in terms of revisions:

(Def –) $K^-_\phi = K \cap K^*_{\sim\phi}$.

Gärdenfors calls this the *Harper identity*. The idea is that since $K^*_{\sim\phi}$ is the minimal change in K needed to accommodate $\sim\phi$, it contains as much as possible of K that does not entail ϕ. Gärdenfors then provides the following representation theorem:

THEOREM 2.4 If the revision function * satisfies (K*1)-(K*6), then the contraction function – generated by (Def –) satisfies (K–1)-(K–6).

Gärdenfors then proves that the revisions and contractions defined via the Levi and Harper identities are interchangeable, in the sense that "if we start with one definition to construct a new contraction (or revision) function and after that use the other definition to obtain a revision (or contraction) function again, then we ought to get the original function back" (p. 70).

1.3 *Contraction Functions and Epistemic Entrenchment*

The last aspect of the AGM model that we will consider is perhaps the most problematic. We have seen that in forming a contraction K^-_ϕ of a belief set K, the criterion of informational economy requires that K^-_ϕ should contain as much as possible from K without entailing ϕ. The problem is that typically there will be several possible ways of forming a contraction that does not entail ϕ. How should we determine exactly which beliefs should be given up?

Gärdenfors provides two solutions that lead to equivalent proposals. The first one is based on the maximal subsets of K that do not contain ϕ.[4] The second one uses the degree of epistemic entrenchment of the sentences in K. If the relative

4. Alchourrón and Makinson (1985) propose a solution based, not on *maximal* subsets of K that do not contain ϕ, but rather on *minimal* subsets of K that contain ϕ. They call it *safe contraction*. See also Gärdenfors (1988, section 4.8).

epistemic entrenchment of the sentences in K can be determined, this information should be used when forming the contraction so that only the propositions with the lowest degree of entrenchment are retracted.

Alchourrón and Makinson (1982) tried to give a solution to the contraction problem by finding *maximal* subsets of K that fail to imply ϕ. A belief set K' is a maximal subset of K that fails to imply ϕ iff (i) $K' \subseteq K$, (ii) $\phi \notin K'$, and (iii) for any ψ such that $\psi \in K$ but $\psi \notin K'$, $\psi \rightarrow \phi$ is in K'. Usually there are several such subsets for any K. The set of all belief sets K' that are maximal subsets of K that fail to imply ϕ is denoted $K \perp \phi$.

Using this idea, one can construct a selection function S that picks out one element $S(K \perp \phi)$ of $K \perp \phi$ for any K and any ϕ whenever $K \perp \phi$ is nonempty. K^-_ϕ can then be defined by the following rule:

(Def Max) $\quad K^-_\phi = S(K \perp \phi)$ when not $\vdash \phi$;
$\quad\quad\quad\quad\quad\quad K^-_\phi = K$ otherwise.

Contraction functions thus defined are referred to as *maxichoice contractions* in AGM (1985). The problem with maxichoice contractions is that they are too large. Alchourrón and Makinson show that if a revision function is defined using a maxichoice contraction via the Levi identity, then, for any sentence ϕ and any belief set K such that $\sim\phi \in K$, K^*_ϕ will be maximal. In other words, one could reach a complete view on every issue expressible in L by contracting using a maxi-

choice function and then replacing what is removed by its negation. In a sense, maxichoice functions *create* maximal belief sets.

Another way of using $K \perp \phi$ is by assuming that K^-_ϕ contains only the propositions that are common to all of the maximal subsets in $K \perp \phi$. More precisely,

(Def Meet) $K^-_\phi = \bigcap(K \perp \phi)$ whenever $K \perp \phi$ is nonempty;
$K^-_\phi = K$ otherwise.

This kind of function is called a *full meet* contraction function. The problem in this case is the opposite of maxichoice contraction—the function is too small. Alchourrón and Makinson (1982) prove that if a revision function is defined using a full meet contraction function via the Levi identity, then, for any sentence ϕ and any belief set K such that $\sim\phi \in K$, $K^*_\phi = Cn(\phi)$. The revision will contain only the sentence and its logical consequences.

Since using only one of the maximal subsets in $K \perp \phi$ yields a contraction set that is too large, and using all of them yields a set that is too small, the natural step is to explore the consequences of using only some of the maximal subsets. That is the idea behind a *partial meet* contraction function. A selection function S can be used to pick out a nonempty subset $S(K \perp \phi)$ of $K \perp \phi$, if the latter is nonempty. The contraction function is then defined as follows:

(Def Part) $K^-_\phi = \bigcap S(K \perp \phi)$.

The intuitive idea is that S picks out the most epistemically entrenched maximal subsets in $K \perp \phi$. Thus it must be assumed that there is a partial ordering of the maximal subsets in $K \perp \phi$. Technically, we need to introduce the notation $M(K)$ to denote the *union* of the family of all the sets $K \perp \phi$, where ϕ is any proposition in K that is not logically valid. It is then assumed that there is a transitive and reflexive ordering relation $\leq$ on $M(K)$. This relation can then be used to define a selection function S that picks out the top elements in the ordering.

(Def S) $\quad S(K \perp \phi) = \{K' \in K \perp \phi\colon\ K'' \leq K' \text{ for all } K'' \in K \perp \phi\}$

A contraction function defined by (Def S) is called a transitively relational partial meet contraction function. The following theorem connects the rationality postulates for contraction with this result:

THEOREM 2.5 For any belief set K, a contraction function – defined over K satisfies (K^-1) - (K^-8) iff – is a transitively relational partial meet contraction function.

The main problem with this approach is that partial meet contractions are not closed under intersection. That means that the meet of two maximally entrenched maxichoice contractions need not be maximally entrenched. We will examine the consequences of this result in Section 2.2.2, when we compare partial meet contractions to the solution offered by Levi (1991).

The second approach to the contraction problem suggested by Gärdenfors is based on the pretheoretical properties of the notion of epistemic entrenchment. By introducing a set of rationality postulates for the qualitative properties of this notion, Gärdenfors is able to provide a constructive definition of contraction (and revision) functions. Gärdenfors does not offer a quantitative measure of epistemic entrenchment—a project that would require a substantial addition to the meager resources of the AGM model—because he believes that "the problem of uniquely specifying a revision function (or a contraction function) can be solved, assuming only very little structure on the belief sets apart from their logical properties" (1992, p. 17).

At the outset it is crucial to bear in mind that the epistemic entrenchment of a sentence is *not* connected to its subjective probability. If a sentence has been accepted in K, it is judged to be maximally probable. Nonetheless, an individual does not regard all of her beliefs as having equal epistemic entrenchment. The informal criterion to determine the degree of epistemic entrenchment of a sentence is "how useful it is in inquiry and deliberation" (p.87). Some true sentences are central to our cognitive endeavors and others are just epistemically inert. Entrenchment is connected to explanatory and predictive power, and to overall informational value.

The epistemic entrenchment of a sentence is relative to the belief set in which it occurs. Different belief sets are associated with different orderings of epistemic entrenchment, even if some of the sentences in the sets overlap. If ϕ and

ψ are sentences in L, the notation $\phi \leq \psi$ is used as shorthand for 'ψ is at least as epistemically entrenched as ϕ', and $\phi < \psi$ for 'ψ is epistemically more entrenched than ϕ'. The relation $\leq$ is defined only in relation to a given belief set K. Gärdenfors (1988, pp. 89-91) offers the following postulates for the relation $\leq$:

(EE1)	For any ϕ, ψ, and γ, if $\phi \leq \psi$ and $\psi \leq \gamma$, then $\phi \leq \gamma$.	(transitivity)
(EE2)	For any ϕ and ψ, if $\phi \vdash \psi$, then $\phi \leq \psi$.	(dominance)
(EE3)	For all ϕ and ψ in K, $\phi \leq \phi \,\&\, \psi$ or $\psi \leq \phi \,\&\, \psi$.	(conjunctiveness)
(EE4)	When $K \neq K_{\perp}$, $\phi \notin K$ iff $\phi \leq \psi$ for all ψ.	(minimality)
(EE5)	If $\psi \leq \phi$ for all ψ, then $\vdash \phi$.	(maximality)

(EE1) is a minimal requirement for any ordering relation. The justification for (EE2) is that, if ϕ entails ψ and either ϕ or ψ must be retracted from K, it is a smaller change to retract ϕ. If we retracted ψ, ϕ would have to be retracted anyway in order to retain closure, so we lose less by retracting ϕ. (EE3) says that, since we must retract either ϕ or ψ to retract $\phi \,\&\, \psi$, the informational loss incurred by retracting $\phi \,\&\, \psi$ is at least the same as that incurred by giving up ϕ or ψ. (EE4) states that when a sentence is not contained in K, it is not entrenched at all and is thus minimal in the ordering. (EE5) states the opposite, namely, that when a sentence is logically valid, and thus contained in every nonabsurd belief set, it is

maximal in ≤. Logical truths have the highest degree of epistemic entrenchment because they will never be given up.

The following condition relates ≤ to contraction functions:

(C≤) $\psi \leq \phi$ iff $\psi \notin K^{-}_{\phi \& \psi}$

The intuitive idea is that if K is contracted by ϕ & ψ, either ϕ or ψ must be given up. ψ should be retracted just in case ϕ is at least as epistemically entrenched as ψ. ϕ and ψ are equally entrenched only when they are both logically valid.

With the aid of (C ≤), we reach the desired representation theorem:

THEOREM 2.6 A contraction function – satisfies ($K^{-}1$)-($K^{-}8$) iff ≤ satisfies (EE1)-(EE5), where $\psi \leq \phi$ iff $\psi \notin K^{-}_{\phi \& \psi}$ for all ϕ and ψ in L.

The theorem states that for any well-behaved contraction function there exists an ordering of epistemic entrenchment that generates the function. The same result applies to the revision function, via the Harper identity. Theorem 2.6 shows that the problem of constructing appropriate contraction (and revision) functions can be reduced to the problem of providing an appropriate ordering of epistemic entrenchment.

Much more remains to be said about the AGM model, most of it critical in nature. Some of the problems will emerge in our discussion of Levi's account of belief revision, and others will be considered in Chapter 3. In the next section I

present Levi's approach to belief revision, which has much in common with the AGM model, but which provides a much more detailed picture of the deliberations that go into an agent's decision to change his mind.

2. Levi's Approach to Belief Revision

Levi treats inquiry as a decision problem. Two desiderata guide an inquiring agent's decisions to modify her present state of belief: (i) the acquisition of valuable new information, and (ii) the avoidance of error. Inquiry proceeds from a background of firmly held beliefs, the agent's state of full belief, which the agent regards as error-free. Given the agent's cognitive goals, an optimal change in belief is one that provides new error-free information. But the avoidance of error and the acquisition of new information are goals that pull in different directions. On the one hand, error is best avoided by remaining in the original state of full belief. On the other, a maximum amount of information is acquired by shifting to states that might not be error-free. "The moral of the story is that those shifts that increase informational value incur risk of error" (1991, p. 12).

In order to provide an account of how these two conflicting cognitive goals can be balanced, Levi has to begin by clarifying what a state of full belief is. When an inquiring agent consciously accepts a set of propositions at any given time, he commits himself to the consequences of these propositions. The agent may not recognize all the consequences of his beliefs at the time, but he cannot rationally

deny them if made aware of them. By consciously accepting the consequences of his beliefs, the agent fulfills his *doxastic commitment*. But in virtue of his limited reasoning abilities, his *doxastic performance* will always fall short of his doxastic commitment. The set of propositions to which an agent is committed at any given time is the agent's state of full belief.

Given this distinction, the phrase 'X fully believes that ϕ' threatens to be ambiguous. It could be a partial description of X's doxastic commitment at a given time, or of X's conscious acceptance of a belief. Levi reserves the phrase 'X fully believes that ϕ' for the former, and uses 'X fully recognizes that ϕ' for the latter. "With this understood, we may say that when X fully believes that ϕ at t, he may not fully recognize that ϕ at that time, even though he is committed at t to doing so" (p. 8).

2.1 *Conceptual Frameworks*

In order to consider the modifications of an agent's state of full belief, some conception of the range of possible changes is required. The set K that contains an agent's potential states of full belief $\boldsymbol{K}$ is the agent's conceptual framework. Potential states of full belief can be compared in terms of their *strength*—if $\boldsymbol{K}_1$ is stronger than $\boldsymbol{K}_2$, the latter is a consequence of the former. The comparison in terms of strength generates a partial ordering from stronger to weaker potential

states of full belief.[5] Levi imposes further conditions on conceptual frameworks. Given any two potential states of full belief $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$, there should be a potential state that is the strongest common consequence of both. This state is the *meet* $\boldsymbol{K}_1 \wedge \boldsymbol{K}_2$ of both states.[6] The availability of the meet of two states in a conceptual framework has an important consequence for our purposes—it makes intersubjective agreements possible:

> Consider two inquirers, X and Y, sharing a common framework. On some occasions, it may be desirable for both X and Y to modify their views by adopting a belief state representing the shared agreement or common ground between them. To do this entails that they both give up informational value and, hence, incur a cost that they seek to minimize. In particular, they do not want to give up any more information than will be needed to bring them into agreement. The assumption of the existence of the [meet] of $\boldsymbol{K}_x$ and $\boldsymbol{K}_y$ allow for the conceptual availability of such a move to both X and Y (p. 13).

For a similar reason, it should be possible to form the *join* $\boldsymbol{K}_1 \vee \boldsymbol{K}_2$ of the two states, that is, the weakest potential state that is as strong as each of the states in the pair. When an agent wants to acquire more information, he can do so by adding another agent's views to his. But in so doing, he incurs the risk of error.

---

5. That is, the relation of strength is reflexive, transitive, and antisymmetric.
6. In this particular passage, Levi uses the terms 'meet' and 'join' to denote the union and intersection of two states of belief, respectively. Throughout the dissertation I will use the opposite, more common usage to avoid confusion. Thus the meet is the intersection, and the join the union of two potential states of full belief.

Since the agent does not want to increase the risk gratuitously by shifting to a stronger state than is necessary, he should shift to the weakest state that incorporates the other agent's views.

The existence of a meet and a join for each pair of elements in K guarantees that a conceptual framework is a *lattice*. Furthermore, a conceptual framework contains a uniquely weakest state which is the consequence of all other states. Such a state, labeled **1**, is the state of maximum ignorance. The conceptual framework also contains a uniquely strongest state **0**, the inconsistent state, that has every other state as a consequence.[7] Given a state of full belief $\boldsymbol{K}_i$ in the framework, there should be another state $\boldsymbol{K}_j$ such that $\boldsymbol{K}_i \wedge \boldsymbol{K}_j$ has **0** as a consequence. Levi assumes there is a weakest $\boldsymbol{K}_j$ having the given relation to $\boldsymbol{K}_i$ and labels it the pseudocomplement $\neg\boldsymbol{K}$.

With these elements in place, and the assumption that a conceptual framework forms a *distributive* lattice, it is possible to say that K satisfies the requirements of a pseudoboolean algebra. Levi further assumes, against the intuitionists, that the join $\boldsymbol{K} \vee \neg\boldsymbol{K}$ is identical to **1**, and $\neg\boldsymbol{K}$ thus becomes the complement of $\boldsymbol{K}$. This latter assumption allows him to say that a conceptual framework is a distributed and complemented lattice, and thus to claim that "the potential states of full

7. The reader must keep in mind that Levi's use of '0' to designate the top of a boolean algebra, and of '1' to designate the bottom, is a departure from common usage.

belief in a conceptual framework K are partially ordered in a manner satisfying the requirements of a boolean algebra" (p. 16).

Intuitively speaking, the set K of potential states of full belief can be described as the set of states that the agent is conceptually capable of adopting at a given time. But the phrase 'conceptual capacity' invites several interpretations. It might refer to the intellectual ability of an individual, in which case the set of states of belief available to an agent will have the structure of a boolean algebra only in a very limited number of cases. Or it may refer to the potential states that, at least in principle, any intelligent agent has access to, in which case no revision of conceptual frameworks would be possible—as Davidson (1984) seems to argue.

Levi wants to suggest a third way of understanding conceptual capacity. Given that the main cognitive goal of any inquiring agent is to acquire new, error-free, valuable information, the agent might restrict the potential states of full belief that he wants to consider to those states that promote his cognitive goals. Even if the agent is able to embed his current state in a partial ordering of belief states constituting on a boolean algebra, the agent will not be interested in potential states of full belief that do not provide the kind of information that she considers valuable. Levi summarizes the point thus:

> The notion of conceptual capacity in the third sense is a notion that is relativized to the aims of the inquiring agent. ... Thus, it seems to me that there is a sense in which different agents may embrace different conceptual frameworks and the same agent may endorse different

> frameworks at different times. But disagreements over and revisions of conceptual frameworks are *au fond* disagreements over and revisions of cognitive values (p. 20).

In order to fully understand the role that cognitive values have in Levi's approach, we must consider the conditions of rationality that he imposes on legitimate expansions and contractions of an agent's state of full belief.

2.2 *Belief Revision*

Before examining Levi's account of belief revision, it will be convenient to consider how states of belief can be linguistically represented. This will allow us to find the parallels between Levi's approach and the work of AGM, but it is important to keep in mind that the linguistic representation of states of belief is not an essential component of Levi's theory of belief revision. A potential state of full belief $\boldsymbol{K}$ can be represented in a regimented language L by a potential corpus K. Potential corpora are identical to Gärdenfors's belief sets. Levi requires L to have a first-order predicate logic, vocabulary and axioms interpretable as set theory and arithmetic, as well as extralogical vocabulary. Regardless of the complexity of L, Levi does not assume that all potential states of full belief can be represented in L. Potential corpora are closed under logical implication, and there are corpora $\boldsymbol{I}$ and $\boldsymbol{0}$ representing the potential states **1** and **0**. The set of all potential corpora in L closed under first-order deduction form a boolean algebra.

Levi identifies four types of belief change: expansions, contractions, replacements, and residual shifts. An *expansion* is a change from a state of full belief to one that is at least as strong. If K_1 and K_2 are a pair of potential corpora such that $K_1 \subseteq K_2$, then K_2 is an expansion of K_1. A *contraction* is a change from a state of full belief to one that is at least as weak. If K_1 and K_2 are a pair of potential corpora such that $K_1 \subseteq K_2$, then K_1 is a contraction of K_2. A *replacement* is a change in which the meet of the initial state and the subsequent state is the inconsistent state. K_1 is a replacement of K_2 iff $K_1 \wedge K_2 = \boldsymbol{0}$ and both K_1 and K_2 are distinct from $\boldsymbol{0}$. Replacements correspond to Gärdenfors's revision functions. A *residual shift* is a change in which the initial state and the subsequent state are not comparable according to the partial order, and their meet is not the inconsistent state. K_1 is a residual shift from K_2 iff K_1, K_2 and $K_1 \wedge K_2$ are all distinct from $\boldsymbol{0}$ and from each other.

There are two fundamental theses that Levi defends in the context of belief change, the commensurability thesis and the commensuration requirement. Levi (p. 65) explains the theses thus:

Commensurability Thesis: Given an initial state of full belief $\boldsymbol{K}_1$ and another state of full belief $\boldsymbol{K}_2$, there is always a sequence of expansions and contractions, beginning with $\boldsymbol{K}_1$, remaining within the space of potential states of full belief and terminating with $\boldsymbol{K}_2$.

Commensuration Requirement: Every legitimate change in belief state from initial K_1 to K_2 is decomposable into a sequence of contractions and expansions each of which is legitimate.

The first thesis was notoriously rejected by Kuhn (1970) and Feyerabend (1962), but it is just a consequence of the condition that the conceptual framework should satisfy the requirements of a boolean algebra. I will return to the commensurability thesis in Chapter 4. The second thesis, which Gärdenfors believes is captured by the *Levi identity*, essentially claims that there are only two legitimate changes in belief states: contractions and expansions. Both replacements and residual shifts are derivative. That means that we need only consider the properties of expansions and contractions.

2.2.1 Expansions

Expansions come in two varieties, routine and deliberate. In the former, an epistemic state is expanded in response to external stimuli such as observation, experimentation, or the testimony of witnesses or experts. In a routine expansion it is assumed that the agent has antecedently committed himself to the reliability of the process, program, or routine by means of which the information is obtained. "Commitment to using a program for routine expansion brings with it commitment to full recognition that the routine is sufficiently reliable" (p. 75).

The second type of expansion is through a deliberative or inferential process. The difference between the two types of expansion is a philosophically important one. A deliberate expansion involves a process of decision and justification. Given a series of possibilities for expansion, the decision to expand in one direction and not in another must be evaluated in terms of how well the expansion promotes an agent's cognitive goals. The type of inference involved in deliberate expansion is not deductive. A deductive inference does not lead to an expansion of a state of full belief; rather, it is a case of fully realizing what one's cognitive commitments are. The inference involved in a deliberate expansion is an ampliative one, and in that sense, Levi says, it can be called *inductive*.

In order to expand her state of belief, the agent must first identify a set of potential expansions. Her present state of full belief $\boldsymbol{K}$ must be used as the standard to decide what constitutes a serious possibility, and her cognitive goals and interests must be used to determine the range of potential expansions. The resulting set, which Levi calls an *ultimate partition*, contains all the doxastic propositions that are consistent potential expansions of $\boldsymbol{K}$ such that one and only one element of the set is true if $\boldsymbol{K}$ is. Identifying an ultimate partition is the task of abduction.[8]

8. Levi points out (p. 172, n. 6) that abduction thus understood is not an ampliative inference, but rather an exercise in conjectural thinking. Pagnucco and other researchers in AI interpret abduction as inference to the best explanation. I will examine Pagnucco's approach in the next chapter.

Once this set of potential expansions has been identified, the agent is forced to balance the desideratum to expand into the state that contains more information, and the desideratum to avoid error. Informational value is assessed using a content function that varies inversely with the probability of the state to be added.[9] The risk of error varies inversely with the credal probability assigned to the potential states in the ultimate partition. Each of these evaluations orders the potential expansions in opposite ways. According to the information measure, the agent should expand into inconsistency; according to the risk measure, she should not expand at all. The resolution of the conflict, Levi argues, is a weighted average of the utility functions representing the two desiderata.[10]

A fundamental difference between the approaches of Gärdenfors and Levi is that the latter tolerates accidental expansions into inconsistency. We have seen that when a proposition ϕ is added to a belief set (potential corpus) K, and $\sim\phi$ is a consequence of K, $K^{+}_{\phi} = K_{\perp}$ ($K \wedge \phi = \mathbf{0}$). In Gärdenfors's view, "$K_{\perp}$ is the inconsistent endpoint that should be shunned at all costs. It is impossible to get out of $K_{\perp}$ by expansion because $(K_{\perp})^{+}_{\phi} = K_{\perp}$ for all ϕ. Figuratively, $K_{\perp}$ is epistemic hell" (1988, p. 51).

---

9. The set of potential expansions of $\boldsymbol{K}$ is a boolean algebra; its maximum is $\boldsymbol{K}$ and the minimum is $\mathbf{0}$. Thus it is possible to define a probability function that generates a partial ordering of its elements.

10. A more detailed presentation of deliberate expansion will be provided in Chapter 4, when we consider the decision process involved in the acceptance of an explanation.

Gärdenfors's fear of epistemic hell is the main reason why he considers revision to be more fundamental than expansion. When $\sim\phi$ is a consequence of K, an expansion by ϕ is legitimate only if the result is a revision of K. Notice, however, that the revision cannot be decomposed into a sequence of expansions and contractions, as demanded by the commensuration requirement, because the initial expansion would be an expansion into inconsistency. In other words, Gärdenfors cannot transform what he calls the *Levi identity*

(Def *) $\quad K^*_\phi = (K^-_{\sim\phi})^+_\phi$

into $K^*_\phi = (K^+_\phi)^-_{\sim\phi}$, which seems a more natural way to represent a revision of the agent's state of belief.[11] In consequence, it is clear that Gärdenfors cannot accept Levi's commensuration requirement, and that what he calls the *Levi identity* is only a partial description of Levi's full requirement.

Levi claims to have found the way out of epistemic hell. In his view, expansion into inconsistency is prohibited only when it is the result of a deliberate expansion. But in routine expansion it is often the case that a new observation or experiment injects inconsistency into an agent's state of belief. In such cases the agent must contract his state of belief in order to "extricate himself from this embarrassment" (p. 94). The contraction may call into question background assump-

11. Sven Ove Hansson (1993) also makes this last claim in an article entitled, appropriately, "Reversing the Levi Identity."

tions and theories, as well as the reliability of the process that brought about the conflicting piece of information. The question is, How can an agent contract from inconsistency? In order to answer this question we must consider Levi's analysis of contractions.

2.2.2 Contractions

An agent's retrenchment from an inadvertent expansion into inconsistency via routine expansion is called a *coerced* contraction. But not all contractions are coerced. In some cases a contraction is deliberately chosen in order to realize the inquirer's cognitive goals. The most important case in which a contraction can be beneficial is when the agent considers the truth of a potential explanation E. If the explanation is inconsistent with presently held beliefs, the agent may want to give the potential explanation a fair hearing. In order to do so, he must contract $\sim E$, and any other belief that has $\sim E$ as a consequence, in order not to beg the question against E. But in order not to beg the question *in favor* of E, the agent must remain in suspense about the truth of E. The decision to adopt E will then be made in accordance with the decision procedures described for the case of expansion.

In the case of coerced contractions, we find a much more difficult problem. When the agent inadvertently expands into the inconsistent corpus $\boldsymbol{0}$, she must try to move into a consistent state as soon as possible. The problem is that, by classical closure, the corpus $\boldsymbol{0}$ contains every sentence in the language L. Thus for every

sentence ϕ in L, ϕ & $\sim\phi$ must be removed by contracting either ϕ or $\sim\phi$. An infinite task awaits those who fall into epistemic hell.

Levi's way out of epistemic hell is by keeping track of your last epistemically virtuous state of full belief. Since the expansion into inconsistency was the result of adding ϕ to K, the inconsistency can be removed in three ways. (i) by removing ϕ, (ii) by questioning the background assumptions in K, or (iii) by questioning both. Informally speaking, the situation can be described as follows. In the first case, ϕ may be removed when doubts arise about the reliability of the program for routine expansion or as to whether the program was properly implemented. The retraction of ϕ must thus be accompanied by the retraction of the belief in the reliability of the program for routine expansion, or in its correct implementation. In the second case, $\sim\phi$ is removed from K, and then K is expanded by ϕ. In order to avoid inconsistency, other beliefs that entail $\sim\phi$ must also be removed.

As we saw in Section 1.3, any contraction of K to a weaker state K' can be done in a number of different ways. So in order to give a more precise account of both coerced and deliberate contractions, Levi must confront exactly the same problems about contraction that Alchourrón, Makinson, and Gärdenfors had to face in their seminal papers of the 1980s.[12]

---

12. The account provided below is the one offered in *The Fixation of Belief and Its Undoing* (1991). It is often referred to as "Levi contraction" in the belief revision literature. In recent years, Levi has modified some of the technical details of the account, although the underlying principles remain unchanged. The technical strategy for contraction that he now accepts is what Rott and Pagnucco (1999) call "severe withdrawal." Levi (1998) prefers the name "mild

Levi begins the technical analysis of the contraction problem by considering *saturatable* contractions, which are weaker than the maxichoice contractions with which AGM (1985) began their analysis. Let $C(K, \phi)$ be the set of contractions of K by removing ϕ. A contraction will be saturatable iff the result of adding $\sim\phi$ is maximally consistent in K. $S(K, \phi)$ is the set of all saturatable contractions.[13]

Since any contraction involves a loss of information, the agent should choose the contraction strategy that minimizes that loss. In the case of expansion, informational value was assessed by means of a content function that varied inversely with the probability of the state in question. But the same method cannot be used in the case of contractions. The reason can be explained as follows. If informational value is determined by means of a probability-based function, the set of optimal contractions—the set of contractions that minimize loss of informational value—will always contain at least one maxichoice contraction, even if there are other equally optimal saturatable contractions. If the agent adopts the strongest contraction available, she should choose a maxichoice contraction. But the adoption of a maxichoice contraction leads to the unacceptable consequences identified

---

contraction." I have opted for the earlier version because the 1991 proposal was originally conceived as a response to the contraction strategies set forth in AGM (1985), and the presentation below tries to preserve the comparative character of Levi's original account. Furthermore, I do not deal with the technical details of contraction in my account of explanation, so the omission of the latest technical developments in the literature are not too important for my purposes.

13. Saturatable contractions are weaker than maxichoice contractions because the third condition for maxichoice contractions, namely, that if $K' \subseteq K$, then for any ψ such that $\psi \in K$ but $\psi \notin K'$, $\psi \rightarrow \phi$ is in K', can fail for some ψ in K. Thus $S(K, \phi)$ is a proper subset of $K \perp \phi$.

by Alchourrón and Makinson that we studied in Section 1.3. Unfortunately, we are no better off by choosing a saturatable contraction because we obtain the same undesirable results.

The solution advocated by Alchourrón, Gärdenfors, and Makinson (1985) was to choose partial meet contractions. The idea behind partial meet contractions is that the inquirer should adopt the intersection of all the maximal subsets that he considers optimal according to the partial ordering of epistemic entrenchment. The problem with partial meet contractions is that they are not closed under intersection. Thus, either there are optimal contractions that are not members of $K \perp \phi$, or contractions may be allowed to be suboptimal. AGM seem to lean towards the second alternative. In some cases we might not be able to find a partial meet contraction that fully respects the principle of informational economy. Nonetheless, a partial meet contraction will be the best approximation to an optimal contraction.

Levi finds this solution unacceptable. On the one hand, the restriction of available options for contraction to maximal subsets of K seems unjustified. On the other, Levi believes that any contraction we adopt should *always* be optimal. The first objection is tied to the type of monotonicity requirement for informational value adopted by AGM. The restriction to maximal subsets is defended on the grounds that (i) informational value increases with an increase in strength, (ii) maximal subsets of K are stronger than saturatable contractions, and (iii) the agent should minimize loss of informational value. Levi justly complains that if (i) were

to be adopted in all contexts, then the best contraction strategy would always be maxichoice contraction, which both AGM and Levi reject. The AGM position is defensible only if they restricts the range of application of the strong monotonicity requirement to the comparison between maximal subsets and saturatable contractions, which seems quite *ad hoc*.

Instead of (i), Levi adopts the weaker monotonicity requirement that stronger states should carry *at least as much* informational value as the weaker states they entail.[14] This means that the range of available options for contraction must also include saturatable contractions. The best contraction strategy, in Levi's view, is the meet of optimal saturatable contractions.

The inclusion of saturatable contractions among the available options for contraction does not take care of Levi's second objection to the AGM contraction strategy. Like maxichoice contractions, saturatable contractions are not closed under intersection. Once again, there might be sets of saturatable contractions, all of which are optimal, but whose intersection fails to be optimal. Levi's challenge is thus to find a condition that insures that the intersection of a set of optimal saturatable contractions is always optimal.

The most natural way to proceed is to modify the measure of informational value for contractions. Instead of the probability-based measure used for expan-

14. The importance of using the weaker monotonicity principle will become more evident when we examine its application in the case of explanatory expansions in Chapter 4.

sions, Levi defines an index M of *damped* informational value. The only difference with a probability-based measure is that in the damped version the informational value of the *intersection* of a set of saturatable contractions, where each minimal saturatable contraction in the set has the same probability-based informational value, is equal to the common informational value. This guarantees at once that the intersection of saturatable contractions can be optimal provided each of the individual contractions is optimal.

These results are summarized in the following definition:

> K^{-}_{ϕ} is the meet of all saturatable contractions of K by removing ϕ that minimize loss of informational value among saturatable contractions as determined by the given M function (p. 130).

Levi's rejection of the strong monotonicity requirement has an important consequence—one of the main postulates of the AGM model, the recovery postulate for contraction,

(K⁻5) If $\phi \in K$, then $K = (K^{-}_{\phi})^{+}_{\phi}$,

fails for Levi's notion of contraction. In fact, Levi's adoption of a weaker monotonicity requirement is largely motivated by this result. In his view, recovery should not be a condition on acceptable contractions in all contexts. Levi provides compelling examples of contractions of K by ϕ in which it is rational to give up more beliefs than allowed by the strong monotonicity requirement. From the re-

sulting state K^-_ϕ it is not possible to expand back to K simply by adding ϕ, thus violating recovery.

Rott (1991) has proven that if a revision function is defined via the Levi identity using Levi's notion of contraction, i.e., without using the recovery postulate, the function will be identical to Gärdenfors's original revision function. This result is also implicit in Theorem 2.2 above. In consequence, it makes no difference whether or not we accept the recovery postulate in deciding what should count as an admissible revision (or expansion). The postulate is only important when dealing with the question of what constitutes a minimal contraction.

I will conclude this section by considering how Levi's solution to the contraction problem can be applied to coerced and deliberate contractions. As we saw above, coerced contractions occur when the agent inadvertently stumbles into inconsistency. Levi identified three strategies that the agent could follow to extricate himself from that situation:

1. The agent can question the background assumptions that contradict the new information ϕ. This is equivalent to finding the admissible contraction $K^-_{\neg\phi}$ of the initial corpus, and then expanding by ϕ. Let this be K^*. This contraction strategy is equivalent to Gärdenfors's notion of revision.

2. The agent can question the new information ϕ that led him into inconsistency. This is equivalent to contracting K^* by ϕ, and then expanding it by $\sim\phi$. Let this be K^{**}.

3. The agent can question both ϕ and $\sim\phi$, thus contracting to a state that is the intersection of K^* and K^{**}.

Levi's suggestion is that an agent should not contract from $\boldsymbol{0}$ to a corpus unless the contraction is attainable by one of these three strategies. The agent "should not end up with more information than he would have gotten had he followed the program for routine expansion without stumbling into inconsistency or had he aborted the program before expanding into inconsistency" (p. 149).

In order to make this idea more precise from a decision theoretical point of view, Levi modifies the M function to make its assessment sensitive to the provenance of the inconsistency.[15] Let $MC(K, \phi)$ be the set of all maximally consistent corpora that are expansions of elements of $\boldsymbol{K^*}$ and $\boldsymbol{K^{**}}$. In order to guarantee that loss of damped informational value is minimized, any contraction from $\boldsymbol{0}$ due to expansion of K by ϕ should be an intersection of elements of $MC(K, \phi)$, "subject to the proviso that the M values of the elements of these sets be the normalizations of

15. The reason why Levi needs to modify the M function is that if M remains fixed, it will always recommend contracting into the same state, regardless of how the agent stumbled into inconsistency. The recommended state will be the weakest consistent corpus carrying maximum damped informational value.

the values initially assigned when all maximally consistent corpora not in the set are assigned the value *0*" (p. 150). In consequence, the recommended contraction from inconsistency will be K^* if its damped informational value is greater than that of K^{**}, and K^{**} in the opposite case. If K^* and K^{**} have the same damped informational value, the optimal contraction will be the intersection of the two.

The full implementation of Levi's strategy for contraction from inconsistency would lead to a considerable complication of the AGM framework. In the next chapter I will consider a relatively simpler strategy to deal with contraction from inconsistency. The main problem with Levi's strategy is that it seems unnatural to demand that we always make backup copies of our belief states before we move on. Inconsistencies in inquiry are often discovered when it is too late to reconstruct our last epistemically virtuous belief state. In such cases the best strategy will be to isolate the inconsistency from the rest of one's beliefs, and solve the problem locally, as David Lewis (1982) suggests. To be sure, the implementation of this strategy also requires the introduction of additional logical mechanisms in order to quarantine inconsistencies. It seems that there is no easy way out of epistemic hell.

I conclude by considering the case of deliberate contractions. This type of contraction occurs, as we saw above, when the agent wants to give an explanatory hypothesis ϕ a fair hearing. In this case, the function M can be used to decide

whether he is justified in removing $\sim\phi$ from K. Let $K^* = K^-_{\sim\phi}$ and $K^{**} = (K^-_{\sim\phi})^+_{\phi}$. There are three possible outcomes of shifting from K to K^*:

1. The agent may decide that $\sim\phi$ was true after all. Here the acceptance or rejection of the recovery postulate makes a slight difference. If recovery is accepted, the agent returns to K; if it is not, he returns to a state weaker than K. In neither case is there a loss or gain of damped informational value. In the course of deciding that ϕ is false, the agent will usually acquire other beliefs due to observations and experiments, so strictly speaking the state he returns to is not identical to K even if recovery is accepted. But if this complication is ignored, there will be neither gain nor loss.

2. The agent may decide that ϕ is true. The informational value of K^{**} can be greater than, lesser than, or equal to that of K^*.

3. The agent may not make up his mind, thus refusing to add ϕ or $\sim\phi$ to his corpus. The agent suffers a loss of informational value.

From a decision theoretical point of view, the best option is to refuse to contract to K^*, unless K^{**} carries more informational value than K. "There is no sense in incurring the risk of ending up with K^* unless there is a chance of benefiting by ending up with K^{**}" (p. 156). Informally, one should not give up a cherished theory unless there is another theory that has the potential to be more informationally

valuable. The same analysis will be true in the case of explanatory hypotheses, as we shall see in the beginning of the next chapter.

3. Alternatives to the AGM-Levi Approach

The AGM-Levi approach to belief revision will be the starting point for the characterization of the concept of explanation that I develop in the next chapter. But before we make use of the approach, there is a final question that must be considered. It is necessary to explain why the AGM-Levi approach is preferable to other ways of representing epistemic states. In this section I examine two other ways in which the beliefs of an individual can be logically analyzed: the Bayesian approach and the modal model. After offering a very brief sketch of each theory, I explain why the AGM-Levi approach is better suited for our purposes.

3.1 *Bayesian Models*

The best-known models of epistemic states are the Bayesian models used in decision theory and game theory. In such models, a state of belief is represented by a probability measure defined over some object language or over some space of events. The intended interpretation is that the probability function provides a measure of an individual's degree of belief in a sentence or proposition. It is generally assumed that all the information that is relevant for decision making is conveyed by such probability measure.

Neither the AGM model nor Levi's theory of belief revision possess a notion of degree of belief, and in that sense they are less informative than the Bayesian model. However, since we are interested only in sentences that are accepted as certain, there will be no need to use the full complexity of probability functions. Furthermore, if we assume, as many Bayesian models do, that a sentence ϕ is accepted as certain relative to P iff $P(\phi) = 1$, it is easy to establish a relation between these models and the AGM model. A probability function P defined over an object language L generates a belief set iff, for all sentences in L, $P(\phi) = 1$ iff $\phi \in K$ (Gärdenfors, 1988, pp. 38-39). Some authors, for example, de Finetti and Carnap, make a distinction between acceptability and maximal probability. The distinction is important in some contexts,[16] but it does not play any role in the present one and will be disregarded.

Finally, the AGM-Levi approach has a comparative advantage over the Bayesian approach. According to traditional Bayesian approaches, rational changes of belief can be described by conditionalization of probability functions whenever this process is defined, that is, when the information to be added is consistent with the probability function. This means that conditionalization is essentially an expansion of the epistemic state represented by the probability function. The problem is that neither traditional conditionalization nor Jeffrey conditionalization are useful

16. For example, if one wants to estimate the value of a real-valued parameter x that ranges between 0 and 1, the hypothesis that x has an irrational value normally has probability 1, but it need not be accepted as true for that reason.

when dealing with contractions and revisions of epistemic states. Although Gärdenfors offers some suggestions as to how the Bayesian approach can be expanded to deal with these types of epistemic change, the simplicity of the AGM postulates, combined with our exclusive interest in sentences that are accepted as certain, make the AGM-Levi approach a better choice for our purposes.

3.2 *Modal Models*

A second way of modeling epistemic states is based on the modal approach introduced by Hintikka in *Knowledge and Belief* (1962). In this type of model, a belief is seen as a relation between an individual and a proposition. The rationale adduced for using propositions is twofold. On the one hand, the objects of belief are not sentences but rather the nonlinguistic content of sentences. On the other, belief is only one of several propositional attitudes, and propositions seem to be required in order to explain what it is that all the attitudes directed towards the same content have in common.

The modal approach requires the introduction of two additional logical operators. The object language is augmented by the epistemic operators B_a and C_a, which are the formal counterparts of '*a* believes that' and 'it is compatible with everything that *a* believes that'. The formal semantics for these operators is then developed in terms of possible worlds (or of model sets, in Hintikka's case). Propositions are generally identified with sets of possible worlds—a sentence ex-

presses a given proposition iff the sentence is true in exactly those possible worlds with which the proposition is identified. An epistemic state is modeled by a subset K of the set of all possible worlds. The intended interpretation is that K is the narrowest set of possible worlds in which the individual is certain to find the actual world. What an individual accepts in a given epistemic state is exactly what is true in all worlds included in K. In consequence, the more you learn, the fewer possible worlds are compatible with what you accept.

The main objection to the modal approach is based on an appeal to Occam's razor. According to both Gärdenfors and Levi, it is possible to offer an account of belief change that achieves the same results as the modal approach without "the bells and whistles that introducing propositions and possible worlds provides" (Levi, 1991, p. 27).

Levi argues that if one takes belief to be a relation between an individual and a proposition, two boolean algebras will be required, one for the set of propositions and one for the set of states of full belief. This duplication can be avoided if one identifies the propositions with the states of full belief, thus reducing two algebras to one.

> When we use statements like "X believes that ϕ," we are not compelled to take this to claim that X bears a belief relation to a proposition that ϕ distinct from a potential state of belief. I suggest that the expression "that ϕ" represents a potential state of full belief. ... "X believes that ϕ" partially describes X's state of full belief by partially locating his state

> of full belief in the boolean algebra of potential belief states according to the partial ordering of potential belief states (p. 23).

The idea is not to give belief a preferential treatment over the other propositional attitudes. But since belief is presupposed in attributing any other propositional attitude to an agent, it is possible to represent the differences between the different propositional attitudes "as differences in the structures they induce on potential expansions of X's state of full belief" (p. 26).

Gärdenfors, on the other hand, also argues for the dispensability of propositions and possible worlds. In Chapter 6 of *Knowledge in Flux* he shows how, instead of starting the analysis of epistemic states with belief sets, which presuppose classical logic, one can start with unstructured epistemic states as the only fundamental entities and then define a proposition as a function from epistemic states to epistemic states. Gärdenfors defines a *belief model* as a pair $\langle K, Prop \rangle$, where K is a set of epistemic states and *Prop* is a class of functions from K to K. Intuitively, these functions represent epistemic inputs that lead to expansions, contractions, and revisions of epistemic states. By introducing a set of postulates for these functions, Gärdenfors is able to define a relation of logical consequence between propositions and to introduce the logical connectives. The class of all models satisfying these postulates yields a boolean algebra. Although Gärdenfors calls these functions "propositions," they bare little resemblance to the propositions used in the modal model. "The ontological base for the construction is meager. The only

entities that have been assumed to exist are epistemic states and functions defined on epistemic states. ... the construction does not in any way use the concept of a possible world" (p. 144).

The two proposals reach the same point through different routes, as Levi points out: "Once one has ensured that the space of potential states of full belief has the structure of a boolean algebra, Gärdenfors's approach and the one I follow come out equivalent" (p. 29). Since there are no good reasons, and plenty of bad reasons to adopt the modal model, my adoption of the AGM-Levi approach to epistemic states seems justified.

In the next chapter I examine the problems and questions that arise when we consider the notion of explanation in the context provided by the AGM-Levi approach to belief revision. The analysis of these elements will clear the field for the account of explanation that I propose in Chapter 4.

CHAPTER 3
EXPLANATION AND BELIEF REVISION

An explanation is an epistemic vehicle. By providing the relevant information, an explanation takes an inquiring agent from a state of belief in which he lacks understanding of a fact to a state in which such understanding is achieved. Such is the simple intuition behind the account of explanation that I present and defend in this chapter and the next. In the present chapter I focus on some preliminary aspects, including a survey of the few accounts of explanation in the belief revision literature. Chapter 4 contains a more detailed presentation of my approach.

I begin the chapter by offering an analysis of the different epistemic contexts or circumstances in which an inquiring agent might decide to accept an explanation. This analysis will allow us to see the complexities involved in characterizing the changes that the agent must make to his epistemic state in order to understand a given phenomenon. There have been very few attempts to characterize explanation as a belief revision operation, and most of them tend to confuse explanation with justification. In Section 2, I analyze and ultimately reject two of these proposals. Finally, Section 3 provides a brief informal presentation of the account of explanation that I develop in more detail in Chapter 4.

1. The Epistemic Contexts of Explanation[1]

To obtain a general picture of the problems involved in providing an account of explanation in terms of belief revision, I will begin by considering all the possible epistemic contexts in which an agent's failure to understand a phenomenon prompts her to consider the acceptance of an explanation.[2] The epistemic context of an explanation relative to a corpus of beliefs K is determined by the following variables: (i) whether the agent antecedently accepts, rejects, or neither accepts nor rejects one or more elements of the explanation that he eventually accepts; (ii) whether the agent's acceptance or rejection of the elements of the explanation is explicit or implicit, (iii) whether the agent fully recognizes that the explanation *is* an explanation, and (iv) whether the explanandum is consistent or inconsistent with the agent's present belief state.

The only fixed element in the epistemic context of an explanation is the explanandum. I require that if the agent wants to explain the fact stated by a sentence ϕ,[3] she must fully *believe* and fully *recognize* that ϕ is true. As we saw in Chapter

---

1. The epistemic context of an explanation should not be confused with the explanatory contexts mentioned in Chapter 1. The latter are dialogical settings in which linguistic and contextual factors are supposed to determine which linguistic transactions can be considered explanations.
2. The analysis presented here focuses on the state of belief of an individual agent, but the explanations that he considers adding to his corpus might be the result of a joint decision. The details about how such joint decisions are made will be provided in Chapter 4. To anticipate: If the explanation is the result of the individual's isolated assessment of the situation, the explanation will be called a "subjective explanation." If it is based partially or totally on the assessment of a learning community, the explanation will be an "explanation in K" or an "objective explanation."
3. For brevity, I will often use the phrase 'the explanation of ϕ' for 'the explanation of the fact stated by ϕ' and 'understanding ϕ' for 'understanding the fact stated by ϕ'.

1, a universally accepted requirement on explanation in the philosophical literature is that the explanandum sentence must be true. The requirement that the agent must fully believe that the explanandum sentence is true is its equivalent in the context of belief revision.[4] I also require that the agent fully *recognize* the truth of ϕ because it is evident that the agent's decision to look for and accept an explanation can only be triggered by a belief that is explicitly accepted. To be sure, an explanation of ϕ will also explain any other sentence entailed by ϕ, whether it is explicitly believed or not,[5] but a sentence that is not explicitly accepted cannot impel the agent to look for an explanation of the fact stated by that sentence.

In order to analyze the epistemic contexts of explanation, we need to characterize the following notions. A *potential explanation* of ϕ in K is a sentence that would partially explain ϕ if it were accepted in K and the resulting belief state were adjusted to preserve consistency.[6] Let Γ be the set of all the partial explanations of

---

4. Since I want to consider cases in which the explanandum is inconsistent with the agent's belief state, the requirement that the agent must accept the explanandum sentence entails that the belief state that constitutes the starting point of an explanation can be inconsistent. In such cases, I will call the explanandum an *anomaly*.

5. This claim will be disputed by those who deny that explanation is closed under logical implication. Peter Lipton (1991, p. 51) mentions the following counterexample, which he attributes to Ted Williamson: The rage for paisley explains why all the men in a restaurant are wearing paisley ties, but the rage for paisley does not explain why they are wearing ties. Since the fact that a person is wearing a paisley tie entails that he is wearing a tie, closure fails. I do not believe the counterexample is conclusive. A more accurate description of the situation is that the rage for paisley explains why all the ties that the men are wearing have a paisley print. This explanandum does not entail that all the men are wearing ties.

6. In most accounts of explanation, a *potential* explanation is a sentence that, if true, would explain a given phenomenon regardless of whether anyone is aware of it or not. In my account, a potential explanation can only become a bona fide explanation if an agent (or a group of agents) judges it to be true. This difference reflects my contention that explanation is essentially an epistemological notion.

ϕ that the agent judges to be true according to selection criteria that will be specified in the next chapter. An agent knows the best (subjective/objective) explanation of ϕ in K available at time t if and only if (i) Γ is a proper subset of the agent's belief set K at t; (ii) the agent fully believes and fully recognizes all the elements of Γ at t; and (iii) the agent fully recognizes that Γ is the best explanation of ϕ in K. These conditions cannot be used to define the notion of understanding because Γ only includes the elements that the agent or his peers have been able to identify, and these might be insufficient to provide a complete understanding of ϕ. Nonetheless, if a potential explanation is judged to be true and accepted in K, it will provide some degree of understanding of the explanandum ϕ.

The notion of explanation used in this definition is left open for now, but I believe that the analysis provided in this section is compatible with almost any definition of explanation in the philosophical literature. Thus Γ can be seen, for example, as a set of D-N explanations, as a set of causal statements, or as a set of probabilistic laws with their corresponding class inclusion statements.

I will consider four main epistemic contexts in which an agent's failure to understand ϕ impels him to look for an explanation of ϕ. I will introduce several simplifying assumptions that will be indicated in the footnotes. It is assumed throughout that the agent's assessment of the risk of error incurred and of the informational value obtained in accepting a potential explanation ψ in Γ has been preceded by a suppositional contraction of K by ψ, so the assessment is not

question-begging, or by $\sim\psi$, so the assessment is not biased against the potential explanation.

Case 1

$K \cup \{\phi\}$ is consistent and the agent neither antecedently accepts nor rejects all of the elements of Γ. We have two subcases:

A. If $K \cup \Gamma$ is consistent, the agent must *expand* his epistemic state by those elements of Γ that are not elements of K.

B. If $K \cup \Gamma$ is inconsistent, the agent must *revise* his epistemic state by those elements of Γ that make $K \cup \Gamma$ inconsistent.[7]

Case 2

$K \cup \{\phi\}$ is consistent, but the agent antecedently rejects all of the elements of Γ. The agent must *revise* his epistemic state by all the elements of Γ in order to make $K \cup \Gamma$ consistent.

Case 3

$K \cup \{\phi\}$ is consistent and the agent antecedently accepts and recognizes all the elements of Γ, but he fails to recognize that they explain ϕ. The agent must fully recognize that the elements of Γ explain ϕ.

---

7. To simplify the analysis, I have not divided cases 1A and 1B into two subcases each to reflect the fact that the elements of Γ that are antecedently accepted or rejected can be explicitly or implicitly believed. If the agent does not fully recognize the elements of Γ that are antecedently accepted or rejected in K, I will require that the agent must live up to his doxastic commitment before he can understand φ. The same goes for Cases 2 and 3 below.

Case 4

$K \cup \{\phi\}$ is inconsistent.[8] The agent must explain the anomaly by first *revising* his epistemic state by ϕ, and then proceeding as in Cases 1, 2, or 3.[9]

There are several important differences and similarities between the four cases. I will begin by considering the four epistemic contexts from a psychological perspective. From the point of view of the agent, cases 1, 2, and 3 are very similar. They describe a series of situations in which the agent lacks the degree of understanding of ϕ that his cognitive interests and goals demand. They range from the case in which the agent fails to understand ϕ completely because he lacks the required information (the extreme form of case 1A) or because he fails to identify, among the beliefs that he accepts and fully recognizes, any set of beliefs that would provide any explanation of ϕ (case 3), to cases in which only some small but crucial elements of Γ are missing (case 1A) or are disbelieved (case 1B). In the extreme case, the explanation just does not fit with the agent's state of belief at all

---

8. As in the previous cases, if the agent does not fully recognize the contradiction, he must first live up to his doxastic commitment by fully recognizing that the explanandum is inconsistent with his present state of belief.

9. Levi (1991) and Hansson (1991) argue that when an epistemic input conflicts with the information that the agent currently accepts, it is by no means obvious that the conflicting input must be accepted and the belief set adjusted to reflect its acceptance, as the AGM postulates for revision stipulate. The conflicting input might be rejected instead. It would seem convenient, therefore, to give the agent the option of rejecting the explanandum φ in order to restore consistency. Nonetheless, rejecting φ would only explain φ away, but it would not explain it in the proper sense of the term. I am only interested in cases in which the agent decides to accept the explanandum despite the fact that it contradicts his present beliefs. In order to explain the phenomenon, however, consistency must be restored by revising K by φ. I return to this problem later on in this section.

(case 2). In most cases the agent will not be able to tell whether his failure to understand ϕ is caused by the lack of information, by the shortcomings in his logical or mathematical training, by a failure of memory, or by emotional distress. For the same reason, if he learns or discovers the information required to explain ϕ, and fully recognizes that the information explains ϕ, he might still not be able to tell whether the information was already implied by his beliefs, whether he did not know it, or whether he had forgotten it. In some cases, having the elements of Γ in front of him will refresh his memory, and in others it will allow him to see that the information was implied by the rest of his beliefs. But in all cases, the agent is able to explain ϕ only because all the elements of Γ are added to the set of beliefs that he fully accepts and recognizes, and because their explanatory relation to ϕ is made explicit.

Case 4 is, psychologically speaking, somewhat different from the others because the agent begins by recognizing that there is a conflict between the explanandum and the rest of his beliefs. Since the agent has decided not to explain away the anomaly, the compulsion to explain ϕ is more strongly felt than in the other cases. Furthermore, the agent's revision of his belief state after identifying the source of the contradiction will usually serve as a guide in his search for an explanation. Thus the reasons for the agent's initial lack of understanding will be more evident to him than in the other cases.

These psychological aspects, interesting as they are, are not the subject of this study. The goal of a philosophical theory of explanation is to prescribe the changes that the agent should make to his belief state in a given epistemic context, regardless of the agent's subjective appreciation of that context.

When we disregard these psychological features, other differences and similarities emerge. An important difference between case 3 and cases 1, 2, and 4, is that in the latter the agent is required to make changes in his doxastic commitments, while in case 3 he is only required to improve his doxastic performance. In case 3, the agent fails to recognize a relation between beliefs that he ought to have recognized in virtue of his commitment to, and full recognition of the elements of Γ. If he contracts by ψ to give $\sim\psi$ a fair hearing, but decides that ψ is true after all, he will not have increased the informational value of his state of belief even though he now recognizes that ψ is a partial explanation of ϕ. He should have done that from the beginning. This case will not play an important role in what follows, but I have made it explicit as a reminder that in order to say that an agent understands a given fact, the agent must fully recognize the explanatory relation between the explanans and the explanandum. It is not enough to possess an explanation of ϕ; one must recognize that it *is* an explanation of ϕ.[10]

---

10. Scriven (1959, p. 461) uses cases taken from the history of science to illustrate that sometimes the derivation of the explanandum from the explanans can present considerable mathematical difficulties. Without such derivation, one cannot say that the phenomenon has been explained.

The most important difference among the epistemic contexts that involve a change in commitment has to do with the decision process involved in the expansion or revision of the agent's state of belief. Although I will discuss the matter more thoroughly in Chapter 4, a few preliminary considerations are in order.

Following Levi, I have adopted the avoidance of error and the acquisition of new valuable information as the two desiderata that guide the changes in an agent's state of full belief. By expanding or revising his belief state in order to explain ϕ, the agent incurs the risk of error. At the same time, he is motivated to expand or revise his beliefs by the potential increase in the informational value of his epistemic state.

As we saw in Chapter 2, the risk of error varies inversely with the credal probability assigned to the relevant potential expansions of K. Since the agent takes his present state of belief as the standard by which the risk of error is assessed, in each of the three epistemic contexts in which such assessments are made the risk of error will be judged differently.

Assessing the risk of error incurred by expanding K by one or more elements of Γ in case 1A is unproblematic because $K \cup \Gamma$ is consistent. Thus after assigning a credal probability distribution to the potential explanations that he is considering, the agent's decision to expand will depend on the explanatory value of the potential expansions. I will consider this case *in extenso* in the next chapter.

But in cases 1B and 2, where the agent explicitly believes that $K \cup \Gamma$ is inconsistent, the situation is entirely different. The credal probability assigned to some or all of the elements of Γ is 0. Thus by expanding K by Γ in cases 1B and 2, the agent is guaranteed to incur error. In order to consider giving up beliefs that he fully accepts, the agent needs a strong epistemic incentive. His decision to revise K will depend on the result of the suppositional comparison of the informational value of his present state of belief with the informational value of the state that would result from revising his present state by one or more elements of Γ. From a decision theoretical point of view, the best option is to refuse to revise K unless the revised state carries more informational value than K. These cases will also be more carefully considered in the next chapter.

Finally, Case 4 involves two separate assessments of risk of error incurred and informational value obtained. The first one occurs when the agent decides to accept the explanandum, and the second one when he decides to accept the explanation. In the first case, by accepting a sentence that is incompatible with his present state of belief, the agent is deliberately incurring error. Despite Gärdenfors's and Levi's insistence that an agent should never deliberately expand into inconsistency, the acceptance of an explanandum inconsistent with the agent's beliefs is a common, and perhaps inevitable phenomenon in the course of inquiry. The cost of rejecting the explanandum may be too high to bear, and a responsible inquirer might decide to live with the problem until he finds a way to explain the anomaly.

When the source of the trouble is identified, the agent can then revise his beliefs and eliminate the inconsistency.

Providing a precise analysis of the decision process involved in deliberately expanding into inconsistency and then contracting back to consistency is beyond the scope of this dissertation. In recent work, Wassermann and Hansson (1999), and Hans Rott (2001) have independently developed the suggestion made by David Lewis (1982), and mentioned in the previous chapter, that the best way to deal with inconsistencies in a belief state is to quarantine the inconsistency from the rest of the beliefs in the belief state, and try to solve the problem locally. Parikh (1999), Chopra & Parikh (2000), and Chopra (2000) follow a different approach in dealing with the same problem. The model for representing belief structures that they propose relies on a notion of partial language splitting and tolerates some amount of inconsistency while retaining classical logic. If an inquiring agent follows one of the methods proposed by these authors, I cannot see why deliberate expansion into inconsistency should not be tolerated when the cost of rejecting the anomalous phenomenon and/or the background assumptions is too high. The problematic cluster of beliefs can always be kept in check, and inquiry can proceed beyond the gates of epistemic hell.[11]

---

11. Levi could argue that tolerating inconsistency as an intermediate step in the search for truth is just another manifestation of the "messianic realism" that he is trying to discredit. But my point is not that inconsistency should be tolerated as part of the search for truth "at the End of Days." My point is that if deliberate expansion into inconsistency is forbidden, inquiry will come to a full stop.

The assessment of risk and informational value in the second stage of case 4, when the agent is trying to decide whether to accept the elements of Γ, is unproblematic. We can assume that the agent has accepted ϕ and that he has already revised his belief state to restore consistency. The second stage of case 4 can thus be treated as an instance of cases 1, 2, or 3 depending on the elements of Γ that the agent already accepts or rejects in the revised belief state.[12]

The analysis presented in this section was intended as an introduction to the problems that I will discuss in the remaining sections of the dissertation. But before I present my own approach to the problem, it will be convenient to take a look at how philosophers, computer scientists, and researchers in artificial intelligence have dealt with the question of explanation in the context of belief revision.

2. Explanation AGM Style

Even though explanations play a central role in shaping an agent's beliefs, there are very few accounts of explanation in the belief revision literature, and the only theory that is philosophically informed is the one offered by Gärdenfors (1988, ch. 8). Before I analyze his approach, I will examine how the concept of expla-

12. If the explanandum is logically entailed by the elements of Γ, that is, if the explanations are deductive-nomological ones, then case 4 cannot be reduced to case 3. The reason is that if Γ was already in K, so was the explanandum via the closure requirement for belief sets, and there never was an inconsistency to begin with.

nation is used by authors in computer science and AI in order to illustrate what a theory of explanation in terms of belief revision should not be.

2.1 *Pagnucco on Abduction*

The terms 'abduction' and 'explanation' are often used by computer scientists and researchers in AI in areas such as diagnostic reasoning, database updates, vision, and text understanding. In dealing with the question of how to compute updates of a logical database, for example, an "abduction" or explanation is used to support the addition of a piece of information ϕ into a database. An update request "insert (ϕ)" can be achieved by finding some formula consistent with the database such that the union of the set of ground facts in the database and the formula yields ϕ as a logical consequence (within the logic programming domain).[13] The formula is generally interpreted as a justification of the new data, a reason to accept ϕ. When used in the solution of diagnostic problems, abduction often becomes a form of causal backtracking. Given a causal network represented by a causal diagram, the purpose of an abduction is to identify a set of disorders whose associated effects explain all the symptoms observed.[14]

---

13. See Kakas, Kowalski, and Toni (1993) for a survey.
14. See, for example, Peng and Reggia (1990) for an account of diagnostic problem solving through abductive inference.

In most cases, the word 'abduction' is used to denote both the process of inference and the result of such an inference, in which case it is used interchangeably with 'explanation', and sometimes with 'justification'.

Pagnucco (1996) offers an account of abduction in terms of belief revision that uses the terms 'abduction' and 'explanation' in a similar vein. He explains the guiding idea of his approach in the following passage:

> Many belief revision frameworks ... aim to solely incorporate the epistemic input and any resulting consequences. However, it is our contention that a more natural and advantageous approach is for the agent[s] to first seek some explanation or justification for the epistemic input in light of their currently held beliefs and to incorporate this explanation together with the epistemic input into their new epistemic state (pp. 4-5).

Pagnucco presents a logic-based notion of abduction that bears some resemblance to the abductive update of databases described above. He defines an abduction or explanation—he uses the two terms interchangeably—for a sentence ϕ with respect to a domain theory K as a sentence ψ that satisfies the following two conditions (p. 79):

(i) $K \cup \{\psi\} \vdash \phi$

(ii) $K \cup \{\psi\}$ is consistent.

In general, there will be many sentences that qualify as abductions or explanations of ϕ. Pagnucco discusses several criteria that can be used to select the best abduction, although in most cases these criteria do not determine a unique choice. The first criterion is minimality: "assume as little as possible in proving a formula ϕ. This expresses the desire to avoid superfluous abductions" (p. 80). Using the consequence relation, Pagnucco introduces a partial order over the set of abductions of ϕ with respect to K. According to the weakness ordering, some abductions will be weaker than ϕ itself. In that case, the result of expanding K by ψ in the light of ϕ will be $Cn(K \cup \{\phi\})$. The differences among those abductions that are weaker than ϕ is effectively obliterated.

The second consideration is that an abduction should not be trivial. An abduction ψ of ϕ with respect to K is trivial iff $\psi \vdash \phi$. The idea is that the abduction should make use of K and not be able to prove the new information on its own.[15] Another way in which abductions may be seen as trivial is in those cases in which $\psi \rightarrow \phi$ is a theorem in K. Using the deduction theorem, $(\psi \rightarrow \phi) \in Cn(K)$ iff $\phi \in Cn(K \cup \{\psi\})$, we can prove that if $(\psi \rightarrow \phi) \in K$, ψ is an abduction of ϕ. "These types of abduction are inherent to the logic in a certain sense and may always be obtained regardless of the domain theory (up to inconsistency)" (pp. 82-83). Pagnucco claims that these trivial abductions can be weeded out using the other selec-

15. This condition is equivalent to the NES (No-Entailment-by-Singular-Sentence) requirement discussed by Achinstein (1983).

tion criteria described here, together with the selection mechanisms associated with the operation of abductive expansion described below.

Finally, Pagnucco argues that it would be desirable to have a criterion to determine degrees of specificity for abduction. Different explanations demand different degrees of specificity. In his view, "abduction can in a sense be viewed as an inference 'backwards' over an implication; from consequent to antecedent. One way to view specificity then, is to treat propositions further 'back' along an implication chain as more specific" (p. 83). He explores several possible ways in which levels of specificity can be formally determined, but none of them lead to a satisfactory definition.

As we saw above, Pagnucco's purpose in defining an abduction or explanation for a formula is to provide an inquiring agent with a reason to accept the new information she has gleaned. Pagnucco defines an operation called *abductive expansion* which captures this idea. An operation of abductive expansion of a belief set K by a formula ϕ adds to K some formula ψ which explains ϕ, that is, a formula ψ which together with K implies ϕ without making the set inconsistent. Pagnucco offers the following definition:

$K^{\oplus}_{\phi}$ is an abductive expansion of K with respect to ϕ iff

$K^{\oplus}_{\phi} = Cn(K \cup \{\psi\})$ for some $\psi \in L$ such that:

(i) $K \cup \{\psi\} \vdash \phi$ and

(ii) Not $K \cup \{\psi\} \vdash \perp$

$K^{\oplus}_{\phi} = K$ if no such ψ exists.

He then introduces the following rationality postulates for abductive expansion:

($K^{\oplus}1$)	$K^{\oplus}_{\phi} = Cn(K^{\oplus}_{\phi})$.	(closure)
($K^{\oplus}2$)	If $\sim\phi \notin K$, then $\phi \in K^{\oplus}_{\phi}$.	(limited success)
($K^{\oplus}3$)	$K \subseteq K^{\oplus}_{\phi}$.	(inclusion)
($K^{\oplus}4$)	If $\sim\phi \in K$, then $K^{\oplus}_{\phi} = K$.	(vacuity)
($K^{\oplus}5$)	If $\sim\phi \notin K$, then $\sim\phi \notin K^{\oplus}_{\phi}$.	(consistency)
($K^{\oplus}6$)	If $K \vdash \phi \leftrightarrow \gamma$, then $K^{\oplus}_{\phi} = K^{\oplus}_{\gamma}$.	(preservation)
($K^{\oplus}7$)	$K^{\oplus}_{\phi} \subseteq Cn(K^{\oplus}_{\phi\vee\gamma} \cup \{\phi\})$.	(supplementary 1)
($K^{\oplus}8$)	If $\sim\phi \notin K^{\oplus}_{\phi\vee\gamma}$, then $K^{\oplus}_{\phi\vee\gamma} \subseteq K^{\oplus}_{\phi}$.	(supplementary 2)

These postulates only impose basic constraints on the operation of abductive expansion. In order to select the best abductive expansions, Pagnucco constructs three selection mechanisms: epistemic entrenchment, partial meet abductive expansion functions, and a construction based on Grove's system of spheres. I will only examine the first two.

Pagnucco's analysis of epistemic entrenchment is based on the notion of an expectations ordering proposed by Gärdenfors and Makinson (1994). An expectations ordering is an epistemic entrenchment ordering that only satisfies conditions (EE1) - (EE3), that is, transitivity, dominance, and conjunctiveness.[16] Maximality

16. See Section 1.3 in the previous chapter.

and minimality are dropped in order to apply the notion of epistemic entrenchment to nonmonotonic reasoning.

Pagnucco defines the notion of an abductive entrenchment ordering by adding a fourth condition to the notion of an expectations ordering:

(AE1) When $K \neq K_{\perp}$, $\phi \in K$ iff $\gamma \leq \phi$ for all $\gamma \in L$. (maximality)

An abductive entrenchment ordering is an expectations ordering in which all of the agent's beliefs are maximally entrenched. This makes it easy to extract the agent's current state of belief from the ordering and consider only the sentences that he does not currently accept. Pagnucco then shows that for any well-behaved abductive expansion function there exists an ordering of abductive epistemic entrenchment that generates the function.

Partial meet abductive expansion functions, on the other hand, are modeled after AGM's partial meet *contraction* functions. The dual of AGM's $K \perp \phi$, the set of all belief sets K' that are maximal subsets of K that fail to imply ϕ, is $K \top \phi$, the set of all belief sets K' that are maximally consistent supersets of K that imply ϕ. Let S be a function that picks out elements from $K \top \phi$. The intended interpretation is that S picks out the members of $K \top \phi$ that rank highest according to some index of explanatory merit. Using the function S, Pagnucco defines a partial meet abductive expansion function thus:

(Def Part $\oplus$) $K^{\oplus}_{\phi} = \bigcap S(K \top \phi)$ whenever $K \top \phi$ is nonempty;
$K^{\oplus}_{\phi} = K$ otherwise

This construction suffers from the same problem that we identified in the case of AGM's partial meet contraction functions: partial meet abductive expansions are not closed under intersection. The intersection of two abductive expansions that rank best according to the selection function need not rank best according to the same function. As Levi justly complains, "there is no effort to show that the meet or intersection is also optimal" (1998, p. 19, n. 14). In contrast, Levi's methods for deliberate expansion do provide a way to choose among the optimal options for expansion as long as the functions representing the risk of error and the informational value of the expansion are determinate.

Pagnucco recognizes that further restrictions can be imposed on these selection mechanisms to make them more precise. For example, if the requirement of minimality is imposed on abductions, the result is a full meet abductive expansion function that turns out to be equivalent to AGM expansion. And if abductions are required to be maximally specific, the result is a maxichoice abductive expansion function. The choice of one restriction over another will depend on nonlogical factors such as the intended use of the abductive operation.

Pagnucco does not offer an epistemological interpretation of the notions of abduction and abductive expansion, but he is aware that his definition of abduction might not capture what he calls "the intuitive notion of explanation." He refers the

reader to Salmon's *Four Decades of Scientific Explanation* (1989) for details about this "intuitive notion," and he mentions that "parallels can be drawn from the discussion of work concerning Hempel and Oppenheim's deductive-nomological model of explanation" (p. 10, n. 8). In the rest of this section I examine whether the accounts of abduction and abductive expansion provided by Pagnucco have anything in common with the notion of explanation as it is understood in the philosophical literature.

The first point that must be clarified is a terminological one. Pagnucco uses the term 'abduction' to denote both an inference to an explanation and the explanation itself, the second use being by far the more frequent. However, Pagnucco's abductions, in the second sense, cannot be considered bona fide explanations. Virtually every account of explanation in the philosophical literature begins with the requirement that both the explanans and the explanandum statements of a bona fide explanation must be true, or must have been accepted as true for the time being. In contrast, in Pagnucco's account neither the sentence ϕ that the agent wants to explain, nor the sentence ψ that, together with K, would entail ϕ, has yet been accepted in K. In consequence, Pagnucco's definition of abduction can only be interpreted as a definition of an element of a *potential* deductive explanation of ϕ in K, that is, as a sentence that would form part of the deductive explanation of ϕ if it were accepted in K.

Potential explanations have important uses. Hempel mentions, for example, that a potential explanation can be used to examine "whether a novel and as yet untested law or theory would provide an explanation for some empirical phenomenon" (1965, p. 338). Harman (1965), on the other hand, argues that the explanatory power of a new theory is part of the evidence that leads us to accept it. But Pagnucco's potential explanations can be put to no such use. As Hempel's comment makes clear, in order to test the explanatory potential of a theory, one must have antecedently accepted the facts that would be explained by the theory. Pagnucco's abductions work the other way around: one must first accept one of the potential explanations of ϕ in order to accept ϕ. There is no independent way of assessing the explanatory value of ψ because its status as a potential explanation is tied to a sentence ϕ about whose truth the agent is still undecided.

Now suppose that the agent expands her belief set by ϕ using the operation of abductive expansion. It seems to me that the explanation used to support the acceptance of ϕ cannot itself be accepted without defeating the whole purpose of Pagnucco's approach. The type of expansion operation that he defines is intended to provide support for every new sentence that is added to K. The problem is that, although ϕ in $K^{\oplus}_{\psi}$ is supported by ψ, the latter sentence is added to K in an abductive expansion without itself having any support other than the fact that it is part of the best explanation of ϕ. In other words, for every justified formula ϕ that the agent accepts, she must accept an unjustified formula ψ whose only credentials are

that it is an element of the intersection of the highest ranking maximally consistent supersets of K that imply ϕ. In fact, the agent might not even know which of the potential explanations of ϕ is the one that ends up in her belief set, as Pagnucco himself acknowledges:

> In abductive expansion, as we have seen, the concern is to determine which beliefs should be incorporated into the current epistemic state using an abductive strategy to identify the appropriate expansion given new information. In so doing however, the process of abduction has become "internalized" in the belief expansion process and therefore, the actual abduction(s) made to effect a change in epistemic state is, in a sense, lost. That is, it may not be possible to determine the abduction selected for a belief set K and epistemic input ϕ in the sense of [the definition of abduction provided above] (i.e., it may not be possible to identify ψ) (p. 135).

Pagnucco argues that once ϕ has been accepted, it will be possible to find an abduction "capable of doing the job." The strategy is to examine the set $K_\phi^\oplus \setminus K_\phi^+$ and determine a finite axiomatization of it. "The conjunction of the elements of this finite axiomatization will suffice as an appropriate abduction" (p.135). There might be many ways to finitely axiomatize the set, but Pagnucco argues that restrictions such as minimality and the like can be used to select the best abduction.

Unfortunately, this strategy does not answer my initial objection. Not only is the agent accepting an unexplained sentence ψ every time she accepts a sentence ϕ

via an abductive expansion. If the agent has to examine $K^{\oplus}_{\phi}$ in order to find out which of the potential explanations of ϕ was used in the abductive expansion, it is absurd to say that the agent used that potential explanation to *justify* or support the expansion of his belief set into $K^{\oplus}_{\phi}$. Although Pagnucco wants to offer an account of expansion in which the agent "first seek[s] some explanation or justification for the epistemic input," the abductive expansion function that he defines does not reflect his declared intention.

Finally, the claim that the conjunction of the elements of the finite axiomatization of $K^{\oplus}_{\phi} \setminus K^{+}_{\phi}$ suffices as an appropriate explanation of ϕ can also be challenged. Pagnucco's account is vulnerable to the same counterexamples that have been raised against approaches, such as Hempel's, in which explanation is understood as an inferential relation. Consider Bromberger's well-known flagpole example. Let ϕ = 'The flagpole in front of the building is 10 meters tall' and ψ = 'The length of the flagpole's shadow is 10 meters and the elevation of the sun in the sky is 45 degrees'. Suppose that the agent wants to add ϕ to his belief state. Using ψ and his mathematical knowledge, the agent can conclude that the flagpole is 10 meters tall. Thus if we examine the conjunction of the elements of the finite axiomatization of $K^{\oplus}_{\phi} \setminus K^{+}_{\phi}$, we will find ψ. So ψ is an appropriate abduction of ϕ according to Pagnucco, but clearly it does not explain why the flagpole is 10 me-

ters tall. It does, however, completely justify the addition of ϕ to the agent's belief set.

Despite the fact that many researchers in AI have understood the notion of abduction as "inference to the best explanation," the notion of explanation involved has little in common with the notion as it is understood in the philosophical literature. Pagnucco's abductions are better understood as justifications, reasons to believe something. A reason is not the same as an explanation. To use another famous counterexample, an agent can have very powerful reasons to believe *that* a storm is brewing without having a clue as to *why* it is brewing. And his request for a justification of the *claim* that a storm is brewing is very different from his request for an explanation of the *fact* that a storm is brewing. Although Pagnucco's approach to abduction is logically impeccable, it does not provide an adequate framework for the formulation of an epistemologically motivated notion of explanation.

2.2 *Gärdenfors on Explanation*

In *Knowledge in Flux* (1988), Gärdenfors offers an account of explanation using the framework for belief revision presented in that book. The result is a highly complex formal theory based on a very simple idea. Most of my objections will be directed against the latter, so we will not need to worry too much about the technical details of his approach.

Unlike Pagnucco, Gärdenfors takes it for granted that if an agent wants to explain the fact stated by ϕ, the agent already believes that ϕ is true. An explanation does not determine whether a person accepts ϕ or not. Nonetheless, Gärdenfors argues that an explanation does have an effect on the way we believe that ϕ: "it is quite clear that the fact that E may be more or less *surprising* or *unexpected*, and the principal effect of a successful explanans is that the surprise at E is *decreased*" (p. 167).

To make these notions more precise, Gärdenfors has to introduce a probabilistic model of epistemic states using a first-order language, instead of the propositional language used in AGM. An epistemic state K suitable for the formulation of Gärdenfors's theory consists of (i) a set W of possible worlds, (ii) for each world w a probability measure P_W defined over sets of individuals in w, and (iii) a belief function B that measures the probability of sets of possible worlds. The measure P_W together with B yield a second-order probability distribution of properties, and using this distribution Gärdenfors defines a first-order expected probability measure.

Gärdenfors's strategy is to take as a basis for the analysis, not the agent's present belief state K, but the state K_ϕ^- in which the explanandum sentence has been contracted from K. Once the explanandum sentence is deleted, it is possible to measure how surprising it would be to find out that it is true. "The surprise value of ϕ is inversely related to the degree of belief associated with ϕ in K_ϕ^-. The

central criterion on explanations is that the explanans in a nontrivial way should decrease the surprise value of the explanandum" (p. 168). In other words, the explanans should make ϕ more believable in K_{ϕ}^{-}. These considerations lead to the following necessary conditions for explanation (p. 178):

(EXP) An explanation of a singular sentence ϕ relative to a state of belief K (where $\phi \in K$) consists of (i) a conjunction T of a finite set of probability sentences and (ii) a conjunction C of a finite set of singular sentences that satisfy the requirements that (iii) $B_{\phi}^{-}(\phi / T \ \& \ C) > B_{\phi}^{-}(\phi)$, where B_{ϕ}^{-} is the belief function in the state K_{ϕ}^{-}, and (iv) $B(T \ \& \ C) < 1$ (that is, $T \ \& \ C \notin K$).

The definition amounts to a counterfactual analysis of an epistemic context in which the agent accepts neither the explanans nor the explanandum. The purpose of introducing the additional context is to determine what information should be accepted in K in order to make the acceptance of ϕ in K less surprising than it would be in the absence of the additional information. Before we examine whether this epistemic context—which I did not consider in the first section of this chapter—contains the key to explanation, we must complete the exposition of the account.

Gärdenfors argues that a consequence of his definition is that there will be degrees of explanation. "The more an explanation increases the belief value of the explanandum, the better it is" (p. 185). From condition (iii), Gärdenfors obtains a

measure of the explanatory power of an explanans. The greater the difference between $B_{\phi}^{-}(\phi / T \& C)$ and $B_{\phi}^{-}(\phi)$, the greater the explanatory power of $T \& C$ relative to ϕ. Since (EXP) allows many different explanations of ϕ, Gärdenfors argues that the explanatory power of $T \& C$ can be used to determine which of all the possible explanations is the best one. Deductive explanations turn out to be the best because they increase the belief value of the explanandum to the maximum value.

Gärdenfors contrasts his approach to Hempel's view according to which the explanans of an inductive-statistical (I-S) explanation shows that the phenomenon described by the explanandum sentence was to be *expected*. The demand that the explanans should make the explanandum *less surprising* is an analogous but weaker claim which Gärdenfors believes is immune to the objections that have been raised against the I-S model. An examination of these objections will allow us to test the adequacy of Gärdenfors's definition.

In *Aspects of Scientific Explanation*, Hempel characterized statistical explanations as inductive inferences or arguments in which "the explanans confers upon the explanandum a more or less high degree of inductive support or of logical (inductive) probability" (1965, p. 385). The probability associated with the explanation determines the strength of our expectations. Since Hempel required that the probability associated with an I-S explanation be fairly close to 1, the explanandum of an inductive-statistical explanation will always be expected "with 'practical' certainty, or with very high likelihood" (p. 389).

In response to criticism by Richard Jeffrey, in the first German edition of *Aspects* Hempel ([1977] 2001) gave up the high probability requirement, together with the claim that the explanans of an I-S explanation should show that the phenomenon described by the explanandum sentence was to be expected.

Jeffrey (1971) presented two objections to the requirement that the probability associated with an I-S explanation be fairly close to 1. In the first place, Jeffrey argued that it is erroneous to set limits to the concept of explanation. We cannot exclude the possibility of explaining any phenomenon, regardless of the probability associated with it. Peter Railton makes the same point: "Virtually impossible events may occur, and they deserve and can receive the same explanation as the merely improbable or the virtually certain" (1978, p. 213).[17]

Secondly, Jeffrey argued that when we try to explain why the explanandum of an I-S explanation did not obtain, we use the same statistical laws and the same initial conditions that we would use in explaining its occurrence. In Jeffrey's view, an explanation of either the occurrence or the nonoccurrence of a statistical phenomenon "consists of a statement that the process is a stochastic one, following such-and-such a law. ... The knowledge that the process was random answers the question, 'Why?'—the answer is, 'By chance' " (p. 24). Understanding the out-

17. This comment was actually a criticism of Jeffrey, who had argued that the I-S model was unobjectionable only in the "beautiful cases" in which the probability is so high "as to make no odds in any gamble or deliberation" (p. 27). Railton argued that the "beautiful cases" were logically identical to the rest.

comes of a stochastic process does not involve a justificatory argument as to why a given outcome obtained; it requires an adequate description of the process involved.

Despite Gärdenfors's allegations to the contrary, Jeffrey's objections can be reformulated in the context of his own theory. I will argue that Gärdenfors's account erroneously limits the scope of the notion of explanation, and that it leads to an excessive relativization of the concept. I will begin with the latter objection.

In his discussion of Scriven's famous paresis example, Gärdenfors argues that if an agent wants to know why, of all people, Nietzsche developed paresis, an acceptable explanation is that he suffered from syphilis and that there is a low but nonvanishing probability that a syphilitic patient will develop paresis. So far, Gärdenfors seems to agree with Jeffrey and Railton. However, if the agent wants to know why, of all syphilitics, Nietzsche developed paresis, Gärdenfors argues that there is no explanation. There is no further factor that would make the explanandum less surprising. A more accurate way of describing the situation would be to say that the agent already knows the explanation, namely, that there is a low but nonvanishing probability that a syphilitic patient will develop paresis, not that there is no explanation. Gärdenfors's account does not allow this formulation because as soon as the explanans becomes part of the agent's state of belief, it loses all its explanatory power. In Gärdenfors's view, a set of sentences is an explanation in some epistemic contexts, but not in others. In that regard, his account

closely resembles van Fraassen's theory of explanation, as Gärdenfors himself acknowledges.[18] I argued in Chapter 1 that this extreme form of epistemic relativity is untenable. In the account of explanation that I will defend, whether a set of sentences is a potential explanation will not depend on whether the agent believes that the sentences in the set are true or not. The assessment of how *valuable* the explanation is, however, will depend on the agent and on the epistemic context.

We now turn to Jeffrey's first objection. Gärdenfors argues that characterizing a successful explanans as one that increases our degree of belief in a proposition does not preclude the explanation of facts that are familiar or not surprising at all because when we ask for an explanation of a well-known fact, "we in a sense pretend that ϕ is surprising" (p. 167). How does one pretend that something is surprising? Contracting one's belief state by the explanandum sentence will not do. If the belief value of ϕ in K_{ϕ}^{-} is very close to 1, it is difficult to see why an agent would want an explanation of ϕ if an explanation is conceived as information that raises the credal probability of ϕ. What additional information could possibly make ϕ less surprising?

Perhaps Gärdenfors could argue that there is a certain threshold beyond which explanations are no longer required. But the following example should dispel that idea. Consider the case of an old sailor who wants to know why the tide

18. "The theory of explanation that comes closest to the present one is van Fraassen's" (p. 170).

rose today at 6 p.m. According to Gärdenfors, the sailor should engage in an exercise of hypothetical belief revision and contract his epistemic state by the belief that the tide rose today at 6 p.m. Because of his many years of experience observing the correlation between the regular ebb and flow of the tides and the position and phase of the moon, the credal probability that he assigns to the belief that the tide rose today at 6 p.m. in the contracted state is very close to 1. That fact, however, cannot preclude him from wondering why the tide rose today at 6 p.m. Furthermore, an explanation in terms of Newton's law of gravitation will not change the sailor's near certainty that the tide rose today at 6 p.m. It will, however, help the sailor *understand* why the tide rose today at 6 p.m.

The problem is not limited to cases in which the initial degree of belief is very close to 1. Suppose a tourist visiting the Sahara is quite surprised to see a storm approaching since it is her belief that it never rains in the Sahara. If the tourist had been told beforehand that the barometer was falling, and that whenever the barometer is falling a storm is approaching, the tourist's credal probability that there would be a storm would have been very close to 1. And yet, the falling barometer does not explain why a storm is approaching. In fact, we can say that even though the tourist would no longer be surprised by the occurrence of the storm, she would still be *puzzled* in a sense that has nothing to do with the credal probability that she assigns to the occurrence of the storm. Her intellectual curiosity would not

be satisfied, and she would demand an explanation. Her epistemic situation would be similar to that of the old sailor who wanted an explanation of the tides.

The source of the problem has long been identified by philosophers of science: "some regularities have explanatory power, while others constitute precisely the kinds of natural phenomena that demand explanation" (Salmon, 1984, p. 121). The regularities captured by the probability sentences in Gärdenfors's definition will often have no explanatory value even if they make the explanandum completely unsurprising. Without further restrictions on the explanans, the account will always be vulnerable to such counterexamples.

Gärdenfors argues that one should choose the explanans T & C that has the highest degree of explanatory power, but this demand does not solve the problem. It is difficult to imagine a set of sentences that would raise the tourist's credal probability more than the information regarding the falling barometer. Sets of sentences that are maximally explanatory in Gärdenfors's sense will often be useful for prediction and practical deliberation, but they will fail to provide understanding.

A deeper problem with Gärdenfors's account has to do with the notion of surprise. If we assume that the prior degree of belief in a sentence can be used to measure how surprising it would be to find out that it is true, Gärdenfors's account will often be inapplicable because the initial credal probability will be indeterminate. What is my degree of belief that it rained today in the Azores? Or that the

economy of Bolivia grew more than 1% in 1987? It would be a mistake to assign any value to my degree of belief because I have no elements whatsoever to judge the issue one way or the other. The rational attitude, it seems to me, is to suspend judgment.

Levi (1988) proposes a much more fruitful way to think of the notion of surprise: "The truth of h is not surprising relative to a body of information if and only if the acceptance of $\sim h$ via inductive inference from that information is not legitimate. The truth of h is to be expected relative to that body of information if and only if its inductive acceptance is legitimate" (p. 207). An inductive inference in Levi's sense is a deliberate expansion of the agent's state of belief that seeks the best trade off between the informational value obtained and the error incurred.

According to these definitions, it would not be surprising for me to find out that it rained today in the Azores, or that the economy of Bolivia grew more than 1% in 1987, because it would have been illegitimate for me to infer the opposite. But the truth of these sentences would not be expected either. If my discovery that a sentence is true is neither surprising nor to be expected, it is because I had not judged the issue one way or the other. But if no credal probability can be assigned to the explanandum in B_{ϕ}^{-}, then Gärdenfors's account of explanation will be inapplicable in the vast majority of cases. In general, if the agent has no prior informa-

tion about the explanandum, it will be meaningless to say that the potential explanans should make the explanandum less surprising.[19]

To be sure, the decision to accept a potential explanans T & C will be affected by my previous beliefs, but only in the sense that they will affect the way that I assess the informational value of the explanans and the risk of error incurred by accepting it. In my account, however, that assessment will be based on the actual belief state that results from the acceptance of the explanandum, and not on the state K^-_ϕ that results from counterfactually deleting the explanandum.

Like Pagnucco's analysis of abduction, Gärdenfors's theory of explanation is better understood as an account of justification. The similarities between the two accounts are not difficult to see. Both accounts begin with a belief state in which the explanandum is absent, either because it has not yet been accepted, or because it has been counterfactually deleted. In both cases, an explanation lends support to the acceptance of ϕ in K (or in K^-_ϕ). In Pagnucco's case, an explanation, together with K, entails ϕ, thereby leading to a new state $K^{\oplus}_{\phi}$; in Gärdenfors's case, an explanation raises the credal probability of ϕ in K^-_ϕ, thus making it more believable in K. Both Pagnucco and Gärdenfors thus identify justificatory or evidential information with explanatory information.

---

19. This objection is not explicitly stated in Levi (1988), but it follows from the definitions quoted above. Levi (personal communication) agrees with the conclusion of my analysis.

In order to offer an adequate account of explanation, we must abandon the idea that *any* information that justifies or serves as evidence for the explanandum also explains it. This idea was introduced by Hempel and Oppenheim (1948) as part of their thesis that there is a logical symmetry between prediction and explanation, and although Hempel never explicitly rejected that part of the symmetry thesis, he later declared it "open to question" (1965, p. 367). Numerous counterexamples, together with the rejection of the high probability requirement for inductive-statistical explanations, suggest that it is simply false.

3. Explanation: The Basic Idea

The account of explanation that I will present in the following chapter is based on a very basic intuition about the notions of explanation and understanding. To understand a phenomenon, I submit, is largely a matter of knowing how it fits into the cognitive system formed by the agent's doxastic commitments, and her cognitive interests and goals. An explanation consists in the information required to integrate the phenomenon into that cognitive system. The challenge is to transform this basic intuition into an account that is philosophically motivated and logically precise. In this section I introduce my approach in an informal way; the complete account is developed in the next chapter.

An inquiring agent has no doubt that all the sentences in her corpus of beliefs are true. Nonetheless, she does not regard all of the facts stated by these

sentences as being equally well-understood. The degree to which an agent understands the fact expressed by a sentence ϕ will depend on how well-integrated ϕ is to the agent's cognitive system. It will not depend on how much support it has or on how epistemically entrenched it is. On the one hand, if a sentence has been accepted in K, it is judged to be true and no further argument is necessary.[20] On the other hand, poorly understood phenomena can be highly epistemically entrenched, and completely useless facts can be very well-understood. Even though the explanatory value of a sentence is one of the main factors that determines its position in an ordering of epistemic entrenchment, this does not entail that all explanatory facts are themselves well-understood. Logic alone is clearly not sufficient to describe an inquiring agent's system of commitments. Many of our most important beliefs about the world are stated in terms of probabilities, and these beliefs must be included in assessing the degree to which an agent understands a given fact. The propositional language used in the AGM framework will thus be replaced in the next chapter by a language that includes a probability sentence for every combination of singular sentences. The probability sentences in K will express the agent's commitment to objective relevance relations between facts.

---

20. It is, of course, possible to understand a state of affairs that one does not fully believe. In such cases the agent supposes, for the sake of argument, that the sentence describing the state of affairs is true, and adjusts his beliefs accordingly. For a discussion of suppositional reasoning, see Levi (1996).

With the aid of probability sentences it will be possible to offer a more precise characterization of the idea that understanding a phenomenon is a matter of knowing how it fits into the agent's system of doxastic commitments. In order to understand the fact that ϕ, where ϕ is a predicate P followed by an individual constant a, the agent must add to her state of belief sentences that capture the idea that there are facts that contribute to make ϕ true, and facts that act against it. Without such information, ϕ will describe a brute fact, isolated from the rest of the agent's beliefs about the world.

Probability sentences are the connecting tissue of an agent's corpus of beliefs. They make evident which facts lowered ϕ's chance of not being true, and which facts raised its chance of being false. The type of probability sentences that we are interested in are not of the form $p(\phi / \psi) = r$, where $r \neq 0$. This type of sentence will not tell the agent what influence ψ has on ϕ. Would r be higher if ψ were not true? Would the absence of ψ make any difference at all? Indeed, in Chapter 4 I will argue that specific probability values have no explanatory value, only descriptive, predictive, and evidential value. The probability sentences that the agent should look for will have the form $p(\phi / \psi) > p(\phi / \sim\psi)$ and $p(\phi / \psi) < p(\phi / \sim\psi)$. These are the sentences that allow the agent to determine the factors that positively or negatively affect ϕ's objective chances of being true.[21]

---

21. There are other probabilistic accounts of explanation in the philosophical literature. The differences between these accounts and my approach will become evident in the next chapter.

I will argue that in order to understand ϕ, the agent must expand (or revise) her state of belief by adding a subset of all singular sentences that state facts that raise or lower the probability of ϕ, together with the corresponding subset of all probability sentences that capture the statistical relevance relations. The union of the two subsets constitutes the explanation of ϕ in K.

If the acceptance of the explanation of ϕ is an isolated decision in which the agent obeys no judgment but his own, the explanation will lack any sort of objectivity. My main purpose is to define a type of explanation that is objective by any reasonable standard of objectivity. Thus the decision process that I will be most interested in is the one that takes place when the agent becomes part of a learning community.

The failure to understand a phenomenon is more likely to occur when it is accepted as a result of what Levi calls a *routine* expansion, that is, an expansion in response to external stimuli such as observation, or the testimony of witnesses or experts. If the agent is immediately unable to place the new information within the network of statistical relevance relations, the newly accepted sentence will state a brute fact until an explanation for it is found. Deliberate expansion is less likely to result in the acceptance of brute facts, mostly because the decision to expand is done against a background of accepted beliefs, some of which might be probability sentences that have the effect of raising or lowering the credal probability assigned

to the potential expansion, and consequently, raising or lowering the agent's assessment of the risk incurred in expanding.

The kernel of truth in Gärdenfors's account of explanation is that the probability sentences that counterfactually raise the credal probability of the explanandum may also result in the acceptance of an explanandum that is better understood. Since the only condition that Gärdenfors imposes on probability sentences is that they should make the explanandum less surprising, there is no guarantee that the probability sentences will not turn out to be spurious correlations. But if we restrict the probability sentences to those that describe genuine relevance relations, a probability sentence that makes the explanandum less surprising will also provide the agent with some degree of understanding. However, there is no guarantee that the level of understanding achieved using such probability sentences will be sufficient to satisfy the agent's cognitive interests and goals.

To conclude, an explanation of ϕ in K will be characterized as the union of a set of singular sentences that state one or more facts that raise or lower the probability of ϕ, and a set of probability sentences that state the corresponding relations of probabilistic relevance. Whether the assessment of the risk incurred and of the informational value obtained in accepting the explanation is made by an individual or by a group of experts will make an enormous difference for the objectivity of the explanation. The purpose of Chapter 4 is to provide a more precise characterization of the ideas presented in this section.

CHAPTER 4
A PRAGMATIC ACCOUNT OF EXPLANATION

In his many writings on belief revision, Isaac Levi has insisted that any analysis of the decisions that a rational agent ought to make in the course of inquiry should take into account the difference between information and informational value. The distinction is captured by the principle of *weak* monotonicity for informational value that he advocates: If an epistemic state K carries more information than an epistemic state K^*, it carries at least as much informational value. In contrast, Alchourrón, Gärdenfors, and Makinson have adopted a *strong* monotonicity principle: If an epistemic state K carries more information than an epistemic state K^*, it carries more informational value.

Levi's main argument for the distinction between information and informational value is that when an agent seeks to expand her beliefs, her interest is restricted to information that promotes her cognitive goals or that is relevant to the problems that she is trying to solve. "The inquirer is not interested just in gratifying her curiosity by strengthening her doctrine, but in doing so in certain respects" (1991, p. 82). Thus some information will be valuable and other will be useless. Or as Catherine Elgin puts it, "truth does not always enhance understanding. An irrelevant truth is epistemically inert. ... We have no reason to credit it; we can make nothing of it" (1996, p. 124).

The account of explanation that I present in this chapter has been conceived with the distinction between information and informational value in mind. I will argue that the goal of an inquiring agent is not to find explanations *simpliciter*; it is to find epistemically valuable explanations. This idea is captured by the three theses that I will defend:

1. Whether a piece of information is a potential explanation of ϕ is entirely a nonpragmatic matter.

2. The epistemic value of a potential explanation of ϕ is mostly a pragmatic matter.

3. In trying to understand ϕ, an inquiring agent should accept the potential explanation(s) of ϕ with the highest epistemic value.

I believe that the combination of these three theses constitutes a compromise between the two opposite approaches to the pragmatics of explanation analyzed in Chapter 1. It incorporates the insights that make each of them compelling individually, but it eliminates most of their shortcomings. The account preserves the idea that there must be an objective basis to an explanation, but it regards the objective basis as an incomplete structure. It recognizes that "explanation is an interest-relative notion" (Putnam, 1978, p. 41), but it avoids the problems associated with defining the context of an explanatory speech act. Whether such a compromise will please either party, I am not prepared to say.

The chapter has two sections. I begin by discussing the type of probability sentences that must be used in an explanation. Unlike other authors, I take the discovery of relevance relations as an end in itself, and not as a means to determine the exact value of the probability that the explanans confers upon the explanandum. Defending this claim will require a discussion of the notion of statistical explanation, and of the alleged difference between objective and epistemically relativized statistical explanations. The result of the discussion will lead to the definition of a potential explanation. I then introduce the notion of an explanation space, which is the set of all potential explanations of a given fact. This completes the analysis of the purely nonpragmatic aspects of explanation. In Section 2, I analyze how the credibility, the informational content, and the epistemic value of an explanation should be understood. I will argue that although an objective measure of the credibility and content of an explanation can be established, a complete determination of its epistemic value will ultimately depend on the interests and goals of individual inquirers. With this analysis in hand, it will be possible to offer a definition of an explanation of ϕ in K.

1. The Objective Basis of Explanation

In this section I defend the first of the three theses stated above, namely, that determining the potential explanations of a given fact is entirely a nonpragmatic matter. Before we begin, it will be convenient to remind the reader of the main idea behind

my approach to explanation. My basic contention is that an explanation should provide the information required to integrate the explanandum into the agent's cognitive system. An explanation should provide some of the factors that contributed to make ϕ a fact, and some of the obstacles that could have, but did not prevent it from being one. The influence of these factors is captured by probability sentences of the form $p(\phi / \psi) > p(\phi / \sim\psi)$ and $p(\phi / \psi) < p(\phi / \sim\psi)$ that indicate that the fact that ψ is statistically relevant to the explanandum.

The following technical clarifications are necessary before we can take a closer look at the content and form of these probability sentences. I will use the same object language that Gärdenfors used in his account of explanation, that is, a first-order language simplified as follows. The language only contains singular sentences and probability sentences. Atomic sentences are built from predicates P, Q, R, etc., and individual constants a, b, c, etc. Singular sentences are defined as atomic sentences or truth-functional combinations of atomic sentences, and denoted ϕ, ψ, η, etc. Probability sentences are of the form $p(A / B) = r$, where A and B are predicates followed by a constant, or truth functional combinations of such expressions, and r is a number between 0 and 1. Singular sentences represent states of affairs, and true singular sentences state *facts*. I take facts to be the relata of explanation.

1.1 *Statistical Relevance and Probability Values*

The notion of statistical or probabilistic relevance has been used by many authors in the analysis of explanation. Besides Hempel's I-S model, the best known examples are Salmon's (1971, 1984) S-R model, Railton's (1978, 1980) D-N-P model, and Fetzer's (1974, 1981) causal-relevance model. My account is different from theirs not only because the pragmatic aspects of explanation play central stage. The main difference lies in the fact that they consider probability values to be an essential part of an explanation. In contrast, I will argue that reference to probability values is largely unnecessary.[1]

Probability values are thought to be important for two different reasons. If a statistical explanation is conceived of as an inductive argument, as it was in Hempel's original I-S model, the degree of expectation that a body of evidence confers upon a given event must be very high.[2] Thus the value of the inductive probability must be kept in check to make sure it does not go below a certain threshold as inquiry proceeds. On the other hand, if a statistical explanation is understood as an objective account of the stochastic process involved, as it is in Salmon's and Rail-

---

1. Paul Humphreys (1989) offers an account of probabilistic causality that makes no use of probability values. The notion of causality thus defined is then used to provide an account of causal statistical explanation. I examine his argument below.
2. Hempel never says how high it should be. He only says that it must be "very high" and that the conclusion should be expected "with very high likelihood" (1965, p. 389).

ton's models, it is crucial to avoid the attribution of false probability values to the probabilistic laws.

We have already seen that in response to criticism by Jeffrey, Hempel gave up the high probability requirement, together with the claim that the explanans of an I-S explanation should show that the phenomenon described by the explanandum sentence was to be expected.[3] Without this claim, however, the first reason to attribute any importance to probability values disappears. If the explanans is not supposed to justify our expectations that the explanandum will occur, there is no need to make sure that the value of the probability remains over a certain threshold.

Before we can evaluate the second reason why probability values are deemed to be explanatory, we must take a closer look at the logical structure of statistical explanations. One of the features of probability theory is that it does not have a weakening principle. In deductive logic, a truth-functional argument will remain sound if additional true premises are inserted. But a sound inductive argument that strongly supports its conclusion can be transformed into one that strongly undermines its conclusion with the insertion of additional true premises. Consider the following example. If I know that Mary has a streptococcal infection and was treated with penicillin, I will be confident that she will recover. But if I

---

3. See Chapter 3, Section 2.2.

find out that Mary is almost a hundred years old and has a weak heart, the evidence will support the opposite conclusion. An individual event can be referred to different reference classes, and the probability of the property associated with the event can vary considerably from one class to another. Hence, a body of evidence may confer a high degree of expectation upon a given event, while another body of evidence may confer a very low degree of expectation upon the same event. This is the problem that Hempel called the *ambiguity* of I-S explanation.

Hempel's partial solution to the problem is the requirement of maximal specificity. The requirement states that an acceptable statistical explanation should be based "on a statistical probability statement pertaining to the narrowest reference class of which, according to our total information, the particular occurrence under consideration is a member" (1965, p. 398). The requirement does not completely eliminate the ambiguity because the narrowest reference class can only be determined in the light of our current knowledge. It does not guarantee that there are no unknown statistical generalizations that can be used to construct a rival argument. In fact, Hempel claimed that "*the concept of statistical explanation for particular events is essentially relative to a given knowledge situation as represented by a set K of accepted sentences*" (p. 402, emphasis kept). I return to this claim below.

Salmon (1971) rejected Hempel's solution to the problem because he strongly believed that the epistemic relativity of statistical explanations was unacceptable, and because the requirement of maximal specificity failed to rule out counterexamples in which irrelevant information found its way into the explanation. Salmon argued that the appropriate reference class for a statistical explanation is one that is *objectively* homogeneous, not one that is *epistemically* homogeneous, and that the crucial characteristic of a probabilistic explanation is statistical relevance, not high probability.

The notion of an objective homogenous reference class amounts to this: For any given reference class A, and for any given property C, there is, *in principle*, a partition of that class into two subclasses $A.C$ and $A.\sim C$. "A property C is said to be statistically relevant to B within A if, and only if, $p(A.C, B) \neq p(A, B)$" (p. 42). Using von Mises's concept of place selection, Salmon defines the conditions to which the reference class must conform:

> If every property $[C_1, C_2, C_3, ..., C_n]$ that determines a place selection is statistically irrelevant to B in A, I shall say that A is a *homogeneous reference class* for B. A reference class is homogeneous if there is no way, *even in principle*, to effect a statistically relevant partition without already knowing which elements have the attribute in question and which do not. (p.43)

Salmon then replaces Hempel's requirement of maximal specificity for the *reference class rule*: "Choose the broadest homogeneous reference class to which the single event belongs." (p.43). This characterization of statistical explanations is supposed to avoid any epistemic relativity because any statement of the form $p(G, F) = r$ that meets the homogeneity condition must be regarded as a fundamental statistical law of nature: its reference class cannot be further specified, not because we do not know how to make a further relevant partition, but because *in principle* it is impossible to make a further relevant partition.

Salmon then defines a statistical explanation as follows. If we want to know why a member of the class A has the property B, the answer will be a S-R explanation that consists of: (i) the prior probability that a member of the class A will have the property B: $p(B/A) = r$, (ii) a partition into homogeneous cells with respect to the property in question: $A.C_1$, $A.C_2$, etc., (iii) the posterior probabilities of the property in cells of the partition $p(B/A.C_1) = r_1$, $p(B/A.C_2) = r_2$, etc., and (iv) a statement of the location of the individual in question in a particular cell of the partition: "a is a member of $A.C_k$."

Two details should be noted about Salmon's definition. The first one is that negative relevance is granted explanatory import. This can be seen by considering that in some of the cells included in the explanation the posterior probability of the property will be lower than the prior probability. Salmon argues, and I agree, that

"negatively relevant factors as well as positively relevant factors contribute to our understanding of the phenomena we are studying" (1989, p. 67). The second one is that he explicitly requires the use of probability values in providing an explanation.

The use of probability values in Salmon's definition of statistical explanation stems from the fact that the S-R model is at bottom a covering-law model. Since any statement of the form $p(G, F) = r$ that meets the homogeneity condition must be regarded as a fundamental statistical law of nature, each of the probability sentences in the explanans of a S-R explanation is a law of nature. And since the factive condition on explanation demands that every element in an explanation must be true, the probability assigned to the explanandum by each of these probability sentences must be the right one.

To see how restrictive this requirement is, consider the following example provided by Humphreys:

> If a man dies from lung cancer, having been a heavy smoker, omitting from a probabilistic explanation any of the following minor relevant factors will result in a false probability claim: cosmic radiation from Alpha Centauri, particles from a chimney in Salem, Oregon, and a smoke-filled room he entered briefly at the Democratic convention eight years ago. It is good to be strict in matters of completeness, but not to the point of absurdity (1989, p. 111).

Humphreys argues that if one insists in providing the exact probability of the explanandum as part of the truth conditions of an explanation, it will be impossible to distinguish between a complete explanation and a true explanation. The omission of absurdly small probabilistically relevant factors, known or unknown, will result in a false explanation.

How can a true but incomplete statistical explanation be provided? Humphreys argues that instead of focusing on probability values, we should focus on causal relevance. An explanation should provide one or more of the factors that are causally relevant to an explanandum. A factor is causally relevant if it changes the propensity for an outcome. More precisely, "B is a direct contributing cause of A just in case ... B increases the chance of A in all circumstances Z that are physically compatible with A and B, and with A and B_0, where B_0 is the neutral state of **B**, i.e., $p(A/BZ) > p(A/B_0Z)$ for all such Z; and BZ and A are logically independent" (p. 74).[4] Similarly, B is a direct counteracting cause of A just in case B decreases the chance of A, that is, $p(A/BZ) < p(A/B_0Z)$.

A causal explanation can then be characterized as follows. Let Y, S, and t be terms referring to, respectively, a property or change of property, a system, and a trial. If one wants to know why Y occurred in S at t, an appropriate explanation will

---

4. A, B, and Z are specific events, which Humphreys defines as the "possession of specific values of a property on a given trial by an individual system" (p. 25). **B** is an event variable.

be, "Y in S at t occurred because of ϕ despite ψ," where "ϕ is a (nonempty) list of terms referring to contributing causes of Y; and ψ is a (possibly empty) list of terms referring to counteracting causes of Y" (p. 101). Humphreys says that ψ is not part of the explanation proper. "The role it plays is to give us a clearer notion of how the members of ϕ actually brought about Y—whether they did it unopposed, or whether they had to overcome causal opposition in doing so" (p. 101). According to my view of explanation as the purveyor of understanding, this restriction is unfounded, especially since counteracting causes are well-defined concepts in Humphreys's account. Despite this shortcoming, Humphreys approach captures the idea that probability values *per se* do not enhance our understanding of phenomena. Probability values have descriptive, predictive, and evidential value, but not explanatory value.[5]

Humphreys's strategy has the advantage that it makes it possible to offer a true explanation of an event by providing a contributing or a counteracting cause "even in cases where the other factors are not known and the true probability value cannot be calculated" (p. 112). So, for example, if the explanation of why Nietzsche developed paresis is that he suffered from untreated tertiary syphilis, the explanation will not be falsified if it is discovered that there is an additional factor

---

5. This does not mean, of course, that the facts involved in a probability sentence cannot be assigned numerical values. Whether Jones took one aspirin or a thousand makes a huge difference in the explanation of his death by drug poisoning.

that increases a syphilitic's chances of developing paresis. Although the explanation will no longer be complete, it will still be true. In Salmon's account, the original explanation would have to be discarded because the class of syphilitics would no longer be homogenous. Being able to offer a true but incomplete explanation also makes the model easier to apply in nonscientific contexts. Since my approach to explanation does not distinguish between scientific and nonscientific explanations, Humphreys model seems better suited for my purposes.

Unfortunately, there is an obvious objection to Humphreys approach which we have not considered. As the many versions of Simpson's paradox illustrate, one or more of the factors that the agent is unaware of can turn a contributing cause into a counteracting cause, or vice versa. Humphreys response to this objection is puzzling. He says: "Of course, epistemically, we can never know for certain that such confounding factors do not exist, but that is an entirely separate matter, although regrettably relative frequentists have often failed to separate epistemic aspects of probabilistic causality from ontic aspects" (p. 114).

It seems to me that it is Humphreys who is guilty of not keeping epistemic and ontic matters in separate baskets. If Salmon's model is too demanding, as Humphreys maintains, it is because we can never know if we have met all the conditions that it imposes on explanation. But Humphreys account suffers from a similar problem. In order for something to be a contributing or a counteracting

cause in Humphreys sense, there cannot be *any* further factor, known or unknown, that will invert the influence of these causes on the explanandum, or that will neutralize them altogether. Thus an agent who offers a causal statistical explanation will always have to relativize the explanation to a knowledge situation. In consequence, a distinction between an epistemic and an ontic notion of causal statistical explanation seems inevitable. Despite the advantages of Humphreys approach, he cannot avoid facing the epistemic relativity of explanation, to which we now turn.

1.2 *The Epistemic Relativity of I-S Explanation*

One of the goals of Salmon's account of statistical explanation was to eliminate the epistemic relativity of Hempel I-S model. Hempel had claimed that I-S explanations were *essentially* relativized, that is, that the relativity could not be eliminated. It is not clear what Hempel had in mind. Salmon thought that this statement reflected Hempel's secret commitment to determinism. Since the only clear cases of objectively homogenous reference classes are those that are trivially homogenous because they occur in universal generalizations, Salmon reasoned, "the thesis of epistemic relativization of I-S explanation implies that all I-S explanations are incomplete D-N explanations" (1984, p. 52). Hempel privately rejected the accusation.

The confusion about Hempel's assertion arises, it seems to me, because he failed to separate two different questions. The problem that Hempel had in mind

was an epistemological one. Unlike D-N explanations, the truth of the premises of an I-S explanation is not enough to make the explanation rationally acceptable. We also need to know that it is impossible to construct a rival argument with true premises and an incompatible conclusion. And since we will never be in a position to do so, to talk about nonrelativized true I-S explanations in the same sense that we talk about nonrelativized true D-N explanations is a mistake. Briefly stated, the problem is that the factivity of an I-S explanation does not warrant its rational acceptability.

Salmon pointed out, correctly, that if the statistical generalization is a fundamental stochastic law, no rival argument can be constructed and the truth of the premises is sufficient to guarantee the rational acceptance of the statistical explanation, Thus Hempel's claim that statistical explanation is *essentially* relativized is mistaken. To this, Hempel's only response was that there was no satisfactory account of the notion of a stochastic law: "I would certainly prefer an absolute characterization of statistical explanation to my relativized one, if the problem of the ontological ambiguity of such absolutely understood explanations can be satisfactorily resolved" ([1977] 2001, p. 175). In other words, eliminating the epistemic ambiguity of statistical explanation requires a satisfactory explication of the notion of a fundamental statistical law.

The problem with Hempel's response is that he transforms the original epistemological question into an ontological one. The original point still remains. Even if we obtain a satisfactory account of what a stochastic law *is*, we will never be in an epistemic position to accept an inductive-statistical explanation based only on the truth of the premises. This is the relativity that cannot be eliminated from Hempel's account of inductive-statistical explanation, and that does not arise in the case of deductive explanations. To make this evident, suppose that you are contemplating two arguments, one inductive-statistical and one deductive-nomological, and an omniscient god suddenly materializes in front of you and tells you that the premises of the two arguments are true. Hempel's original point is that even in that case you will only be able to unconditionally admit the deductive one. The statistical one will still have to be relativized to a knowledge situation because unfortunately the omniscient god disappeared into the philosophical ether without telling you whether there was a rival inductive-statistical argument with true premises and the same conclusion.

The accounts offered by Salmon and Humphreys avoid the epistemic relativity of Hempel's inductive-statistical explanations by introducing a condition that effectively rules out the possibility that a bona fide statistical explanation will be defeated by a rival statistical claim. The reference class of a S-R explanation is required to be objectively homogenous, and a contributing cause in Humphreys's

causal statistical explanations cannot be transformed into a counteracting cause by any hidden cause. The factivity condition thus recovers its importance in their accounts of probabilistic explanation.

1.3 *The Objectivity of Explanation*

The conclusion that we seem forced to draw from the foregoing analysis is that if truth is to play a role in an account of statistical explanation, epistemic relativity must be avoided at all costs. This claim has led many to say that the goal of inquiry is not truth but empirical adequacy or "epistemic optimality," as Hempel ([1990] 2000) called it in his later years. Instead, I want to argue that truth and epistemic relativity are not incompatible features of explanation.

The apparent incompatibility arises because the factivity condition, as it is understood by Salmon, Humphreys, and many others, is an expression of their commitment to the correspondence theory of truth. To require that the sentences in an explanation must be true is to require that they must correspond to the facts. But what are the facts that the explananda and the explanandum must correspond to? And how do we know when they so correspond? I have followed Mellor (1995) in defining a fact as whatever is stated by a true sentence, but Salmon *et al.* cannot use the same strategy to say what a fact is without begging the question. Perhaps they could argue that facts are particulars that have properties, but then they will have to say what a property is and what a particular is in a noncircular way. That

is, they cannot define a property *P* as whatever *x* has if '*x* is *P*' is true. These and other well-known difficulties stand in the way of a defensible correspondence theory of truth.[6]

The fact that I am using truth to define facts might suggest that I have a different account of what makes a sentence true. I do not. But I do not need to either. All that is required in order to implement the factivity condition for explanation is that we have an *objective* way of judging when a sentence is true and when it is false. And the belief-doubt model set forth by Peirce, and adopted by Levi, provides a way to do so, albeit an "epistemically relativized" one.

According to the belief-doubt model, an inquiring agent presupposes that everything he is currently committed to fully believing is true. This does not mean that truth or falsity is relative to what the agent believes. But the agent's *judgments* of truth are relative to what he believes. If the agent is concerned with establishing true explanations of phenomena, his decision to accept an explanation can only be made relative to the judgments of truth available to him. Naturally such decisions will lack any sort of objectivity.

---

6. See, for example, Davidson (1969; 1980), Putnam (1981). Davidson's argument against the correspondence theory, the infamous "slingshot," is also directed against any theory of causation whose relata are facts. Davidson, it seems, has no objection to using facts as the relata of explanation. If we consider the sentence 'the fact that ... *explains* the fact that ...', the substitution of equivalent sentences for the contained sentences, or of coextensive singular terms or predicates in the contained sentences, will not preserve truth. Without substitution *salva veritate*, Davidson's slingshot does not get off the ground. For a defense of facts as the relata of the relation of cause and effect, see Mellor (1995). I return to this issue below.

An agent who wants to claim objectivity for the explanations that he accepts must first make sure that the explanation is consistent with the basic facts accepted in the field. But that is not enough. The objectivity of our conjectures lies, as Popper correctly points out, "in the fact that they can be intersubjectively tested" (1959, p.44). The intersubjective test that an explanation must pass is the evaluation of its credibility and of its explanatory value in the eyes of the experts.

Suppose a group of inquirers—a community of experts in the field—wants to consider the adoption of an explanation. To do so, they must first adopt a belief state K^* representing the shared agreement between them. Such a belief state is available to them because the states of belief in a conceptual framework are partially ordered in a manner satisfying the requirements of a boolean algebra. In consequence, it will be possible to form the meet of their individual states, i.e., the strongest common consequence of all their states of belief. Obviously, such a state will contain more than just singular sentences representing facts and probability sentences representing empirical generalizations. It will also include sentences that state which are the most relevant problems in the field, what type of experiments and observations are considered more reliable, in addition to basic methodological and reasoning principles.[7]

---

7. These are some of the elements that Kitcher (1993, ch. 3) identifies as the basis for a *consensus practice*. Kitcher offers a meticulous model of the dynamics of scientific debate in which individual scientists driven by impure and noncognitive motives manage nonetheless to

Once the members of the community of experts have accepted a common corpus, they must take K^* as the basis for establishing a set of potential explanations of the problem at hand, that is, an ultimate partition in Levi's jargon.[8] For example, suppose a group of historians is trying to establish why Germany lost World War II. They must initially agree on a set of ground facts and low level hypotheses. Statistical data and the chronology of the war will be easy to agree upon. The explanation of some incidents, such as the failed Nazi attempt to occupy St. Petersburg, can be noncontroversially accepted, while the explanation of others will be a matter of heated debate. After the historians have agreed on a common corpus of beliefs K^*, they can put together a set of explanatory options which will include all the factors consistent with K^* that might explain the defeat of *die Wehrmacht*. At this stage of inquiry it does not matter whether the potential explanations are uncontroversial or completely outlandish, as long as they are somehow relevant to the problem at hand.

It is possible for a group of agents to share the same information and yet disagree about the degree of credal probability that they assign to the explanations in the set of explanatory options. Since the agents do not want to beg the question

---

form groups that develop in epistemically progressive ways. His approach uses a naturalistic background to formulate a prescriptive theory of scientific progress. Levi's (1984, ch. 6; 1997, chs. 7-9) approach to consensus forming is drawn along purely decision theoretical lines, and it is not limited to the case of science.

8. In Section 2 I provide a more precise characterization of the ideas informally presented here.

by assigning the highest marks to their favorite explanations, they must adopt a common credal probability measure. A common strategy to eliminate the conflict between different credal probability distributions is to represent the shared agreement as the weighted average of the distributions in conflict. In the case of explanation, it is to be expected that several explanatory hypotheses will rank highest in the distribution because typically there will be several factors that all the experts consider relevant to the explanandum. Thus the historians will all give high marks to the hypothesis that the Allied invasion of France on D-Day contributed to the defeat of the German army, and to the hypothesis that the fact that the Red Army entered the war had the same consequence. In contrast, the hypothesis that the position of the planets on the day that Hitler was born spelled disaster for Germany should get the lowest mark.

On the other hand, different historians will disagree in their assessment of the importance of the explanations contained in the set of explanatory options. A historian whose interest lies in the role of espionage during the war will judge the explanatory value of a potential explanation differently than one who is interested in the role of civilian opposition to the occupying German forces. Levi argues, and I agree, that despite these differences there must be a minimal objective criterion to measure the informational value of any potential expansion. That criterion is the

content or information carried by the potential expansion, which Levi identifies with its logical strength.

The problem is that not all potential explanations of a given fact are comparable in terms of content. The community of experts might then try to reach an agreement on the methodological criteria that they will use to evaluate the explanatory value of these residual explanations, as I will call them, and in most cases they will reach consensus on certain ground rules. I believe, however, that reaching a *complete* agreement cannot be a precondition for objective inquiry. As long as the experts agree to base their assessment of the risk of error incurred in accepting a potential explanation on the weighted average of their individual credal probability distributions, and as long as the evaluation of the explanatory value of an explanation is based on its informational content and on the basic methodological principles accepted in K^*, the objectivity of the explanation will be guaranteed and the factivity requirement will be fulfilled.[9] It seems to me that to

---

9. This approach is, I suspect, too conservative for someone committed to Kuhn's views about theory change. Kuhn (1970, 1977) argued that methodological principles often conflict with each other and are too imprecise to determine a unique choice. It seems to me, however, that Kuhn's claims about the nonspecificity of methodological criteria are largely exaggerated. There are many methodological principles that are widely accepted by one community or another, and that can be applied to concrete cases without imprecision or ambiguity. The requirement of consistency is an obvious example. So is the requirement of predictive accuracy. Kuhn's claim regarding the inevitable conflicts that arise between these criteria is also an overstatement. There are obvious tensions between criteria such as simplicity and accuracy, but such tensions do not prove that any family of methodological rules will be internally inconsistent, or that an agreement can only be reached by means of an irrational compromise. Again, there are examples of methodological principles, such as Mill's methods, that do not have a tendency toward the kind of conflict that Kuhn describes.

ask for unanimity in the assessment of the explanatory value of every single potential explanation is to unduly restrict the range of explanations that the individual agents can rationally accept. Objective inquiry can take place even when the members of a community do not assess the importance of all explanations and theories in exactly the same way. As John Earman correctly points out, all that is needed to guarantee the objectivity of inquiry is that "the community of experts share a paradigm in the weak sense of agreement on the explanatory domain of the field, on the circumscription of the space of possible theories to be considered serious candidates for covering the explanatory domain, on exemplars of explanatory success, and on key auxiliary hypotheses" (1993, p. 31).

The foregoing analysis indicates that the implementation of the factivity requirement for explanation will always be "epistemically relativized" in more ways that Hempel ever intended. But it is hard to see how such a relativization could be eliminated if we want to provide a coherent picture of the role of explanation in inquiry. If we adopt the view that truth and epistemic relativity are incompatible, as Salmon and Humphreys believe, we lose one of the main incentives in the search for explanations. Why would anyone want to incur the cost and effort involved in inquiry if the results cannot be assumed to be true in future decisions and deliberations? Levi uses this line of argument in defense of the belief-doubt model of inquiry:

> If inquiry cannot be motivated by a concern to remove doubt, what is its rationale? If we cannot incorporate the solutions we come close to establishing into the evidence and background information for future investigations, why should we care that we come close? The truth of the well-established conjecture remains an open question and a legitimate issue for future investigation. Inquiry never settles anything and, hence, inquiry—even inquiry into a specific problem—never legitimately terminates because the matter is settled but only, so it seems, because the investigators are tired or bored or have run out of funds. No matter how minute a question might be, if inquiry into that question is free of costs, it should go on forever (1991, p. 2).

A different type of objection to the belief-doubt model can be raised by someone who, like Kuhn, contends that an initial shift to a neutral state representing the shared agreement between the participants in inquiry is impossible. Kuhn (1970) denies that there is a non-question-begging point of view from which an agent can judge the truth of two competing hypothesis. As we saw above, Kuhn believes that no system of shared assumptions will be precise enough or strong enough to serve as a basis for a decision. Furthermore, in the "revolutionary" cases, the two parties cannot even understand each other because the two theories are incommensurable. The shift in an agent's loyalties from one theory to another will always be the result of an individual assessment of the epistemic virtues of the

competing theories based on the individual's cognitive interests and goals, combined with the operation of psychological and sociological forces.

But if each individual has a different set of reasons to accept a given hypothesis or set of beliefs, it becomes a mystery how an agreement is ever reached. Kuhn struggled with this issue in *The Structure of Scientific Revolutions* (1970) and in subsequent writings without ever offering a satisfactory solution. In fact, in the last chapters of *Structure*, Kuhn adopted a rather traditional approach to the problem: "To say ... that paradigm change cannot be justified by proof, is not to say that no arguments are relevant or that scientists cannot be persuaded to change their minds" (p. 152). The arguments that Kuhn mentions are based on the fact that a theory has better empirical support, is more fertile, resolves more problems than its competitors, and so on. Further down, Kuhn adds: "Because scientists are reasonable men, one or another argument will ultimately persuade many of them. But there is no single argument that can or should persuade them all" (p. 158). In *The Essential Tension*, Kuhn goes even farther when he argues that "the criteria or values deployed in theory choice are fixed once and for all, unaffected by their participation in transitions from one theory to another" (1977, p. 335). When one compares these passages with Kuhn's initial remarks about the underdetermination of theories by methodological rules, one has to conclude, in Laudan's words, that "it is unclear what all of Kuhn's earlier fuss about incommensurability and the ab-

sence of shared standards amounted to. He cannot have it both ways" (1984, p. 19).

The problem of consensus formation is a difficult one, and Kuhn's struggle with the issue is not surprising. But to deny at the outset that shared agreements can be the basis for the search for new error-free valuable information, as Kuhn and Salmon do from opposite sides of the spectrum, is to invoke a philosophical prejudice and not a principled argument.

1.4 *Potential Explanations and Explanation Spaces*

What, then, is the objective basis of explanation? My suggestion is that we should adopt a noncausal version of Humphreys's approach. The relation between causation and explanation is a highly disputed matter which I will not attempt to elucidate here. Humphreys restricts his account to causal explanations because he believes that the multiple uses of the word 'explanation' make it "preferable to work from causes to causal explanations rather than from a general sense of explanation down to a subcase" (1989, p. 100). I prefer to take the general sense of explanation that I presented in the previous chapter as the starting point of the analysis, and remain uncommitted about whether the account of explanation that I will provide has a causal interpretation.

Following Humphreys, the probability sentences that I will use do not mention the value of the probabilities they assign to the explanandum. Instead, they

state the positive or negative relevance of a factor to the explanandum.[10] Here we must make a choice. I have been using probability sentences of the form:

(1) $p(\phi/\psi) > p(\phi/\sim\psi)$ and $p(\phi/\psi) < p(\phi/\sim\psi)$

to capture the relations of positive and negative relevance. But there is another option:

(2) $p(\phi/\psi) > p(\phi)$ and $p(\phi/\psi) < p(\phi)$.

The unconditional probability $p(\phi)$ is a weighted average since it is given by

(3) $p(\phi) = p(\phi/\psi)p(\psi) + p(\phi/\sim\psi)p(\sim\psi)$.

The question is whether we should use the weighted average of the explanandum as the contrast class, or whether we should compare two conditional probabilities. The following example makes evident why the latter option is preferable.

Suppose there is a country where 95% of adults smoke. Call it France. Suppose that the probability that a French smoker will develop lung cancer $p(Cx/Sx)$ is

---

10. I am assuming that the probability used in these sentences is a real-valued function. An alternative approach would be to use imprecise probabilities (see Walley [1991] for a survey). The resulting account of explanation would differ substantially from the one presented here, and there are numerous problems associated with the notion of imprecise probability that must be dealt with before this avenue can be fully explored. Imprecise probabilities could also be used when assessing the credibility of the resulting intervals of probability values; see Gaifman (1986) for an account of higher order probabilities.

0.35, and suppose that for some unknown reason French nonsmokers have no chance of developing the disease, that is, $p(Cx/\sim Sx) = 0$. In consequence, the weighted average for lung cancer $p(Cx)$ among the French is 0.3325. Now, if we want to explain why François (f), a French adult, developed lung cancer, pointing out that he was a smoker should be considered an acceptable explanation of his illness. But if we compare the value of the conditional probability with the value of the weighted average, we find that the difference is 0.0175, a value that does not reflect the difference that smoking makes in François's life. What we want to know is whether the fact that François was a smoker makes a difference when we compare it to the situation in which he does not smoke at all. If we use inequality (1), we obtain the right result: $p(Cf/Sf) > p(Cf/\sim Sf)$, that is, $0.35 > 0$.[11]

Another important clarification about the use of probabilities in an account of explanation is that we ought to include a mechanism to weed out spurious correlations. The correlation between a rapidly falling barometric reading and the imminence of a storm does not explain why there is a storm coming, but the two correlated facts can be screened off from each other with the help of a third fact that can be used to explain both of them. We will say that the fact that *γ screens off*

---

11. The problem is not limited to cases in which the explanans is necessary for the explanandum. The difference in the numerical values that result from using (1) instead of (2) become even more striking when we move beyond the binary case. See Humphreys (1989, p. 41) for an example. Although my account does not use specific numerical values, it seems preferable to adopt probability sentences whose form allow for a more accurate representation of the situation.

ϕ from ψ just in case $p(\phi / \psi \,\&\, \gamma) = p(\phi / {\sim}\psi \,\&\, \gamma)$. In other words, γ makes ϕ and ψ statistically irrelevant to one another. Thus the probability that there will be a storm given that there is a drop in atmospheric pressure is the same whether the barometer is falling or not. After all, barometers may malfunction.

We are now ready to offer the definition of a potential explanation. Let K^* be the belief set that represents the shared agreement between a community of experts, and let ϕ be a sentence in K^*.

> A set of sentences Ψ is a *potential explanation* of the fact stated by ϕ relative to K^* just in case:
>
> (i) $K^* \cup \Psi$ is consistent.
>
> (ii) $\Psi \not\subset K^*$
>
> (iii) There is a ψ ($\psi \in L$) such that $\psi \in \Psi$.
>
> (iv) Either $p(\phi / \psi) > p(\phi / {\sim}\psi) \in \Psi$ or $p(\phi / \psi) < p(\phi / {\sim}\psi) \in \Psi$.
>
> (v) There is no $\gamma \in K^*$ such that $p(\phi / \psi \,\&\, \gamma) = p(\phi / {\sim}\psi \,\&\, \gamma)$.
>
> (vi) ϕ and ψ are logically independent.
>
> (vii) Nothing else is an element of Ψ.

A potential explanation is thus a set containing a singular sentence and a probability sentence that states the potential statistical relevance of the state of affairs described by the singular sentence to the explanandum. The first condition

states that a potential explanation must be consistent with the corpus of beliefs in which the explanandum is accepted. The second condition states that the potential explanation cannot be already accepted in K^*. The third condition says that the potential explanation must include a singular sentence that describes a potentially relevant factor. The fourth condition states that ψ is positively or negatively relevant to the fact that ϕ. The fifth condition guarantees that ϕ and ψ will not be spuriously correlated, as far as we know. Condition (vi) guarantees that ϕ will not explain itself. It also prevents the inclusion of trivial cases in which $p(\phi / \psi) = 1$ because $\psi \vdash \phi$. The last condition ensures that each potential explanation contains only one relevant factor. Bona fide explanations will typically mention several relevant factors.

Using our definition of a potential explanation, we can now characterize the notion of an *explanation space*. An explanation space can be understood as the subset of the set of sentences by which K^* can be consistently expanded that contains all the potential explanations of ϕ, regardless of whether the inquirers are aware of them or not.

(E_ϕ) For every sentence ϕ in K^*, there is a set $\{\Psi_1, \Psi_2, \ldots \Psi_k\}$ such that Ψ_i is an element of the set iff it is a potential explanation of ϕ. The set, denoted E_ϕ, is the *explanation space* of ϕ.

The explanation space will contain logically equivalent and empirically equivalent potential explanations. On the one hand, if $\Psi_1 = \{\psi, p(\phi / \psi) > p(\phi / \sim\psi)\}$ and $\Psi_2 = \{\eta, p(\phi / \eta) > p(\phi / \sim\eta)\}$, where ψ and η are logically equivalent, then Ψ_1 and Ψ_2 are logically equivalent potential explanations. If an agent accepts Ψ_1, she is thereby committed to Ψ_2. On the other hand, if ψ and η contain coextensive singular terms or predicates that occupy the same places in ψ and η, Ψ_1 and Ψ_2 will be empirically equivalent potential explanations. However, the explanatory value and the credibility of Ψ_1 and Ψ_2 will not be assessed in the same way unless the agents who assess them are aware that the singular terms or predicates are coextensive. Consider the following example. If a group of archeologists want to know why the door of an ancient temple where predawn sacrifices were performed faces east, they might all agree that it was because the high priest wanted to see the Morning Star while he performed the sacrifice. If no one in the group knows that the Morning Star is the planet Venus, they will not conjecture that the orientation of the temple's door can be explained by the high priest's desire to see Venus. But the latter explanation remains a potential explanation even if it is not identified by anyone. The elements of the explanation space are epistemically relativized only in

the sense that the consistency condition for potential explanations must be fulfilled relative to a knowledge situation.[12]

2. The Epistemic Value of Explanation

The explanation space provides all the possible explanations of a fact relative to a given knowledge situation, from the most recondite to the most obvious. However, rarely are we in a position to identify all of its elements. Our options are usually restricted to those potential explanations that our efforts and ingenuity have allowed us to find. We must therefore restrict the options available for expansion to a subset of the explanation space. This subset will be the basis for our assessment of the risk of error incurred and of the explanatory value obtained when we accept a potential explanation.

2.1 *The Credibility, Content, and Explanatory Value of Potential Explanations*

A set of explanatory options relative to E_ϕ, denoted O_ϕ, is the subset of the explanation space of ϕ that contains all the potential explanations of ϕ that the agents whose shared agreement is represented by K^* have been able to identify. To simplify the analysis, each potential explanation Ψ_i in the explanation space will be

12. Condition (iv) also introduces an element of epistemic relativity because the nonexistence of a screening off factor can only be guaranteed relative to K^*.

represented in O_ϕ by the conjunction of its elements, that is, by the conjunction of a singular sentence and a probability sentence, and denoted E_i.

Since O_ϕ will contain quite disparate and incompatible potential explanations, each agent must divide the explanations into those that seem credible, those that seem controversial, and those that seem outlandish. In other words, every agent must assign a different credal probability distribution to the sentences in O_ϕ. As we saw in the previous section, the conflict between the different credal probability distributions of O_ϕ can be solved by representing the shared agreement as the weighted average of the distributions in conflict. The resulting credal probability function C determines the objective risk of error incurred in accepting a potential explanation in O_ϕ. For every potential explanation E_i, the risk of error is $1 - C(E_i)$.

The risk of error incurred in accepting an explanation in the set of explanatory options must be compensated by the explanatory value thereby obtained. The potential explanations in O_ϕ will be more or less explanatorily valuable depending on how much information they contain and on how valuable that information is. How should we evaluate the informational content of a potential explanation? The idea that I will adopt here was first presented by Popper in *Logik der Forschung* (1935), and it has been defended in one way or another by Carnap, Bar Hillel, Levi, and others. Popper claimed that the content of a hypothesis, its "degree of

falsifiability," should be measured by how many other hypotheses it rules out. Likewise, I will argue that the informational content of a potential explanation in O_ϕ should be measured by how many other potential explanations it rules out. Content alone will not be sufficient to obtain a complete ordering of the elements of O_ϕ with respect to explanatory value, but it will provide an objective basis for the ordering.

Consider the following example. A building in Harlem has burned to the ground and the experts of the FDNY are trying to understand what happened. The potential explanations they are considering are:

E_1 = 'A fire bomb exploded in the basement' & The probability that the building will burn to the ground when a fire bomb explodes in its basement is higher than when there is no such explosion'.

E_2 = 'There was an accidental short-circuit in the basement' & ...

E_3 = 'The owner set the building on fire to claim the fire insurance' & ...

E_4 = 'There were old newspapers accumulated in the basement' & ...

E_5 = 'There was no sprinkler in the building' & ...

E_6 = 'There was a traffic jam near the fire that prevented the fire truck from arriving on time' & ...

Assuming that the experts have the required background information, and if we ignore for the moment the problems associated with preemption and overdetermi-

nation, they can conclude that E_4, E_5, and E_6 are compatible with any of the first three potential explanations, that E_1 and E_3 rule out E_2, and that E_2 rules out E_1 and E_3. So E_2 rules out more explanations than any other and is thus the potential explanation in O_ϕ with the highest informational content.[13] If E_2 also happens to be the most credible potential explanation in the eyes of the experts, then the most rational course of action for the experts seems to be to add E_2 to the corpus of beliefs representing their shared agreement. The set of explanatory options would then be reduced to those potential explanations in O_ϕ consistent with E_2, that is, to E_4, E_5, and E_6, and the evaluative process would continue based on this smaller set of options.

We must keep in mind that the set of explanatory options might include counteracting factors that are negatively relevant to the explanandum. In the example above, some feature of the fire might lead the FDNY experts to conjecture that the basement had been painted with fire-retardant paint. This conjecture might rule out the short circuit hypothesis, for example. In most cases, however, these poten-

---

13. Although I am not committed to the thesis that explanations are causal, some readers might be concerned with the difficulties associated with the notions of preemption and overdetermination. The latter is usually connected to sufficiency views of causation. It can be handled in my account by adding to O_ϕ an additional potential explanation formed by the conjunction of the two facts involved. The potential explanation will then be logically stronger and thus less informationally valuable than its conjuncts, as we shall see below. The problem of preemption, on the other hand, is connected to sine-qua-non analyses of causation. Since my analysis includes both facts that raise the probability of the explanandum and facts that lower it, the explanation cannot include two facts which are, respectively, sufficient to produce and prevent the explanandum.

tial explanations will have the lowest degree of informational content. Since we know that the explanandum is a fact, knowing what factors decreased its probability of being a fact will not rule out most of the possible factors that contributed to make it a fact. Potential explanations that state counteracting factors are more useful in those cases in which the factors have received numerical values. For example, if we want to explain why a certain variable in an experiment has the value observed, the factors involved can be divided into those that raise the probability of obtaining that value, and those that decrease it. A counteracting factor can then be used to rule out the contribution of different potential factors.

Popper's idea can be made more precise in the following way. Levi (1984, ch. 5) proposes a method to measure the informational content of the potential expansions in the ultimate partition that captures the idea that the content of an explanatory hypothesis should be measured by how many other explanations it rules out. Levi's proposal can be adapted without modification to the case of potential explanations. To measure the informational content of an explanation, we introduce a measure M that assigns nonnegative values to the elements of O_ϕ summing up to 1 and such that the M-value of a disjunction of elements of O_ϕ is equal to the sum of their individual M-values. The increment in informational content in expanding by adding E_i to K^* is the sum of the M-values of the elements of O_ϕ that are rejected. In other words, the informational content of E_i, denoted $Cont(E_i)$, is

$1 - M(E_i)$. As Levi points out, the M-function has the formal properties of a probability function.

Carnap, Bar Hillel, and others also used probability based notions of content. But the M-function was interpreted by these authors either as a credal probability function or as a measure of the degree of confirmation of the hypotheses. What is new in Levi's approach is that he rejects these interpretations and introduces a distinction between expectation and content determining probability. The latter is based on the partial ordering of the potential expansions of $\boldsymbol{K}$ introduced by a classical consequence relation. The set of potential expansions of $\boldsymbol{K}$ is a boolean algebra in which the maximum is $\boldsymbol{K}$ and the minimum is $\mathbf{0}$, the inconsistent state. If M is defined over this set, it will generate a partial ordering of its elements, and if the only element that has probability 0 is $\mathbf{0}$, potential expansions will strictly increase in probability with a decrease in strength. When the M-function is defined over the set of potential expansions of interest to the inquirer, i.e., over an ultimate partition, we obtain a measure of the informational content of the relevant potential expansions, which I have transformed into a measure of the informational content of a potential explanation.

As we saw at the beginning of the chapter, Levi establishes a distinction between informational content and informational value, a distinction that is captured

by the weak monotonicity requirement. I will adopt an analogous requirement in the case of explanation:

(WMR) If a potential explanation E_1 in O_ϕ carries at least as much new information as another potential explanation E_2 in O_ϕ, E_1 carries at least as much new explanatory value as E_2.

The weak monotonicity requirement generates a quasi-order of the potential explanations in O_ϕ which is based on the partial order generated by the *M*-function, but is not necessarily identical to it. Thus an agent might consider that a potential explanation E_1 that is a consequence of another potential explanation E_2 according to the partial order carries the same explanatory value than the stronger potential explanation because the additional information in E_2 has no value to him. But the stronger potential explanation cannot carry *less* explanatory value than the weaker one. Although some information is useless, it is never worthless.

Here we encounter the crux of the problem of objective explanatory value. On the one hand, even though the quasi-ordering generated by the weak monotonicity condition is based on the partial order generated by the *M*-function, different agents might assess informational value in different ways, all compatible with the *M*-function and in accordance with weak monotonicity. On the other hand, typically there will be elements in O_ϕ that are not comparable in terms of strength. In the example above, the potential explanations E_4, E_5, and E_6 are con-

sistent with K^*, but logically independent from each other and from the other potential explanations in the set of explanatory options. Thus they are not partially ordered by the M-function. They are examples of residual potential explanations.

Since the community of experts wants to adopt the best explanation available to them, they might invoke further criteria in order to assess the explanatory value of the elements of O_ϕ. Levi regards the criteria that are usually invoked to judge the epistemic virtues of a hypothesis or theory "as considerations that complete, to some degree, the quasi-ordering with respect to informational value generated by the [weak] monotonicity condition from the partial ordering with respect to information carried" (1991, p. 83). The question is to what degree can the quasi-ordering be completed and what considerations would be relevant in the case of explanation.

There are several explanatory virtues mentioned in the philosophical literature. Friedman (1974) and Kitcher (1989), for example, argue that explanations improve our understanding through the unification of our knowledge. Explanations that reduce the number of independent assumptions we have to make about the world are to be preferred to those that do not. This suggests that the potential explanations in O_ϕ could be ordered according to some set of rules that determines their unifying power.

The problem is that neither Friedman nor Kitcher have provided an account that can be applied to explanations generally. Friedman's original argument was intended as an account of the explanation of scientific laws. Friedman argued, for example, that the kinetic theory of gases is explanatory because it unified different laws and properties of gases that were previously disconnected. Friedman's only attempt to formalize and generalize the idea of explanation as unification was incisively criticized by Kitcher (1976) and Salmon (1989).

But Kitcher's account is no more helpful that Friedman's. According to Kitcher, the explanatory worth of candidates cannot be assessed individually. In his view, a successful explanation earns that name because it belongs to the explanatory store, a set that contains those derivations that collectively provide the best systematization of our beliefs. "Science supplies us with explanations whose worth cannot be appreciated by considering them one-by-one but only by seeing how they form part of a systematic picture of the order of nature" (1989, p. 430).

The idea that a virtuous explanation should have the potential to unify our beliefs is one that I wholeheartedly embrace. But no one, to my knowledge, has provided a general account of explanation as unification that is not restricted to the case of scientific laws or scientific explanatory exemplars.

Mellor (1995) provides an account of explanatory value that is better suited for our purposes. Mellor approaches explanation via his theory of causation. The

theory requires every cause to raise the chances of its effects. That is, a fact C causes a fact E iff $ch_C(E) > ch_{\sim C}(E)$. When causes are used in the explanation of a given fact, Mellor argues that the explanans must necessitate its explanandum, or at least raise its probability as much as possible, thereby reducing its chance of not existing. Thus, he concludes, "the more C raises E's chance the better it explains it" (p. 77).

The main problem with Mellor's proposal is that when we examine a genuinely stochastic process, the value of the chance that the cause confers on the explanandum will be irrelevant. As Jeffrey has convincingly argued, the information required to explain E is the same information used to explain $\sim E$, regardless of the value of the chance. Furthermore, if E is sometimes randomly caused by C and sometimes randomly caused by C^*, and $ch_C(E) > ch_{C^*}(E)$, there is no reason to think that C is a better explanation than C^*.

Mellor will respond to the objection by claiming that chances measure possibilities. "The less possible $\sim E$ is, i.e. the less $ch(\sim E)$ is and hence the greatest $ch(E)$ is, the closer the fact E is to being necessary. This is the sense in which a cause C may explain E better or worse, depending on how close it comes to making E necessary, i.e. on how much it raises $ch(E)$" (p. 77). Independently of whether we can make sense of such concepts as *almost necessary* or *nearly impossible*, it is not clear how such notions would enhance our notion of explanation.

Probabilities are important in statistical contexts because knowing that *C* raises the chance of *E* allows us to know what makes *E* possible, and because the chance that *C* gives *E* allows us to adjust our expectations of *E*'s occurrence. But it seems to me that mixing chances and possibilities adds nothing to our understanding of why *E* is a fact.

A third candidate for judging the epistemic value of an explanation is Whewell's (1837) notion of *consilience*. Consilience is intended to serve as a measure of how much a theory explains, and it can therefore be used to compare the explanatory value of two different hypotheses. One hypotheses has more explanatory value than another if the former explains more of the evidence than the latter. Thagard (1978) provides compelling evidence that this idea is often used by scientists in support of their theories. For example, Fresnel defended the wave theory of light by saying that it explained the facts of reflection and refraction at least as well as did the particle theory, and that there were other facts involving diffraction and polarization that only the wave theory could explain. Translated into my account, this means that if E_i raises the probability of more facts connected to the explanandum than E_j, then E_i is a better explanation than E_j.

The problem with consilience is that, once again, the account works well in the explanation of laws, but it will not work in the explanation of singular facts. Whether the traffic jam on 125th St. explains more facts connected to the fire than

the old newspapers accumulated in the basement of a Harlem building is hard to say. We would have to define what a fact "connected to the explanandum" is, and it is doubtful that a nonpragmatic formalization of this notion can be found. Besides, sometimes a theory can explain too much. Lavoisier accused the phlogiston theory of this particular crime.

Are there any other criteria that will allow us to assess the explanatory value of the potential explanations in O_ϕ? We still have not examined the values that are usually mentioned in the context of theory choice: simplicity, accuracy, fruitfulness, and so on. But such an analysis is unnecessary. If the criterion is such that the community of experts can agree on its importance and on how it should be applied in particular cases, they can incorporate it in the state representing their shared agreement. They will then be able to complete, to some degree, the quasi-ordering generated by the monotonicity condition with respect to the *M*-function. But to expect a complete agreement in the way that all the agents engaged in common inquiry assess the explanatory value of different potential explanation is to expect a heterogeneous group of inquirers to agree on what aspects of reality they find interesting or useful.

If a decision is required nonetheless, the community of experts can adopt the following compromise. The agents must first identify the elements of the set O_ϕ that can be completely ordered because they are comparable in terms of

strength or because they can be compared using the criteria to evaluate explanatory value that they have incorporated to K^*. The agents can then agree to disagree about the explanatory value of the remaining elements of $\mathbf{O}_\phi$. Let $\mathbf{O}^*_\phi$ be a set of explanatory options such that $\mathbf{O}^*_\phi \subseteq \mathbf{O}_\phi$ and such that the M-value of each element of the set is determined. Combining the credal probability function C that determines the objective risk of error incurred in accepting a potential explanation in $\mathbf{O}^*_\phi$ with the M-function defined over the elements of $\mathbf{O}^*_\phi$, we obtain a value that the community of experts can use to select the best explanation of ϕ. I will call this result the *objective epistemic value* of a potential explanation:

(4) $V^*(E_i) = \alpha C(E_i) + (1 - \alpha)M(E_i)$.

If we assume that $q = (1 - \alpha)/\alpha$, we obtain the following positive affine transformation of V^*:

(5) $EV^*(E_i) = C(E_i) - qM(E_i)$,

where q is the *index of boldness*. Since the explanations that the experts want to accept should not be false regardless of how much explanatory value they have, we must require that $\alpha \geq 0.5$, and thus that $0 \leq q \leq 1$. And since a determinate degree of boldness must be established for the group to be able to make decisions, the index should be the average of their individual indices. Once this is settled, the ex-

perts should reject a potential explanation in $\mathbf{O}^*_\phi$ if $EV^*(E_i)$ is negative, remain uncommitted if it is 0, and refuse to reject it if it is positive. Any potential explanation in $\mathbf{O}^*_\phi$ with positive objective epistemic value is an *objective explanation* of ϕ in K^*, denoted O_ϕ. The disjunction of all such objective explanations is *the* objective explanation of ϕ in K^*:

(OE_ϕ) The *objective explanation* of ϕ in K^*, denoted OE_ϕ, is the disjunction of all the potential explanations in $\mathbf{O}^*_\phi$ with positive objective epistemic value.

The account of objective explanation that I have provided has been relativized to K^*, the corpus representing the shared agreement of a group of experts in the field, but our ultimate goal is an account of objective explanation that is relativized to K, the individual agent's corpus of beliefs. This will require only a minor modification of the foregoing definition.

2.2 *Explanations in* K

An objective explanation in K^* is based on a compromise set of explanatory options because in most cases there will not be a complete agreement regarding the explanatory value of all the potential explanations in $\mathbf{O}_\phi$. But an individual agent does not have to compromise her own view of what is important and what is not. She is free to evaluate the explanatory value of the remaining elements of $\mathbf{O}_\phi$ according to her interests and goals. In other words, she can complete the quasi-

ordering generated by the weak monotonicity condition any way she pleases, with the proviso that her assessment of explanatory value must take into account the prior assessment of the community of experts. As a member of a learning community, the agent can only claim objectivity for the explanations that she adopts if they are based on: (i) the set of explanatory options O_ϕ adopted by the community (ii) the weighted average of the different credal probability distribution that the members of the community assign to the elements of O_ϕ, (iii) the partial order introduced by the *M*-function, and (iv) the criteria for evaluating explanatory value that the members of the community have incorporated to the state representing their shared agreement. When all these elements are taken into account, and if we assume that the agent has assigned *M*-values to all the elements of O_ϕ, we obtain a value that the agent can use to select the best explanation of ϕ. I will call this result the *individual epistemic value* of a potential explanation:

(6) $V(E_i) = C(E_i) - qM(E_i)$.

Since the standards of rationality should be no different for the individual and for the group, the agent should reject a potential explanation in O_ϕ if $V(E_i)$ is negative, remain uncommitted if it is 0, and refuse to reject it if it is positive. A potential explanation in O_ϕ with positive individual epistemic value is an *explana-*

tion in K of ϕ, denoted E_ϕ. The disjunction of all such explanations in K is *the* explanation in K of ϕ:

(EK_ϕ) The *explanation in K* of ϕ, denoted EK_ϕ, is the disjunction of all the potential explanations in $\mathbf{O}_\phi$ with positive individual epistemic value.

For generality, it will be interesting to define a notion of explanation in which the only standards of objectivity that the agent respects are the weak monotonicity condition and the requirement that the explanation must be consistent with his beliefs. I will call these type of explanations *subjective explanations*, and they can only be assigned *subjective epistemic value*. A subjective explanation is based on a corpus of beliefs K, relative to which the agent establishes a set of explanatory options $\mathbf{S}_\phi$. The agent then assigns credal probability values to the elements of $\mathbf{S}_\phi$ and completes the quasi-ordering with respect to explanatory value generated by the weak monotonicity condition from the partial ordering yielded by the M-function. We obtain the subjective epistemic value of the elements of $\mathbf{S}_\phi$:

(7) $S(E_i) = C(E_i) - qM(E_i)$.

Any potential explanation in $\mathbf{S}_\phi$ with positive subjective epistemic value is a *subjective explanation* of ϕ in K, denoted S_ϕ. The disjunction of all such potential explanations becomes *the* subjective explanation of ϕ in K:

(SE_ϕ) The *subjective explanation* of ϕ in K, denoted SE_ϕ, is the disjunction of all the potential explanations in $\mathbf{S}_\phi$ with positive subjective epistemic value.

A subjective explanation S_ϕ can become an objective explanation, or an explanation in K, if the agent contracts his corpus of beliefs by giving up S_ϕ and agrees to submit it to the tribunal of intersubjectivity. If the result of the evaluation by the members of the learning community to which the agent belongs is that S_ϕ has positive objective epistemic value, what once was a subjective explanation now becomes an objective explanation. If the community of experts cannot agree on how to evaluate the explanatory value of S_ϕ, the agent can use the partial evaluation of the community to obtain a certificate of objectivity and accept S_ϕ as an explanation in K.

The transformation of a subjective explanation into an objective one, or into an explanation in K, assumes that it is rational for the agent to contract his state of belief. A complete analysis of contraction is beyond the scope of this dissertation, but something must be said about the agent's decision to give up some of his beliefs in order to accept the verdict of the community of experts. Suppose an agent, call him Einstein, fully believes S_ϕ, and he submits his explanation to the community of experts. Now suppose S_ϕ is judged to be maximally credible and maximally valuable by the community, and becomes O_ϕ. Does Einstein understand *more* now

that his explanation has been certified by others? It seems to me that he does not. But if the agent does not obtain more understanding from this recognition, why should anyone seek objectivity for an explanation that he or she already believes?

Part of the answer is that the belief-doubt model is not a recipe for dogmatism. A seldom noted fact about inquiry is that most newly suggested explanatory hypothesis do not survive the test of intersubjective scrutiny. If the agent is aware of this fact—and he should if he is a responsible inquirer—it would be imprudent for him to give his full assent to an explanatory hypothesis that contradicts firmly established theories and findings without obtaining at least a partial intersubjective assessment of its merit. Whether Einstein fully believed that GTR was true when he presented it to his peers, nobody knows. But he did not need to fully believe that it was true to obtain the understanding that the theory provided. Any inquirer can explore the consequences of a hypothesis by assuming, for the sake of argument, that it is true. If the hypothesis is judged to have positive objective epistemic value by a community of experts, the inquirer will then be fully justified in giving it his full assent.[14]

---

14. By accepting GTR, or any other theory that alters the foundations of a discipline, the members of the community are incurring an enormous cost, namely, the rejection of an important cluster of deeply entrenched beliefs. From a decision theoretical point of view, an agent will only be justified in giving up her prior beliefs if the benefits of doing so outweigh the cost. The acceptance of an explanatory theory that unsettles prior convictions is an example of those epistemic contexts considered in Chapter 3 in which an agent rejects some or all of the elements of the explanation prior to his decision to accept it.

But the question remains. If the agent does not obtain new understanding in the approval that he receives from his peers, why should he seek their approval? What prevents an agent from individually assessing the explanatory value of the corpus that results from accepting the explanation, and deciding to fully believe the explanation if the explanatory value of his corpus is thereby increased? In other words, why should objectivity matter?

The answer is that objectivity itself is a property of information that some agents find valuable and some do not. An agent who decides to be a member of a learning community does so because she believes that her beliefs will be more valuable if they are objective. Other agents will find that objectivity adds no value to their corpus of beliefs. Just as there is a difference between objective and subjective explanation, there is an analogous distinction between objective and subjective understanding. The latter is the type of understanding that Hempel correctly believed should be shunned at all costs from an account of *scientific* explanation. But the reason it should be shunned is not that it is an inferior type of understanding. The reason is that the members of a scientific community are among the many agents who find objectivity valuable. Therefore, an account of scientific explanation should avoid any reference to an evaluative process in which the agent shows no concern for the views of others.

Objective explanations provide the type of understanding that Friedman had in mind when he claimed that there is "a sense on which what is scientifically comprehensible is constant for a relatively large class of people" (1974, p. 8). If we remove the term 'scientifically', which is philosophically insignificant, we obtain the type of objective understanding that the agents involved in common inquiry, broadly conceived, regard as the highest epistemic good.

In the concluding section of the first chapter I said that we should take Friedman's objective notion of understanding seriously and explore the possibility of characterizing, in logically precise terms, a notion of explanation that is both objective and pragmatic, that does not depend on the idiosyncrasies of the individuals involved but that regards their epistemic states as a fundamental part of the analysis. The account of explanation presented in this chapter is the fulfillment of that promise.

BIBLIOGRAPHY

Achinstein, P. (1983). *The nature of explanation*. New York: Oxford University Press.

Alchourrón, C. E. & Makinson, D. (1982). On the logic of theory change: Contraction functions and their associated revision functions. *Theoria 48*, 14-37.

Alchourrón, C. E. & Makinson, D. (1985). On the logic of theory change: Safe contraction. *Studia Logica 44*, 405-422.

Alchourrón, C. E., Gärdenfors, P. & Makinson, D. (1985). On the logic of theory change: Partial meet functions for contraction and revision. *Journal of Symbolic Logic 50*: 510–530.

Belnap, N. D. & Steel, J. B. (1976). *The logic of questions and answers*. New Haven: Yale University Press.

Bromberger, S. (1966). Why-questions. In Colodny, R. G. (Ed.). *Mind and cosmos*. Pittsburgh: University of Pittsburgh Press.

Carnap, R. (1950). *The logical foundations of probability*. Chicago: University of Chicago Press.

Chopra, S. (2000) *Belief structures and sequences: Relevance-sensitive, inconsistency tolerant models for belief revision*. Ph.D. Dissertation. The City University of New York.

Chopra, S. & Parikh, R. (2000). Relevance sensitive belief structures. *Annals of Mathematics and Artificial Intelligence 28*: 259-285.

Danto, A. (1981). *The transfiguration of the commonplace*. Cambridge: Harvard University Press.

Davidson, D. (1969). True to the facts. *Journal of Philosophy 66*, 748-764.

Davidson, D. (1980). Causal relations. In *Essays on actions and events*. New York: Oxford University Press.

Davidson, D. (1984). On the very idea of a conceptual scheme. In *Inquiries into truth and interpretation*. New York: Oxford University Press.

Dretske, F. (1972). Contrastive facts. *Philosophical Review 81*, 411-437.

Earman, J. (1993). Carnap, Kuhn, and the philosophy of scientific methodology. In Horwich, P. (Ed.). *World changes. Thomas Kuhn and the nature of science*. Cambridge, MA: MIT Press.

Elgin, C. Z. (1996). *Considered judgment*. Princeton: Princeton University Press.

Ellis, B. (1979). *Rational Belief Systems*. Oxford: Blackwell.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Fetzer, J. H. (1974). A single case propensity theory of explanation. *Synthese 28*, 171-198.

Fetzer, J. H. (1981). *Scientific knowledge*. Dordrecht: Reidel.

Feyerabend, P. (1962). Explanation, reduction, and empiricism. In Feigl H. & Maxwell, G. (Eds.). *Minnesota studies in the philosophy of science, Vol. III*. Minneapolis: University of Minnesota Press.

Friedman, M. (1974). Explanation and scientific understanding. *Journal of Philosophy 71*, 5-19.

Fuhrman, A. (1991). Theory contraction through base contraction. *Journal of Philosophical Logic 20*, 175-205.

Gaifman, H. (1986). A theory of higher order probabilities. In Halpern, J. (Ed.). *Theoretical aspects of reasoning about knowledge*. Los Altos, CA: Morgan Kaufman.

Gärdenfors, P. (1988). *Knowledge in flux. Modeling the dynamics of epistemic states*. Cambridge, MA: MIT Press.

Gärdenfors, P. (Ed.). (1992). *Belief revision*. Cambridge: Cambridge University Press.

Gärdenfors, P. & Makinson, D. (1994). Nonmonotonic inference based on expectations. *Artificial Intelligence 65*, 197-245.

Garfinkel, A. (1981). *Forms of explanation*. New Haven: Yale University Press.

Hansson, S. O. (1991). *Belief base dynamics*. Ph.D. dissertation. University of Uppsala.

Hansson, S. O. (1993). Reversing the Levi identity. *Journal of Philosophical Logic 22*, 175-203.

Harman, G. (1965). The inference to the best explanation. *The Philosophical Review 74*, 88-95.

Harman, G. (1986). *Change in view: Principles of reasoning*. Cambridge, MA: MIT Press.

Harper, W. L. (1977). Rational conceptual change. In Suppe, F. & Asquith P. D. (Eds.). *PSA 1976*. East Lansing, MI: Philosophy of Science Association.

Hempel, C. G. (1965). *Aspects of scientific explanation*. New York: The Free Press.

Hempel, C. G. ([1977] 2001). Postscript 1976: More recent ideas on the problem of statistical explanation. In Fetzer, J. H. (Ed.). *The philosophy of Carl G.*

Hempel. Studies in science, explanation, and rationality. New York: Oxford University Press.

Hempel, C. G. ([1990] 2000). The irrelevance of the concept of truth for the critical appraisal of scientific theories. In Jeffrey, R. (Ed.). *Carl G. Hempel. Selected philosophical essays*. New York: Cambridge University Press

Hempel, C. G. & Oppenheim P. (1948). Studies in the logic of explanation. In Hempel (1965).

Hintikka, J. (1962). *Knowledge and belief: An introduction to the logic of the two notions*. Ithaca: Cornell University Press.

Hitchcock, C. R. (1995). The mishap at Reichenbach Falls. Singular vs. general causation. *Philosophical Studies 78*, 257-291.

Humphreys, P. (1989). *The chances of explanation. Causal explanation in the social, medical, and physical sciences*. Princeton: Princeton University Press.

Jeffrey, R. (1971). Statistical explanation vs. statistical inference. In Salmon, W. C. (Ed.). *Statistical explanation and statistical relevance*. Pittsburgh: Pittsburgh University Press.

Kakas, A. C., Kowalski, R. A. & Toni, F. (1993). Abductive logic programming. *Journal of Logic and Computation 2*, 719-770.

Kitcher, P. (1976). Explanation, conjunction, and unification. *Journal of Philosophy 73*, 207-212.

Kitcher, P. (1984). *The nature of mathematical knowledge*. New York: Oxford University Press.

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In Kitcher & Salmon (1989).

Kitcher, P. (1993). *The advancement of science*. New York: Oxford University Press.

Kitcher, P. & Salmon W. C. (1987). Van Fraassen on explanation. *The Journal of Philosophy 84*, 315-330.

Kitcher, P. & Salmon W. C. (Eds.). (1989). *Scientific explanation. Minnesota studies in the philosophy of science, Volume XIII*. Minneapolis: University of Minnesota Press.

Kuhn, T. S. (1970). *The structure of scientific revolutions*. 2nd edition. Chicago: The University of Chicago Press.

Kuhn, T. S. (1977). Objectivity, value judgment, and theory choice. In *The essential tension*. Chicago: The University of Chicago Press.

Lakatos, I. (1978). *The methodology of scientific research programmes. Philosophical papers, volume 1*. Cambridge: Cambridge University Press.

Laudan, L. (1984). *Science and values*. Berkeley: University of California Press.

Levi, I. (1977). Subjunctives, dispositions, chances. *Synthese 34*, 423-455.

Levi, I. (1980). *The enterprise of knowledge*. Cambridge, MA: MIT Press.

Levi, I. (1984). *Decisions and revisions*. Cambridge: Cambridge University Press.

Levi, I. (1988). Four themes in statistical explanation. In Harper W. L. & Skyrms B. (Eds.). *Causation in decision, belief change, and statistics*. Boston: Kluwer Academic Publishers.

Levi, I. (1991). *The fixation of belief and its undoing*. New York: Cambridge University Press.

Levi, I. (1996). *For the sake of the argument*. Cambridge: Cambridge University Press.

Levi, I. (1997). *The covenant of reason*. Cambridge: Cambridge University Press.

Levi, I. (1998). Contraction and informational value. Unpublished manuscript. Available at www.columbia.edu/~levi.

Lewis, D. (1982). Logic for equivocators. *Noûs 16*, 431-441.

Lewis, D. (1986). Causal explanation. In *Philosophical papers*. Volume II. New York: Oxford University Press.

Lipton, P. (1991). *Inference to the best explanation*. London: Routledge.

Mellor, H. D. (1995). *The facts of causation*. London: Routledge.

Pagnucco, M. (1996). *The role of abductive reasoning within the process of belief revision*. Ph.D. dissertation, University of Sydney.

Parikh, R. (1999). Beliefs, belief revision, and splitting languages. In Moss, L., Ginzburg, J. & de Rijke, M. (Eds.). *Logic, Language, and Computation*. CSLI Publications.

Peirce, C. S. (1932). The fixation of belief. In Hartshorne, C. & Weiss, P. (Eds.). *Collected Papers of Charles S. Peirce*. Cambridge, MA: Belknap Press.

Peng, Y. & Reggia, J. A. (1990). *Abductive inference models for diagnostic problem-solving*. Berlin: Springer-Verlag.

Popper, K. (1935). *Logik der Forschung*. Vienna: Springer.

Popper, K. (1959). *The Logic of Scientific Discovery*. London: Hutchinson.

Putnam, H. (1978). *Meaning and the moral sciences*. London: Routledge & Kegan Paul.

Putnam, H. (1981). *Reason, truth, and history*. New York: Cambridge University Press.

Railton, P. (1978). A deductive-nomological model of probabilistic explanation," *Philosophy of Science 45*, 206-226.

Railton, P. (1980). *Explaining explanation*. Ph.D. dissertation. Princeton University.

Rescher, N. (1970). *Scientific explanation*. New York: The Free Press.

Rott, H. (1991). Two methods of constructing contractions and revisions of knowledge systems. *Journal of Philosophical Logic 20*, 149-173.

Rott, H. (2001). *Change, choice, and inference. A study of belief revision and nonmonotonic reasoning*. Oxford: Oxford University Press.

Rott, H. & Pagnucco, M. (1999). Severe withdrawal (and recovery). *Journal of Philosophical Logic 28*, 501-547.

Ruben. D. H. (1990). *Explaining explanation*. London: Routledge.

Salmon, W. C. (1971). Statistical explanation. In Salmon, W. C. (Ed.). *Statistical explanation and statistical relevance*. Pittsburgh: Pittsburgh University Press.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton: Princeton University Press.

Salmon, W. C. (1989). *Four decades of scientific explanation*. Minneapolis: University of Minnesota Press.

Salmon, W. C. (1998). *Causation and explanation*. New York: Oxford University Press.

Scriven, M. (1959). Truisms as the grounds for historical explanation. In Gardiner, P. (Ed.). *Theories of history*. New York: The Free Press.

Thagard, P. (1978). The best explanation. Criteria for theory choice. *Journal of Philosophy 75*, 76-92.

Toulmin, S. (1961) *Foresight and understanding. An enquiry into the aims of science*. New York: Harper & Row.

van Fraassen, B. (1980). *The scientific image*. Oxford: Clarendon Press.

Van Fraassen, B. (1980a). Rational belief and probability kinematics. *Philosophy of Science 47*, 165-187.

Von Wright, G. H. (1971). *Explanation and understanding*. Ithaca: Cornell University Press.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. London: Chatman & Hall.

Wassermann, R. & Hansson, S. O. (1999). Local change. Research Report PP-1999-17. Institute for Logic, Language, and Computation. University of Amsterdam. To appear in *Studia Logica*.

Whewell, W. (1837). *History of the inductive sciences*. London: Parker.

Woodward, J. (1984a). A theory of singular causal explanation. *Erkenntnis 21*, 231-262.

Woodward, J. (1986). Are singular causal explanations implicit covering-law explanations? *Canadian Journal of Philosophy 16*, 253-280.