Understanding Interests and Causal Explanation

Petri Ylikoski

Academic Dissertation
Department of Moral and Social Philosophy
University of Helsinki
2001

ISBN 952-91-3373-3 (nid.)
ISBN 951-45-9942-X (PDF)
http://ethesis.helsinki.fi

Contents

Acknowledgements

I am grateful to all my friends and colleagues for their help and support while working on this thesis. I wish to thank Jeroen Van Bouwel, Mika Kiikeri, Tomi Kokkonen, Aki Lehtinen, Kaarlo Miller, Rebecca Schweder and Robrecht Vanderbeeken for useful comments on parts of the manuscript. I am especially grateful to Pekka Mäkelä and Susanna Snell for their numerous suggestions and comments in all phases of this work. Susanna also designed the layout of this book.

Professor Martti Kuokkanen and Professor Matti Sintonen made insightful comments on the penultimate version of the manuscript. Professor Raimo Tuomela provided a valuable opportunity to work in his research project while writing this thesis. Professor Henry Fullenwider skillfully revised my English. I wish to thank all of them.

Finally, special thanks are due to my parents Hilkka and Matti Ylikoski and my brother Tero. This book is dedicated to Susanna.

Helsinki, April 11, 2001

Petri Ylikoski

Chapter 1
Introduction

This work consists of two parts. Part I will be a contribution to a philosophical discussion of the nature of causal explanation. It will present my contrastive counterfactual theory of causal explanation and show how it can be used to deal with a number of problems facing theories of causal explanation. Part II is a contribution to a discussion of the nature of interest explanation in social studies of science. The aim is to help to resolve some controversies concerning interest explanation by explicating the concept of interest and its explanatory uses by using the account of explanation developed in Part I.

In this introductory chapter I will go through the main arguments of this work and discuss some general issues concerning its structure and philosophical motivation that are not addressed elsewhere in the work.

Theory of explanation

My basic account of explanation is outlined in Chapter 2. I call it *the contrastive counterfactual model of explanation*. The same basic ideas can be used to analyze other forms of explanation, but in this work I will concentrate on a singular causal explanation.

There are a couple of reasons for this choice. First, my original motivation for developing an account of explanation was a desire to find some philosophical tools that would help to resolve controversies concerning explanation in social studies of science. One such controversy concerns the nature of interest explanation, which will be the topic of Part II. Explanations in social studies of science are typically singular and causal, which explains my special interest in singular causal explanation.

Secondly, most philosophical discussions of explanation are about singular explanation (cf. Hempel 1965; Ruben 1990). A number of phi-

losophers have criticized this emphasis and pointed out that, especially in the advanced sciences, explanations have generalizations as their *explananda* rather than singular facts (Friedman 1974: 5; Woodward 1979: 63; Tuomela 1980: 217). I agree that philosophy of science has probably concentrated too much on singular explanation. However, it does not follow that interest in singular causal explanation is illegitimate. To the contrary, I think that although no individual singular fact *per se* is of interest to pure theoretical science, the ultimate test for any scientific theory is its applicability to (singular) facts about the world. At any rate, the fact is that most philosophical discussion is about singular causal explanation. This fact makes it prudent to develop one's account of singular causal explanation first, since it makes the evaluation of the suggestion easier. A comparison with the existing accounts provide an important standard of evaluation: the new theory should do at least as well as any of the earlier theories. Success in this competition is a good starting point for the development of accounts of other kinds of explanation using the same basic ideas.

The contrastive counterfactual model consists of a number of theses that are discussed and defended in Chapter 2. The first thesis is that the *explanandum* is best seen as contrastive. An explanation is an answer to a question that asks why *f* occurred instead of *c*. It tells us what made the difference between *f* and *c*. I defend a strong version of the contrastive thesis by arguing that *f* and *c* should be exclusive alternatives. I also argue that that a contrastive *explanandum* is not reducible to a non-contrastive statement.

The second thesis is that an adequate explanation has counterfactual form. I argue that counterfactual theory of causal explanation is distinct from counterfactual theory of causation, and that this idea can be made to work only when combined with the idea of contrastive *explanandum*.

My third thesis arises from discussion of presuppositions of sensible contrast. After reviewing a number of alternatives, I end up requiring that an explanation should specify the mechanism that ensures that *f* occurs instead of *c*. The mechanism is an answer to a how-question underlying every why-question. The requirement of such a mechanism provides us also with a very simple way of dealing with cases of causal pre-emption and overdetermination.

Note that my account makes use of some ideas from pragmatic theories of explanation, but it is not theory of the pragmatics of explanation. I am interested here in the product of explanation, not in the process of explanation. This focus sets my work apart from pragmatic theories of explanation (cf. Gärdenfors 1980; Tuomela 1980; van Fraassen 1980; Achinstein 1983; Sintonen 1984), which concentrate on speech acts of explanation and on the conditions of successful communication of explanatory information. In this work my question is:

given that an explanation does not fail for pragmatic or factual reasons, what does it explain?

Chapter 3 will focus on four issues relevant to the plausibility of my account of explanation. The first issue concerns indeterminism and its implications for a theory of explanation. Wesley Salmon has argued that the Leibniz principle does not hold in indeterministic contexts. The principle says that if *a* explains *f* in one case, and *c* and *f* are incompatible, then *a* cannot explain *c* in another (similar) case. Since the contrastive theory is based on this principle, Salmon's argument challenges the validity of the whole approach.

I argue that Salmon has not produced convincing reasons for giving up the Leibniz principle. The existence of indeterministic processes does not directly imply anything about explanation. Secondly, I argue that Salmon's arguments to the effect that the Leibniz principle is not needed to distinguish between explanations and pseudo-explanations fail. Thirdly, I suggest that the real motivation for accepting the Leibniz principle is that it reflects our conviction that explanations should trace objective relations of dependency. Finally, after these auxiliary arguments, I challenge Salmon's main argument. This argument claims that if we do not give up the Leibniz principle, we cannot explain indeterministic events at all. I reply by arguing that the supporter of the contrastive approach can explain as much about indeterministic events as Salmon can. Furthermore, I argue that the contrastive approach is an indispensable tool for clarifying what we are capable of explaining in indeterministic contexts.

After discussing indeterministic processes I proceed to discuss probabilistic versions of contrastive theory. These versions claim that an explanation does not need to show why *f* had to occur instead of *c*. It is only required that the explanatory factor makes *f* more probable. I argue against these views by showing that they face some difficult conceptual problems, especially in the case of singular causal explanation. I conclude that we should not dilute our standards of explanation, since intuitions behind probabilistic theories can be saved by using the concept of partial explanation.

The third challenge concerns the role of deduction in explanation. I discuss Wesley Salmon's arguments against what he calls the third dogma of empiricism. This 'dogma' says that all scientific explanations are arguments. I show that Salmon's formulation of this thesis is ambiguous, and that the third dogma can be interpreted in four different ways. Salmon's arguments only apply to the two strongest versions of the thesis. After some further discussion I conclude that one should stick to the weakest formulation of the dogma. According to this view, all explanations can be reconstructed as deductive arguments. However, the actual derivation does not have any constitutive role in an explanation.

The final section of Chapter 3 will deal with the role of laws in singular causal explanation. I start by observing that the deductive ideal in itself does not require that the reconstruction of an explanation as a deductive argument includes laws as its premises. Then I proceed to evaluate various arguments to the effect that laws should have an essential role in an singular causal explanation. I show that all these arguments are wanting and argue that there are some good reasons not to require that there are covering laws for all singular causal explanations. However, I do not find that laws and generalizations completely irrelevant to singular causal explanations. I argue that they can have an important role in the *search* for explanations.

The outcome of Chapter 3 is that the contrastive counterfactual account of causal explanation can be defended successfully against some standard arguments that are usually considered to support positions competing with it.

Chapter 4 discusses the problem of macro explanation, that has been widely discussed in philosophy of mind. The problem is that under some widely held assumptions, the explanatory power of macro explanations turn out to be epiphenomenal. The respective micro explanation seems to do all the macro explanation does, plus much more. In order to illustrate the problem, I discuss the biological concept of fitness. I argue that fitness has an ineliminable explanatory role, and that explanations in terms of it cannot be replaced with explanations in terms of its causal bases. The problem of macro explanation is solved by paying more careful attention to the *explananda* of micro and macro explanations.

These results fit nicely with the counterfactual contrastive account of explanation. In order to further illustrate the strength of this approach, I critically analyze the program model of explanation developed by Frank Jackson and Philip Pettit. I argue that my account can accommodate all the intuitively right results of their model, while avoiding its conceptual problems. For example, it turns out that the 'program' and the 'process' explanations are not different explanations of the same *explanandum*, but basically similar explanations of slightly different *explananda*.

The final section of Chapter 4 sketches an account of intentional explanation. I show that the contrastive counterfactual model of causal explanation applies in a straightforward manner to intentional explanation. The central explanatory mechanism in intentional explanation is practical reasoning by the agent. An intentional explanation shows how the agent's beliefs, desires and intentions led her to the choice she had made via a process of deliberation. The idea is that differences in the inputs for the process of practical reasoning make differences in the outputs, the agent's actions. Had the agent had different beliefs or desires, she would have behaved differently.

Causality and causal explanation

Some basic comments about the relation between causality and causal explanation are appropriate here. After all, it is legitimate to ask about this relation when one encounters an account of causal explanation which is not built on any explicit theory of causation.

First, I do not attempt to give an account of causation in terms of explanation, as Kitcher (1989: 420, 436-437) does. To the contrary, my account presupposes some causal concepts. Although my account does not presuppose any specific theory of causation, it presupposes that there are singular causal processes in the world.

Second, my account does not include a claim that all causal talk is to be understood as explanatory (*contra* Scriven 1975). Causal language and concepts are used in various ways. Notions of agency like manipulation, anticipation and responsibility are primary from the point of view of the development of our causal concepts. The explanatory use of causal language builds on these concepts, but it has some distinctive characteristics of their own. Although one could claim that all these uses of causal concepts build upon the idea of 'making a difference', it is too far fetched to claim that this makes them all explanatory uses. In this work I will only be interested in explanation. I will not develop any substantial theses about causation itself or about other uses of causal language.

A number of philosophers have emphasized the importance of considerations of causality in explanation. This has led them to develop causal theories of explanation. The most notable of these is the theory developed by Wesley Salmon (1984, 1998). A number of characteristics of my position can be illuminated by comparing it with Salmon's theory. Let us start with the observation that Salmon's theory is a causal theory of explanation, whereas mine is a theory of causal explanation. Two issues are related to this difference.

First, Salmon is trying to provide a general account of explanation in terms of causation. In contrast, I am only trying to provide a theory of causal explanation. I recognize that there are explanations that are not causal. Among such are mathematical explanations, philosophical explanations, explanations of a property instantiation and explanations of laws. It is true that my account builds on ideas that I claim can be used to make sense of other kinds of explanation, but I do not claim that causality is among these basic ideas.

Secondly, Salmon's methodological assumption is that one can develop a satisfactory theory of explanation by providing a theory of causation. My approach is the opposite. I think it is possible to provide an interesting account of causal explanation without committing oneself to any specific theory of causation. As Christopher Hitchcock (1995) has convincingly argued, the central problem in Salmon's approach, as it currently stands, is that it does not include any considerations of ex-

planatory relevance. The central argument against the covering-law theory has been that it allows irrelevant information to explain. The same charge can be raised successfully against Salmon's theory (Hitchcock 1995: 309-312). The point I am making is not that Salmon's theory of causation is false. Rather, I am arguing that it is incomplete: a theory of explanatory relevance is required to make it answer the questions that a theory of *explanation* is addressing. Here I intend to develop such a theory of explanatory relevance. This theory is intended to be fully compatible with the sort of theory of causation developed by Salmon (1984, 1998) and later by Phil Dowe (2000).

Now, although my theory is compatible with Salmon's theory of causation, I do not want to bind its fate to the fate of any specific theory of causation. Although, the idea of identifying causation with a physical process is attractive (cf. Braddon-Mitchell 1993), it is not wise to make such a commitment in one's theory of explanation. There are still good reasons to doubt that Salmon's (or Dowe's) theory of conserved quantities or other transfer theories of causation have successfully addressed all the questions that a theory of causation should address (cf. Kitcher 1989; Woodward 1989; Dowe 1995; Hausman 1998).

Secondly, the wide array of theories of causation suggested by philosophers suggests that they are not all addressing the same problem and that causation might be such a notion that it cannot be given a unified account. It seems that the principles governing the attribution of causes in different contexts of causal talk are not the same. For example, the principles used in the context of explanation are different from those used in contexts of attributing agency, responsibility, causal contribution or causal production. This suggests that the piecemeal approach I have adopted might work: let us concentrate on causal talk that is intended to be explanatory and see whether we can find any principles governing it.

I think that these arguments justify the use of my working hypothesis that one can make significant advances in analysis of causal explanation without a strong commitment to any individual theory of causation. As far as I can judge, my theory is compatible with various process theories of causation, among them theories advanced by Salmon (1998) and Dowe (2000). Similarly, my account is fully compatible with important theories of causation developed by J. L. Mackie (1974) and by Daniel Hausman (1998). All these theories accept my minimal assumption that there are singular causal processes.

What if, highly surprisingly, some sort of reductive analysis of causation turns out to be true? We have three possible candidates: an agency theory (cf. Menzies and Price 1993), a counterfactual theory (Lewis 1986, 2000) and some sort of probabilistic theory (cf. Davis 1988) that would have the regularity theory as its special case. I believe that the basic ideas of my account could be saved either with an agency theory or

with a counterfactual theory. As I will explain in Chapter 2, my account fits nicely with the idea that our explanatory concerns arise from the context of agency. This suggests that it can be modified to be compatible with an agency theory of causation. A similar observation applies to the counterfactual theory of causation. Since my theory is based on counterfactuals, accommodating it to a counterfactual theory of causation should not be too difficult. The probabilistic theory is more challenging. I would be forced to take back my critique of probabilistic versions of contrastive theory and modify the account developed in Chapter 2 to be probabilistic. These would be substantial changes, but the basic ideas of my approach to explanation would remain.

Interest explanation

Scientific controversies about explanation constitute an important test case for philosophical theories of explanation. A theory of explanation that does not provide any help in understanding and solving in real life controversies is of very limited interest. On the other hand, if scientists find the philosophical analysis appealing, and the analysis is fruitful in clarifying controversial issues, it has an argument in its favor. It suggests that the theory has captured something essential about the explanatory practice.

I have chosen interest explanation as a topic of my case study because interest explanations have an important role in social studies of science. Despite this importance, there has not been much theoretical discussion about interest explanation. This is a surprising observation, since interest explanation has been a controversial topic for the last 25 years. There is disagreement both about the legitimacy of this pattern of explanation and about its nature. This sort of situation provides a natural basis for a philosophical contribution.

Most problems related to interest explanation arise from not making explicit the theoretical background of explanation. In many case studies in social studies of science the intended *explanandum* is very vague. Sometimes it is even an open question whether the author intends to explain anything. Similarly, the theoretical apparatus doing the explanatory work is very imprecisely characterized. The unfortunate consequences of this practice are easy to see. People outside the field, for example philosophers of science, have had great difficulties understanding claims made by sociologists of scientific knowledge. This has led to the circumstance that many attacks on sociologists of scientific knowledge rest on false premises (cf. Laudan 1977, 1981; Roth 1987, 1996; Brown 1989). However the problems are not limited to outsiders. Insiders have also had trouble understanding interest explanations (cf. Woolgar 1981; Latour 1988a, 1988b; Pickering 1995). In some cases these problems derive from more general confusion about the nature of causal

explanation.

The general aim of my work here is to provide an account of causal explanation that can work as a reference point in these discussions. I believe that if causal explanation is understood as I present it, most arguments against it will lose their relevance. This would lead to a general recognition that social studies of science are indeed in the business of providing explanations. The more specific aim of the second part of this work is to apply my general ideas to interest explanation. I will concentrate on the explication of interest explanations, their *explananda*, and the underlying theories. I wish to spell out the explanatory model used in interest explanations in order to see what they can or cannot explain.

My discussion of interest explanation will be organized as follows. Chapter 5 will concentrate on the concept of interest and argue that sociologists and non-sociologists use the concept differently. I will also discuss the *explananda* that interest explanations address. Chapter 6 will illustrate the explanatory mechanisms behind interest explanations by discussing various examples. Finally, in Chapter 7 I will comment on some recent discussions of rational choice theory and functional explanation from the perspective of interest explanation. The aim of my discussion is to provide help in relating interest theory to current discussions in the philosophy of social sciences.

Chapter 5 begins by a general characterization of interest as a folk psychological concept. I argue that interests are closely related to agents' goals. They are *goal-dependent*: there are no interests without goals or aims. However, there are some important differences between goals and interests. I argue that it is possible that an agent is ignorant of her interests, a situation which is not possible with her goals. Second, interests are objective in the sense that once the goals are fixed, what is in an agent's interest is also fixed. The things that are in one's interest depend on causal facts about one's action environment, not on one's wishes or desires.

After a general characterization of the basic concept, I proceed to discuss how scientists use the concept of interest. My claim is that there are important differences between the ways scientists and sociologists of knowledge use the concept. Interests belongs to *the contingent repertoire* of scientists. This repertoire is used to explain away discrepancies between what is believed by the agent herself to be true or rational and what the other people claim to be the truth. This usage is asymmetrical. The beliefs that are believed to be true or rational are not thought to be in need of explanation, whereas 'false' beliefs are accounted for in terms of 'non-cognitive' factors like psychological idiosyncrasies, incompetence, social or career interests. Interests are understood as biasing factors that harmfully interfere with proper scientific conduct.

The sociological concept of social interest is much broader. In the everyday talk of scientists, the term interest is only used for goals that are non-epistemic. The sociologists of scientific knowledge do not equate social with non-epistemic, but treat it as a general category which includes both epistemic and non-epistemic factors. For them to say that research is influenced by interests means that it is goal-directed. It does not imply that such research is inadequate, unscientific, or biased. These sociologists aim to pursue the program of value-free social science consistently. In the case of study of science this means that they fully abstain from evaluating participants' positions. Evaluative concepts like truth, rationality, successful, or progressive, are treated consistently as actor's categories. Similarly, the concept of interest is completely lacking any evaluative connotations.

I also argue that no substantive theory about interests has yet been presented in the sociology of scientific knowledge, not even by supporters of the Strong Program. Apart from the epistemological thesis of instrumentalism, which says that knowledge-producing activities are goal-directed processes, detailed theories of interests in science are lacking. In practice, sociologists of scientific knowledge have used basically the same folk psychological concept of interests as everybody else.

Chapter 5 ends by going through some typical *explananda* of social studies of science. The idea is to illustrate how a wide array of *explananda* are related to interest explanations, and to show that contrastive analysis is of help in making these *explananda* more explicit.

Chapter 6 is dedicated to the discussion of examples of interest explanation. The idea is to explicate the explanatory mechanism by looking in detail at some exemplary interest explanations. On the basis of my discussion of these examples I will show that there is no unique pattern of interest explanation. I distinguish three forms of interest explanation that are based on different explanatory mechanisms. The first is based on the practical reasoning of an agent. The idea is that agent's choices are rational responses to the challenges her action environment sets for the advancement of her objectives. My examples include both social and professional interests as *explanantia* of choices by scientists. I also draw attention to the role of unintended consequences of action in more complex interest explanations.

The two other forms of interest explanation are intentional and non-intentional filtering explanations. In intentional filtering explanation, certain agents are able to control or filter the work of other agents in a way that is conducive to their goals. As a consequence, the actions of the controlled agents can be explained by interests of the controllers even if the two groups do not share the same interests. In non-intentional filtering explanations, the explanatory work is done by a cultural or social selection process analogous to natural selection. In such explanations, interests can have an important role in the selection envi-

ronment that determines the fate of beliefs and practices to be explained. My thesis is that these three patterns of explanation cover most of the informative uses of interest explanation in social studies of science.

The second part of Chapter 6 explicates the model of cycles of credibility first advanced by Bruno Latour and Steve Woolgar and later taken up by many others. I argue that this model provides a theoretical elaboration of the sociological notion of professional interest. Since the model serves as a theoretical background for many sociological studies of science, I put some effort into its explication. The model depicts scientists as entrepreneurs trying to maximize their scientific capital, credibility. The model can be used both to explain some general characteristics of modern academic science and to understand the relationship between social and epistemic goals in scientific work. Especially, it can be used to show that sociological explanations of science in terms of professional interests are not explanations by purely non-epistemic goals.

Chapter 7 aims to deepen the understanding of interest explanation by comparing 'interest theory' with rational choice theory. After briefly discussing some of the differences between these approaches, I concentrate on two suggestions aimed at broadening the explanatory scope of rational choice theory. The first suggestion, made by Philip Pettit, contains an idea of interests as standby causes. I will argue that Pettit's 'rational interest theory' makes stronger motivational assumptions than the standard interest approach. Furthermore, I argue that Pettit's valuable idea of resilience as an *explanandum* can be separated from his thesis of the relative importance of self-regarding motives in human action. The second suggestion, made by Debra Satz and John Ferejohn, proposes an externalist interpretation of RCT. In this interpretation, preferences are understood as descriptions of the structures of interaction rather than of the subjective mental states of individual agents. The externalist interpretation is interesting since it seems to fit quite nicely with the standard explanatory practice in the RCT tradition. Secondly, it describes RCT as structural, non-individualistic, social theory.

In the final section I will discuss Jon Elster's critique of functional explanation in the social sciences. This discussion is relevant since Elster's critique threatens the legitimacy of the pattern of non-intentional filtering explanation sketched in Chapter 6. I demonstrate that Elster's critique is flawed in a number of ways, and I therefore argue that explanations referring to non-intentional filtering mechanisms are indeed legitimate causal explanations in the social sciences.

Part I

Understanding singular causal explanation

Chapter 2
A contrastive theory of causal explanation

In this and the following two chapters, I will develop my version of the contrastive theory of explanation. I will restrict my discussion to singular causal explanation, but the basic ideas and the arguments can easily be applied also to other kinds of explanations. I concentrate on singular explanation for two reasons. First, I intend to apply this account of explanation to explanations in the social studies of science. These explanations are typically explanations of singular occurrences. Second, the fact is that most philosophical discussion of explanation concentrates on singular causal explanation. Applying my ideas about explanation first to singular causal explanation makes it easier to compare them with other approaches and theories. If this is successful, it should be possible to extend the account to other kinds of *explananda*.

In this chapter, I will present the basic components of my account. After describing some philosophical starting points of my approach, I will begin by introducing the idea of contrastive *explanandum* and then elaborate my own version of it. I start by presenting the intuitive idea of contrastive questions, and then proceed to distinguish the *explanandum* of singular causal explanation from other contrastive *explananda*. The development of my account begins with the introduction of a technical notation to indicate contrasts in an unambiguous way. Then I proceed to distinguish between various senses of explanatory failure. I also discuss typical criteria for the choice of a contrast and suggest a novel way to see the difference between scientific and everyday explanatory questions. Finally I argue against Peter Lipton's suggestion that contrasts in explanation-seeking questions do not need to be incompatible.

In section 3, I will discuss the major criticisms presented against contrastive theories of explanation in order to further clarify my position. I argue that all *explananda* can be analyzed as contrastive and

that this is a fruitful approach in understanding explanatory questions. I also argue that the contrastive thesis should be understood as a claim about what an explanation *can* explain, not as a thesis about what the *explainee* has in her mind. Finally, I defend the thesis that a contrastive *explanandum* can be reduced to a non-contrastive *explanandum* against the arguments presented by Dennis Temple and John W. Carroll.

In the fourth section, I will develop a counterfactual theory of causal explanation to accompany my views on contrastive *explananda*. I will start by considering some problems related to the existing counterfactual theories of explanation. I argue that the contrastive approach has certain advantages over the more traditional non-contrastive approach in dealing with the problems concerning the choice of the right level of description, the choice of explanatory information, and the fixing of the causal field presupposed by the explanatory claim. After briefly suggesting that the counterfactual approach to causal explanation offers a nice way to handle the problem of negative causes, I proceed to discuss the presuppositions of a sensible contrastive question. I first review the suggestions found in the existing literature, and then present my own idea, which emphasizes the importance of how-questions underlying the causal why-question. I argue that a causal explanation presupposes an account of a mechanism that ensures that the fact to be explained occurs instead of its contrastive alternative. The idea of a mechanism also helps in distinguishing between explanatory and symptomatic factors and in solving problems related to causal preemption and overdetermination.

1. Three philosophical starting points

The account of explanation presented in this work rests on three philosophical assumptions. As these ideas serve as the starting points of my discussion, it is useful to describe them briefly. I will not argue extensively to support them, since I do not take them to be very controversial: most writers in the philosophical literature would subscribe to them, and these ideas also agree with our common sense.

The first assumption is that *the aim of explanations is to track objective relations of dependency in the world*. Let us call this *the realist conviction*. According to it, explanations are about the things in the world. The realist conviction is incompatible with fully subjective accounts of explanation, but it does not deny that explanations also have some subjective or pragmatic elements. Especially, it does not require, for example, equating causation and causal explanation. The conviction simply states what explanation is about.

There are various kinds of dependency and, consequently, various kinds of explanation. I am not among those who claim that all explana-

tion is causal explanation. When we are explaining events, properties or laws, we appeal to different kinds of dependencies, as the metaphysical relations within these categories vary. For example, explaining why a glass broke is different from explaining what makes it fragile. In the former we are explaining an event and looking for its causes. In the latter we are explaining a property instance, and we are inquiring as to the constitution of the glass in order to account for its fragile nature. This explanation is obviously not causal. The molecular structure of the glass does not cause it to be fragile; it constitutes its fragility. In this work I will concentrate on causal explanation, but the same basic ideas about explanation can be used also in these other forms of explanation.

My second basic assumption is that *an explanation can only relate things described or conceptualized in a certain way*. In other words, we always explain specific aspects of events or phenomena, and not these events or phenomena themselves or on the whole. This idea traces back at least to Aristotle (Ruben 1990: 87-88), and most philosophers of explanation have subscribed to it, most notably Carl Hempel. This makes it natural to think that the *relata* of an explanation are facts.[1] This feature distinguishes causal explanation from causation.[2] I follow a strong philosophical tradition in holding that causation is a natural relationship that holds true in the world between particular circumstances or events independently of our conceptualization. The concept of causation *refers* to processes taking place in the world. As causality is a natural relation, simple claims about causation are not sensitive to the way the cause and the effect are described. It suffices if the description succeeds in picking up the right event or circumstance as a cause, it does not matter how it is described. (Strawson 1985.) If we are making a simple, non-explanatory, causal claim it does not matter whether we say 'the explosion caused the waterpipe to break', or 'the event reported in *Helsingin Sanomat* caused the event that made Jim wet'. The only thing that matters is that these two descriptions refer to the same event. Notice that this insensitivity to the description concerns only causal claims that are about participation in causal relations or processes. In much of our causal talk we are not making these simple causal claims. Often our causal claims are intended to be explanatory, and this changes the situation radically.

In explanation, it is important that causes and effects are described appropriately. Causal explanations face *the problem of explanatory selection* (Hesslow 1983). We have to pick the right aspects of the causal process to be included in the explanation. Not all causal information is explanatory information. Usually, the causal history of an event includes a vast number of elements and aspects that are not explanatorily relevant to the explanation-seeking question we are addressing. We want only the items that make a difference in the things that we are interested in. But the problem of explanatory selection is wider than that,

for it also requires that the items we have picked as explanatory must be described in the right way. We can describe things in various ways and at various levels of abstraction, and the challenge is to find the right way and the right level for the explanation at hand.

It can be speculated that our explanatory preferences stem from our nature as active interveners in natural, psychological, and social processes. We want to know where to intervene to produce the changes we want, and this knowledge often presupposes answers to some why- and how-questions. Without this knowledge we would not know when the circumstances are suitable for our intervention. We would not be able to predict the results of our interventions. I am not claiming that we can reduce the notion of explanation to its instrumentalistic origins. We also want explanations for things that we cannot manipulate. The point of this speculation is not to relate our curiosity directly to our instrumental needs, rather it is to explain why our explanatory preferences are as they are.

The third assumption is that *an explanation is an answer to a question*. The idea is that an explanation is adequate when it answers the explanatory question satisfactorily. Various writers have used this idea to get hold on the concept of explanation. Some, like Hempel (1965), have used it as an informal starting point for their discussion, whereas others have built whole theories of explanation around it.[3] I believe that, treating explanations as answers to question offers a natural way of specifying the object of explanation.[4] This idea will work as my heuristic in the following. I will not commit myself to any specific theory about explanation-seeking questions in this work, and I have very little to add to the existing analyses of explanation-seeking questions.

Some advocates of the contrastive approach, for example, van Fraassen (1980), subscribe to the thesis that all explanatory questions are always *why*-questions. My theory does not include any such commitment. The thesis that explanations are answers to questions should be kept separate from the thesis that all explanation-seeking questions are *why*-questions. As I see it, the same explanatory request can often be made using various linguistic devices. (Scriven 1959: 451; Faye 1999: 68-72.) For example, in some cases a how-question is a more natural way of making a contrastive explanatory request than a why-question. From the point of view of my account, it is not essential that every explanation-seeking question be a why-question or that all explanation-seeking questions can be paraphrased as why-questions. As Markwick (1999: 191) has noted, there is no deep commitment to any specific question-theoretical approach among most supporters of the contrastive theory of explanation. I will continue this tradition of non-commitment.

2. The contrastive explanandum

A famous joke about the bank robber Willie Sutton introduces the basic idea of contrastive explanation. When Willie was in prison, a prison priest asked him why he robbed banks. Willie answered, "Well, that's where the money is." The joke is based on a confusion, for Willie was not answering the question the priest was asking. The priest had in his mind the question: "why do you rob banks, instead of leading an honest life?", whereas Willie answered the question: "why do you rob banks, rather than gas stations or grocery stores?" This is the basic insight of the contrastive approach. We do not explain simply 'Why *f*?' rather, our explanations are answers to the contrastive question 'Why *f* rather than *c*?'. (Garfinkel 1981: 21-22.) Instead of explaining plain facts, we are explaining contrastive facts. Several philosophers of explanation have used the same basic idea.[5] I will follow their lead and try to develop the contrastive idea a little bit further.

The idea that explanations are contrastive is a natural consequence of my basic assumptions. If one considers how to trace relations of dependency in terms of explanatory questions, one easily comes up with the idea that explanations are answers to 'what-if-things-had-been-different' questions (Woodward 1993: 252). We wish to know how the change in the causes brings about the change in the effects. We want to know what makes the difference and then to leave out factors that do not. Furthermore, if there is no change in the effects when we make changes to our putative explanatory factors, then we do not have truly explanatory factors, since there is no appropriate relation of dependency.

The contrastive idea works nicely with our preferred form of causal inquiry: the scientific experiment. When we are looking for explanations using the methods of experimental inquiry, we are basically working in a contrastive setting. For example, we contrast the control group with the experimental group or the process after the intervention with the process before the intervention. In both of these cases, we are trying to *account for the differences* between the outcomes. The basic idea is to keep the causal background constant and bring about changes in the outcomes by carefully controlled interventions. The same contrastive setting also works in comparative research, which is our second option if experiments are impossible. Clearly, by adopting the contrastive idea, we are starting with a very intuitive and central feature of our cognition.[6]

The advantage of the contrastive approach is that it allows us to be specific about the *explanandum*. As a consequence, it can be profitably used in the analysis of apparently competing explanations. As I will show in subsequent chapters, many apparently competing explanations are in reality complementing explanations.

Forms of contrastive explanation

It is important to note that there are various forms of contrastive explanation, depending on the nature of the *explanandum*. These various forms share two basic ideas of the contrastive explanation. In all of them, the explanation traces objective relations of dependence and is seen as an answer to a contrastive question. These forms differ in terms of their *explananda* and in terms of the kind of dependence relationship they are tracking.

First, there is singular causal explanation, which is the topic of this work. In singular causal explanation we are explaining facts about events in terms of facts about the earlier development of the causal process in question. For example, we explain facts about a car accident by referring to the facts about events and circumstances that occurred before the accident. The explanation aims to select the relevant facts from the causal history, the measure of relevance being their contribution to the difference between the fact and its foil. I will return to this explanatory setting shortly.

Second, there is a singular explanation for an instantiation of a property. For example, we might explain the fragility of a glass by referring to some facts about its molecular structure. This explanation does not directly refer to causal processes, or causal dependency, between molecular structure and fragility, but to the relations of dependency between the properties. The fact that a glass has a certain molecular structure constitutes the fact that it is fragile. Certain facts about the molecular structure make the difference between being fragile and being something else. They also determine the specific way in which the glass is fragile. As all facts about the molecular structure of the glass are not relevant to its fragility, we have a similar problem of explanatory selection as in the case of singular causal explanation. Furthermore, as there might be, and probably are many different ways to constitute fragility, our singular explanation of property instantiation is not necessarily a general explanation of the property of fragility. (For an account of this kind of explanation, see Cummins 1983, Bechtel and Richardson 1993, and Ylikoski 1997.)

Regularities and laws are the third important class of *explananda* where the contrastive model of explanation is applicable. In such cases we are interested in the dependence between regularities and more fundamental laws and mechanisms. The laws or regularities to be explained are the way they are because the more fundamental laws and mechanisms are the way they are. The contrastive approach also works here: when explaining laws and regularities, we are explaining why they are the way they are, rather than otherwise.

This short survey of forms of contrastive explanation is not intended to be exhaustive.[7] The central point here is that the intuitions behind

the contrastive approach are general intuitions about explanations, and not *ad hoc* specifications made to suit my theory of singular causal explanation. The fact that the same basic pattern of explanation applies also outside causal explanation gives support to the proposed analysis of causal explanation.

Indicating the contrast

There are various linguistic means to indicate the contrast in an explanation-seeking question (van Fraassen 1980: 128; Garfinkel 1981: 25, 29; Sober 1994: 176). For example, in English we can ask:

Why *f* rather than *c*?

Why *f* and not *c*?

Why *f* instead of *c*?

Given that *f* or *c*, why *f*?

The contrast can also be indicated by the combination of emphasis and non-linguistic contextual cues. Sometimes the contrast is so obvious from the context that there is no need to indicate it at all. The existence of alternative linguistic means even within a single language suggests that there is no single privileged way of indicating the contrast. (*contra* Carroll 1997, 1999.)

In order to simplify the presentation and to avoid possible problems of generalizing the curiosities of one particular linguistic way of indicating the contrast, I will denote the contrastive *explanandum* by *f* [*c*]. Here *f* is a fact, and *c* is a non-compatible alternative (a foil) to *f*. This notation allows one to state the intended *explanandum* more clearly, or at least more economically. The usual linguistic devices can be ambiguous and clumsy in complicated situations that arise in philosophical discussion. The use of technical notation also underlines the difference between understanding the structure of the *explanandum* on one hand, and the pragmatics of locutions like '… rather than …' in English language on the other hand.

To ease the discussion I would like to introduce some other technical terms. Some *explananda* have more than one contrast. In such cases we will be speaking of *a contrast class*. It can be denoted by a list of contrasts, *f* [*c*, *c**, *c***, …], or by a class of contrasts *f* [*Cc*]. The listing of contrasts is more informative, but in some cases it is simply inconvenient. For example, *Cc* may include all the determinants of the determinable *X*, for example color, except the determinant *f*, for example, a certain hue of red. As there is an innumerable number of different shades of colors, it is impossible to list them all. It is practical to have both ways of indicating the contrast available.

The *explananda* with more than one contrast can always be un-

derstood as clusters of simpler *explananda*. The *explanandum f* [*c*, *c**, *c***, …] can always be partitioned to more simple *explananda*: *f* [*c*], *f* [*c**], *f* [*c***], …. The fundamental unit of explanation is then the simple contrastive structure: *f* [*c*].

Some of the members of the contrast class might be unnecessary or *redundant*. This is the case when the same *Ea* (adequate explanatory answer) explains both *f* [*c*] and *f* [*c**]. In cases of redundancy, the redundant contrasts can be dropped from the contrast class without causing any harm.

To make it easier to compare my account of explanation with some other accounts of explanation, the concepts of *complete explanation* and *ideal explanatory text* can be characterized. When the *Cc* include all the possible contrasts for fact *f*, and we have given an adequate explanation to the question *f* [*Cc*], we have *the complete explanation of fact f***.** So, the complete explanation of *f* involves explaining why *f* against all possible contrasts. In practice this means describing the whole causal history of *f*. This is already an unrealistically demanding task, but we can have an even more ambitious goal. This goal can be characterized with the concept of the ideal explanatory text introduced by Peter Railton (1980, 1981). When we have adequate explanatory answers for all *f* [*Cc*] which are true about the particular event or phenomenon *e*, we have *the ideal explanatory text for e*. The ideal explanatory text literally explains everything about *e*.

I don't claim that my definitions fully reflect the intuitions of the theorists who use these two concepts. It might be that the contrastive vocabulary can not capture the intended essence of the concepts. However, I hope that these stipulative definitions can at least elicit more positive characterizations of these vague concepts, as this could allow a comparison of the different theories of explanation. The basic point is that there are various ideas of *completeness of explanation* that are not clearly distinguished in the literature. With these characterizations it is possible to start discussing the aims of different theories of explanation, after which we can then proceed to draw conclusions concerning their compatibility. However, this will be a topic of some other study.

Explanatory adequacy and failure

We have *an adequate explanation* of *f* [*c*] when we have explained why *f* rather than intended contrast *c*. An explanation is inadequate when it does not explain *f* [*c*]. An explanation can fail in various ways (Lewis 1986: 226-228). We can distinguish between a broad and narrow sense of explanatory failure. My interest in this work is in the narrow sense. An explanation fails in the narrow sense due the fact that the offered explanation does not fulfill the requirements of an adequate explanation. These requirements will be characterized in the latter part of this

chapter. Before proceeding further, it useful to take a look at failures in the broad sense. In the broad sense an explanation can fail in two different ways.

First, the explanation might provide misinformation. It can claim things that are not true. Let us call this *a factual failure*. Although in practice it can often be very difficult to determine the facts of the matter, the case of misinformation is not a big theoretical problem for explanation theory. A distinction between a possible explanation and a true explanation can be useful for avoiding confusion. A possible explanation satisfies all the other criteria of a good explanation except for the truth requirement. It fails for purely factual reasons. If it were true, it would explain the *explanandum*.

Another kind of factual failure is the failure to correct incorrect presuppositions of the explanatory question. In such cases the explainer fails to point out to the recipient that her question rests on premises that are not true. This failure can also be classified as an example of pragmatic fallacy, but I think it is clearer to treat it as an example of factual failure.

The second category of explanatory failure in the broad sense is *pragmatic failure*. In these cases the explainer does not provide what the recipient of the explanation wants. In one way or another, there is a communication breakdown between the explainer and the recipient. These failures are similar in all forms of communication, and consequently, they are not unique to the communication of explanatory information.

Again, a pragmatic failure can occur in a number of ways. First, the explanation can answer the wrong question. The explanation could be perfectly good, but it does not address the question that the recipient wants to be answered. It might be that the recipient already knows that answer or she simply does not care about that particular question. Second, the explanatory information might be in a form that the recipient cannot understand. The explanation might be so technical, or the vocabulary so full of jargon, that the recipient cannot cope with it. The explainer might also presuppose background knowledge that the recipient does not have, which leads to a failure to understand. The third way to fail pragmatically is to provide the explanatory information in such a form that the recipient cannot separate the explanatory information from all the other information provided. In such cases, the explainer provides the explanatory information and so much other information that the recipient cannot disentangle them. A similar failure happens when I ask for Frank's telephone number, and someone provides me all the numbers in the telephone book.

Factual and pragmatic failures do not interest me here. In the following, I will be considering explanations that do not fail for factual or pragmatic reasons. What I am looking for is *explanatory adequacy*. There is more to the explanation than that the recipient is satisfied with it, as

the extreme pragmatic theory of explanation would have it. It is possible that the recipient is *wrong* in accepting a certain answer as an explanation. Although the question is wholly up to the recipient, the adequacy of the answer is not.

An explanation can also fail by being *partial*. In such a case, the provided information is explanatorily relevant and true, but the explanation needs to be supplemented to be fully adequate. This notion will become important in Chapter 3, where probabilistic explanation is discussed. However, its use is not limited to probabilistic contexts.

How do contrasts arise?

I will start this discussion of the nature of contrasts with Eric Barnes' (1994: 37) warning about taking a linguistic approach to the generation of contrasts. He notes that most writers on contrastive explanation have used substitutional transformation of sentences to generate the contrast classes.

> Halonen won the 2000 Finnish presidential election
> [Aho won the 2000 Finnish presidential election,
> Hautala won the 2000 Finnish presidential election, …]

This way of presenting contrasts is obviously both pedagogically and stylistically practical, but it can give a misleading impression that this is the right or the only way to generate contrasts. For example, the following two suggestions are sensible contrasts, but they cannot be generated by substitutional transformation:

> Halonen won the 2000 Finnish presidential election
> [The 2000 Finnish presidential election ended in a tie]

or

> Halonen won the 2000 Finnish presidential election
> [The results of the 2000 Finnish presidential elections were nullified]

Clearly, our focus should be on the contrasted states of affairs, not on their linguistic representations. But how do we arrive at these contrasting states of affairs? There is more than one way to form contrasts. The contrast can arise either by imagination or by comparison, or by a combination of the two.

In everyday life, explanatory questions most often arise as a consequence of an abnormal or unexpected incident. Something that is not normal happens and raises our curiosity. We want to know why it happened. In such situations the choice of the contrast is obvious: we will contrast the abnormal occurrence with the normal or expected state of affairs. (Hart & Honoré 1959: 31-38.) In his important 1983 paper, Germund Hesslow distinguishes five different ways of forming the contrast. It is useful to summarize these cases, since they illustrate various senses of 'normality' that are used in generating the contrasts for our

explanations.

First, the contrast can be *the statistically normal* case. For example, when we ask why a particular barn caught fire, we are asking what distinguishes the barn under consideration from other barns. And when we consider these other barns, which are made of similar materials, placed similarly and used similarly, we find that most barns of this type have not burned. This fact is our contrast. We will be looking for causal factors that are present in the case of our burned barn, but not in the statistically normal case. (Hesslow 1983: 95.)

Second, the contrast can be *the temporally normal* case. When we ask why the barn caught fire at some particular time, we are comparing the time of the fire with the barn at earlier times. So here we are not comparing different but similar objects, but the same object at different times. In this case we will be looking for some changes in the conditions to account for the change in the states of affairs we are interested in. (Hesslow 1983: 95.)

The third possible contrast is *a theoretical ideal*. Here the contrast does not arise from the observation of a difference, but from the predictions or assumptions of our theory. A theoretical account gives us a kind of 'default' contrast. The use of this kind of contrast facilitates the systematization of the field covered by the theory. Hesslow mentions Max Weber's 'ideal types', the equilibrium models of the perfect market in the neo-classical economics, the definition of a 'wild type' in genetics, and the physiology of the healthy organism in medicine as typical examples of such theoretical ideals. (Hesslow 1983: 95-96.) These all provide scientists with a standard of comparison that helps in picking the things to be explained. Hesslow also compares theoretical ideals with what Stephen Toulmin calls ideals of natural order. Toulmin writes:

> Our 'ideals of natural order' mark off for us those happenings in the world around us which do require explanation, by contrasting them with 'the natural course of events' – i.e. those events which do not. Our definition of the 'natural course of events' is therefore given in negative terms: positive complications produce positive effects, and are invoked to account for deviations from the natural ideal, rather than conformity to it (Toulmin 1961: 79).

Since both Hesslow and Toulmin are very brief in characterizing them, it is difficult to say whether their concepts are the same, but at least one can say that their function in the context of explanation seems to be the same. Both work as generators of contrasts for scientific explanatory questions.

The fourth source of a contrast for Hesslow is that which is *subjectively expected*. Here the contrast is what the agent was expecting to happen. These explanations show how the fact to be explained could

have occurred against the expectations we had on the basis of knowledge of earlier conditions. (Hesslow 1983: 96.) This source of contrast is interesting because it can be related to the intuition behind the original covering-law theory. The intuition in question is that the function of the explanation is to make the *explanandum* expected. (Hempel 1965: 337.) Hempel later interpreted this intuition as a requirement that the explanation must make the *explanandum* highly probable. The contrastive analysis gives an alternative way of interpreting this intuition: the explanation is related to our expectations because typically we are explaining facts that do not accord with our expectations. Showing why the unexpected happened corrects our background beliefs and in this sense reduces the unexpectedness of the *explanandum*. If this is right, we can give up the requirement that the explanation must make the *explanandum* highly probable, without losing the intuition that, at least sometimes, the function of an explanation is to reduce surprise.

The fifth possible source of a contrast is *a moral ideal*. Sometimes an action is contrasted with a normative account of how a person should have acted. In such a case we are asking for the explanation for the deviance from this standard of conduct, and we choose as explanatory causes such conditions that should not have been present. (Hesslow 1983: 96-97.)

I don't think that we should take this as a complete list of all possible ways in which a contrast can arise. Certainly there are also other ways of generating contrasts. It is essential to see that there are various ways in which a contrast can arise, but having an exhaustive list of all the ways they can arise is not. However, there is one way in which Hesslow's list can be misleading. All the *explananda* in his list are either abnormal or unexpected. This goes nicely with our everyday practices of explanation. However, we sometimes also want to explain the normal case. For example, we might want to explain why grass is (normally) green rather than red. The contrastive approach works here also. It might be that *explananda* in which the fact is the normal case and the foil abnormal are quite rare, but this does not reduce the legitimacy of such questions. At least some of these questions are sensible.

Indeed, this observation suggests one way of characterizing the difference between everyday reasoning and science. The difference is in the explanatory questions asked: in science we also try to explain the normal case, whereas in everyday reasoning we are only interested in explaining the abnormalities. Everyday reasoning takes the normal course of events as granted and only wants to account for a deviation. Usually, we are not able to explain the normal course of events, but we are not bothered by this. But if something unusual happens, we want to know why. A deviance calls for an explanation. Scientists also explain deviance, but they want to know also about the normal case, and this can be reconstructed as turning around of the usual way of using contrasts. In this sense they are like children.

The question of compatible contrasts

Peter Lipton (1990) has claimed that not all contrasts are incompatible. As this thesis puts my analysis of the contrastive *explanandum* in danger, let me consider whether Lipton's thesis stands up to critical scrutiny.[8] Recall the famous example of paresis. Paresis is a form of neurosyphilis, and no one contracts this dreadful disease unless he had latent, untreated syphilis. However, the evolution of the disease is unknown, and only a small percentage of those who have untreated syphilis get paresis. Now we have two persons, John, who has had latent syphilis and who now has paresis, and his brother Tom who does not have latent syphilis. We can explain why John, rather than his brother, contracted paresis, for only he had syphilis. This is something that everybody accepts. However, Lipton notes that the fact that John has syphilis is compatible with the possibility that also his brother Tom has syphilis. These two facts are not incompatible. It just happens that Tom does not have syphilis. Lipton concludes that in this case the fact and its foil are not incompatible. This state of affairs is not in any way related to the curiosities of this particular example. The situation is quite common: we want an explanation for the difference between two apparently similar situations which turned out differently, but which might have turned out similarly. (Lipton 1990: 250.) This is typical in cases of explaining differences, so if Lipton's thesis is right, we might expect that a good many contrasts are compatible.

However, I do not think that we should accept Lipton's thesis. As he notes in his 1993 paper, the reference to the brother is a surrogate for a counterfactual claim about John. We are really interested in John's illness, not his brother's health. (Naturally, the question can also be about Tom, and not John, and the same points will apply.) We are asking why John has paresis rather than being like Tom, who does not have paresis? The exclusive alternative in the *explanandum* is John's not having paresis, not his brother not having it. It is easy to see that the reference to the brother offers a convenient way of picking the desired contrast about John's health, and the explanation is the thing that makes the difference between the two cases. But it is important to remember that we are contrasting two causal scenarios of John's health. The first scenario culminates in John's having paresis, and the second in John's being like Tom, that is, without paresis. When subjected to a more careful analysis, the apparently compatible contrasts turn out to be incompatible.[9]

My claim is that all apparently compatible contrasts turn out to be incompatible when inspected more carefully. I do not deny that we can have completely sensible contrastive statements in which the contrasted states of affairs are compatible. We can also have contrastive questions that have compatible contrasts. My point is about the kinds of explanation-seeking questions that can have an *explanatory* answer. The ques-

tions have to be reconstructed in the above manner in order to give them a properly contrastive answer. Otherwise the explanatory counterfactual could not do its job. There are some presuppositions for a sensible contrastive explanation-seeking question, and I will elaborate these presuppositions shortly.

Here we can see how looking too closely at the linguistic form of the contrastive statement can lead to a misguided analysis. The basic idea in the contrastive approach to explanation is to look for the implied contrasts, instead of being satisfied with the usual statement of *explanandum* (which is often non-contrastive). The same approach should be used here: one should not be satisfied with just any contrastive statement. Instead, one should look behind linguistic formulations and try to capture the real contrast.

3. Arguments against the contrastive idea

A very common way to deny the philosophical relevance of contrastive questions in theory of explanation is to claim that not all explanations are contrastive (Ruben 1990: 40; Humphreys 1989: 137). The point implied by this claim is that a philosophical account of explanation need not concern itself with the contingent features of some explanations. Since only some explanations are contrastive, the contrastive approach does not seem suitable for analyzing explanation in general. This position denies that the contrastive idea could have any heuristic value in developing an account of explanation. If this is indeed the point of the critics, the contrastive approach can defended by showing that it can produce interesting results.

But what about the thesis that not all explanations are contrastive? Responding to this critique is somewhat tricky. The challenge assumes that the supporters of contrastive approach are making universal statements about all *explananda*. I have not found any writer on this topic who is committed to this bold claim. For example, Lipton claims to be agnostic about the issue (Lipton 1990: 261). The reason for this is easy to see. Besides the logical problems with proving a universal statement, there is a problem concerning the vague boundaries of what counts as an explanation and what not. People do have conflicting intuitions about the explanatoriness of a good many explanations, and often the putative counterexamples to the contrastive thesis belong to this heterogeneous class of explanations.

I think that a more fruitful approach to understand the contrastive approach would be to interpret it as claiming that *all explananda can be analyzed as contrastive*. If the contrastive analysis is generally applicable and if in most cases it provides fruitful results, we have an argument for it. I hope that this work, together with earlier literature on contrastive explanation, shows that interesting results can indeed be

achieved by the use of the contrastive approach. It might be that in some special cases it does not help much, but this remains to be shown. And, of course, the critics can try to come up with examples where the use of the contrastive idea is a hindrance to the analysis of explanation. I have not seen such examples yet.

The position I am taking can be further clarified by comparing two ways of understanding the contrastive approach to explanation. The first attaches the contrastive idea to a pragmatic theory of explanation. In this approach the contrastive thesis is about what people really have in their minds when they present explanation-seeking questions. The explanation-seeking question is thought to reflect the *explainee*'s epistemic state, and the contrastive suggestion is understood as a way of specifying what the *explainee* wants to know. I take this to be 'the mainstream approach' in the literature on contrastive explanation. Its best known representative is Bas van Fraassen (1980).

Dennis Temple has noted a problem with this position. He writes: "… in many cases a speaker who asks 'Why P?' is simply puzzled about P, and without having any particular contrary in mind" (Temple 1988: 147). Temple has a point. Sometimes we do not always have any specific contrast in mind. For example, the *explainee* can be confused, or she might want answers to many different questions at the same time. Of course, the supporter of the pragmatist approach can reply that his theory is about an ideal or a rational *explainee*, and not about ordinary people, who are often confused. This is a strange way to defend a *pragmatic* theory of explanation.

The defender of the mainstream approach can respond by pointing out that asking for a contrast is a natural and effective way of clarifying or improving the intended *explanandum*. We can ask the *explainee* to specify her request by suggesting possible contrasts, such as: "Do you mean why *f* rather than *c* or why *f* rather than *c**." This is a useful and a very common way to settle the question. (Hesslow 1983: 94.) This is a good point, but I think we do not need the mainstream pragmatic approach to make it.

The alternative interpretation, which I support, takes the contrastive thesis to be a central contribution to a theory about what an explanation can achieve. It is not concerned primarily with the *explainee*'s epistemic states, but *with the things that a normal explanation can explain*. It asks, given that somebody has provided an explanation, what it explains or which question it answers. It does not make claims about the usual format of the explanation-seeking question, but about the questions for which our explanations could be satisfying answers. We should allow that people could be confused or just simply unclear about their intended *explananda*. The contrastive thesis should be about what an explanation can explain, not about what kind of questions people have, or can have, in their minds.

In this alternative account, it is pragmatic and contextual factors that determine which questions we want to ask or which contrasts we choose, but after fixing the explanatory question the adequacy or inadequacy of the given explanation is a non-pragmatic matter. The theory of contrastive *explanandum* is not a pragmatic theory of explanation in any interesting sense. (Hesslow 1983: 97-98; Woodward 1993: 276.) Of course, it can be naturally extended with pragmatic components, but that is a different matter. A complete theory of explanation should include also pragmatics, but a theory of explanatory adequacy need not do that.

Are contrastive explananda reducible?

The contrastive approach has also been criticized by claiming that a contrastive *explanandum* can be reduced to a non-contrastive form. Dennis Temple and later John W. Carroll have suggested that 'Why *f* rather than *c*' is equivalent to 'Why *f* and not-*c*' (Temple 1988; Carroll 1997, 1999). Temple claims that a consequence of this reduction of the contrast is that the contrastive approach has no advantage over the traditional 'propositional approach', which sees the *explanandum* as a (non-contrastive) proposition. Let us call Temple's position 'the conjunctive view'.

Temple makes it sound like the contrastive approach has nothing new to say about the *explanandum*. This is not true. The traditional *explanandum* has been the plain '*f*', but now if we accept that the right representation of the *explanandum* is the conjunction '*f* and not-*c*', we have made a substantial point about explanation. Earlier it was not thought that the *explanandum* is complex in this way.

There are two ways of reading Temple's suggestion. The difference between these readings is whether we accept the following three propositions to be equivalent:

(1) '… explained the fact that *f* and not-*c*'

(2) '… explained the fact that *f* and explained the fact that not-*c*'.

(3) '… explained the fact that *f* rather than *c*'

The weak reading accepts that (1) and (3) are equivalent, but it does not accept that (2) is equivalent to them. The strong reading accepts that (1), (2), and (3) are all equivalent. The non-conjuctivist position naturally denies all claims of equivalence between these propositions.

Let us first take a look at the strong reading of the conjunctive view. Although neither Temple nor Carroll says it, this reading is really a *reductio ad absurdum* of the contrastive approach. It claims that contrastive explanations are really conjunctions of two non-contrastive explanations. This would make contrastive *explananda* quite superficial phe-

nomena. This is ironic: the contrastive suggestion was originally coined to make sense of our ordinary way of explaining 'why *f* ?'. But now it turns out that we cannot do that because the suggestion presupposes that we already know how to answer this question.

Does this reductive thesis hold water? There are two principal arguments against it. The first argument was advanced by Peter Lipton, who has pointed out that explaining 'Why *f* rather than *c* ?' requires less than explaining 'Why *f* and not-*c* ?' (Lipton 1990: 252-253). Recall the example of John's paresis. No one contracts this dreadful disease unless he has latent, untreated syphilis. However, the evolution of the disease is unknown, and only a small percentage of those who have untreated syphilis get paresis. Now, we can (fully) explain why John, rather than his brothers, contracted paresis, for only he had syphilis. However, we cannot explain why he, among all syphilitics, got it. Under the assumptions of the example, we don't know why some, but not all, with untreated syphilis contract paresis. The strong reading of the conjunctive view would require that we first explain why John contracted paresis, which we cannot do, and then to explain why his brothers did not contract it, which we can do. So, with the conjunctive view we cannot explain something that we intuitively think we can explain. This suggests that the strong reading does not work. We cannot infer (2) '… explained the fact that *f* and explained the fact that not-*c*' from (3) '… explained the fact that *f* rather than *c*'.

The second argument is by David Ruben. The acceptance of the strong reading requires that there are no limitations on possible contrasts. This can be seen by considering any arbitrary *f* and not-*c*. Let *f* be 'snow is white' and let *c* be 'grass is red'. Suppose that I explain both *f* and not-*c*. Have I then explained the fact that snow is white rather than grass is red? Clearly something is missing here. We presuppose that there is some sort of relevance between *f* and *c* when we contrast them, but the normal truth-functional 'and' does not include any considerations of relevance. (Ruben 1990: 42.) The strong reading does not respect the requirements of relevance of the '… rather than…' locution, which makes it a failed reduction.

But what about the weak reading? It does not accept the equivalence between (1) and (2), and so Lipton's and Ruben's arguments cannot refute it. However, these arguments show that the weak reading is of very limited interest. Ruben's argument shows that if we rephrase '*f* rather than *c*' as '*f* and not-*c*', the 'and' does not work in the normal, truth-functional way. The normal 'and' does not require any relevance, but the '*f* rather than *c*' requires some relevance between *f* and *c*. Consequently, the weak reading uses 'and' in non-standard way. There is more than a simple conjunction. This makes Temple's suggestion just an alternative way of indicating the contrast. And this is not big news. As already noted above, the contrast can be expressed by alternative

linguistic means.

The fate of Temple's argument teaches us an already familiar lesson: the analysis of contrastive explanation should not focus too closely on linguistic issues. The central focus of interest should be on the cognitive setting, not the linguistic means to express it. Our interest should be in contrastive facts, not in contrastive statements. The danger is that the philosophical analysis regresses to the study of the pragmatics of some locutions in English. And this is not what *philosophical* analysis should do. After all, explanations are also given in other languages.

4. The counterfactual theory of explanation

Having discussed extensively the nature of contrastive *explanandum*, I can turn to adequate answers to these questions. In the following, I will provide my account of the adequate explanatory answer to a contrastive question. I will limit my discussion to the case of singular causal explanation, but the basic ideas can be easily transferred to other kinds of explanation.

Recently, counterfactual theories of causal explanation have become popular (cf. Horgan 1989, Schiffer 1991, Woodward 1993; Ruben 1994, Baker 1995 among others). This is not surprising. Although counterfactual theories of causation (for example, Lewis 1986) have their problems, most of them can be avoided if these theories are reinterpreted as theories of causal explanation, not as theories of causation. For example, it is not a problem if an account of causal explanation presupposes some causal notions or causal information, although this would be a problem if one is trying to give a reductive account of causation.

The counterfactual theory of causal explanation, in its simplest form, says something like:

a causally explains *f* if and only if

(a) *a* was a part of the causal history of *f*, and

(b) *f* would not have taken place if *a* had not taken place.

This formulation is not identical to any in the literature, but it gives the essential idea. It also shares with recent counterfactual accounts of causal explanation their central problem, which is not with the causal assumptions of the analysis, but in the sketchiness of these assumptions. These writers do not seem to be certain how to spell out their theory, apart from its counterfactual component. There are basically two problems. The first concerns the right level of description and the second the fixing of the background assumptions. I think the addition of the contrastive element to these theories can solve both of these problems.

Let us start with the second. The problem is that counterfactual

explanations always take for granted some causal background, but counterfactual theories of explanation usually lack the resources to characterize this background. The concept of *causal field* is often considered useful in this context (Anderson 1938; Mackie 1974). The idea is the following. A causal explanation is understood as an answer to a causal question. The question 'what caused this explosion?' can be expanded into 'what made the difference between those cases, within a certain range, in which no such explosion occurred, and this case in which an explosion did occur?'. Both causes and effects are seen as differences within a field, and anything that is part of the assumed causal field (the description of which is commonly left unstated) is ruled out as a candidate for the role of cause. (Mackie 1974: 35.) So causes are always causes within a certain field, and the same goes for explanations.

The problem with the counterfactual theories that disregard the idea that explanation is always an answer to a question is that they do not have any resources to fix the parts of the causal field that are taken for granted in the causal explanatory statement. As a consequence, they have to make unhelpful references to 'the pragmatic factors'. With the theories that combine the counterfactual theory with the contrastive idea, there is no such problem. (Hesslow 1983; Woodward 1993.) The contrast helps to fix the causal background in a very definite way: all the elements of the causal history that *f* and *c* share are assumed to belong to the causal field, and the concern is with their differences. One could say that what is shared between *f* and *c* belongs to the conditions and their differences to the causes.

A contrast is a very economical way of fixing the causal field since it does not require listing of all its components beforehand. We do not have to make explicit or even know all the details of the field at the beginning. If the question arises, we can check the answer with the cues provided by the contrast. This is especially clear when we are explaining the difference between two real outcomes. In such cases the causal field can be determined simply by comparing the processes.

The problem of the right level of description is the old problem of explanatory selection. In this case it can be called *the problem of overgeneral descriptions* (Schiffer 1991; Ruben 1994). Suppose that Hugo burps in Regina's presence, and this causes him to become embarrassed. Hugo's burping is explanatory of his becoming embarrassed, because if Hugo's burping in Regina's presence had not been a burping by Hugo in Regina's presence, it would not have caused this instance of his embarrassment. Now, consider the following property: a doing of something in the presence of someone. If Hugo's burping in the presence of Regina had not been a doing of something by Hugo in the presence of someone, it would not have caused his embarrassment. So, it seems that a simple counterfactual account of explanation makes this second prop-

erty explanatory when it intuitively is not. (Ruben 1994: 470.)

Ruben and Schiffer conclude that this makes the simple counterfactual condition an insufficient condition for a property's being explanatory. Schiffer adds three further pragmatic conditions having to do with manipulability, epistemic accessibility and being part of a reliable predictive practice (Schiffer 1991: 14-15). Unfortunately, Schiffer never offers a detailed account of how these pragmatic considerations work and how they solve the problem. According to Ruben these pragmatic elaborations are unnecessary, since the problem can be handled by the intuitive idea of the degree of determinateness of property (Ruben 1994: 470-471).

I will not go to the details of Ruben's solution, which is quite complicated (Ruben 1994: 475). The point I want to make is that we can solve this problem in a much simpler way. We do not have to have recourse to the hierarchies of properties with varying level of determinateness. All we need is to turn our attention to the contrast in the *explanandum*. This route is not taken by Ruben, since he thinks that all *explananda* are not contrastive and that contrasts belong only to the pragmatics of explanation (Ruben 1990: 40-44).

A fact is explanatory if it causes *f* to occur instead of *c*, otherwise it is not explanatory. The same goes for alleged overgeneral descriptions: they either satisfy this test or not. There is no real problem here. The contrast to Hugo's being embarrassed seems to be Hugo's remaining his normal, but nervous, self. Hugo's burping in Regina's presence causes his embarrassment, but his breathing, which is also something Hugo does in Regina's presence does not. Since the mere doing of something in the presence of Regina does not make the difference between the fact and the foil, it is not the explanatory factor we are looking for. So instead of adding more requirements to our analysis of *explanans*, we can solve the problem easily by taking a more careful look at the *explanandum*. This helps us to preserve the intuitive simplicity of counterfactual account of explanation. Why to make an analysis complicated when it is not necessary? Maybe Ruben should reconsider his views about the philosophical relevance of the contrastive idea.

Negative causes

Counterfactual analysis of causal explanation has an advantage of making sense of negative causes that are often taken to be troubling entities in the metaphysics of causation. Negative causes, like gaps, omissions, preventions, absences and privations, are all used quite naturally in our causal cognition and talk. The problem they create for causation theory is that if their existence is taken at face value they introduce dubious ontological categories into one's ontology. Negative properties, non-existences and non-actions are not the kind of things that a philosopher

likes to populate his world with. They just do not feel real and they also create problems of overpopulation for those who take parsimony to be a virtue in ontology.

A counterfactual account of causal explanations makes it possible for us to admit their legitimacy in causal cognition, without accepting them as fully legitimate ontological categories. Of course, nothing in the account says that they cannot be accepted, if one so wishes (Schaffer 2000). It is always possible to interpret causal claims that make reference to these negative entities as claims that are intended to be explanatory. There is no problem with accepting counterfactual claims that refer to these negative entities. After all, these claims are contrary to fact. Secondly, there is no problem with their negativity, since they are not intended to be part of the causal processes themselves, but *instead arise from the comparison* between processes. They just report the ways in which two processes are not alike. The world remains free of negative entities, but our explanatory discourse can have them in a quite innocent way.

The presuppositions of a sensible contrast

The contrast fixes everything that is common between the fact and the foil to the causal background field, but how much in common should these two have? It seems that there are some limits. Not all contrasts are sensible, after all. There should be some kind of relation of relevance. As Elliot Sober (1994: 178) notes, the following does not seem to be a sensible contrastive question:

> Why is Kodaly a Hungarian rather than a vegetarian?

Kodaly's being a Hungarian and his being a vegetarian do not seem to be incompatible alternatives to each other. But there are also contrasts that are incompatible alternatives to each other, but that do not make sense. Consider the following:

> Gregor Mendel participated in a pea-growing contest in 1890
>
> [Gregor Mendel died in 1884,
>
> Gregor Mendel was never born]

The trouble with this contrast class seems to be the fact that the propositions deny each other's presuppositions. If Gregor Mendel was never born, it does not make sense to ask about the date of his death or his hobbies. Similarly it does not make sense to ask about activities of a person after his death. It seems that we can conclude that the alternatives must have something in common. (Barnes 1994: 39.) But what should this something be?

Alan Garfinkel has made one suggestion. He says that the *f* and the members of *Cc* should have a common presupposition. He clarifies this

idea by saying that there should be a P such that *f* and every member of *Cc* entails P (Garfinkel 1981: 40). Now, as Temple has rightly observed, Garfinkel's formulation is too loose for the following reason. For any choice of *f* and its contrasts it is possible to find such a P that they entail it. For example, the disjunction of *f* and the members of *Cc* is such a common presupposition. The examples Garfinkel uses show that he probably had something more substantive in mind, but it is difficult to say what. (Temple 1988: 145.)

The second suggestion is by Elliot Sober. He has suggested that this something in common is a common cause. According to him, we should make the following requirements (Sober 1994: 178):

> 1) *f* and not-*c* *trace back* to a common cause, and
>
> 2) this common cause *discriminates* between *f* and *c* (i.e., makes *f* more probable than *c*).

Sober's idea seems intuitive, but, Eric Barnes has argued that there are contrasts that satisfy 1) and 2), but do not contrast. Consider the following pair:

> Halonen won the Finnish 2000 presidential election on 6 February 2000
>
> [Halonen was depressed shortly after 6 February 2000]

This pair satisfies Sober's requirements, but it does not contrast appropriately (Barnes 1994: 43). According to Barnes, the contrast does not work because the election process and the psychodynamics of an electoral candidate are different kinds of causal processes. Two processes are not similar enough. If this argument works, we need something more than just a common cause.

Following a similar intuition, Peter Lipton has suggested the following requirement:

> 3) *f* and not-*c* should have largely similar causal histories.

This presupposition is closely related to *the difference condition*, which Lipton takes to be the central criterion of explanatory relevance. The principle goes as follows: to explain why *f* rather than *c*, we must cite a causal difference between *f* and not-*c*, consisting of the cause of *f* and the absence of a corresponding event in the history of not-*c* (Lipton 1990: 256).

Here the corresponding event is to be understood as something that would bear the same relation to *c* as the cause of *f* bears to *f*. In a later paper Lipton replaced the term 'corresponding event' with a more general notion of the 'corresponding token' (Lipton 1993: 43). Although the term 'token' is more general, there is a reason for the change. It is not clear that the explaining factor should belong to the ontological category of events; for example it might be a mechanism. Lipton's idea seems to be that similarities should be so extensive that there should

be only one or very few differences between the causal histories. This would allow us to pick the differences as explanatory factors.

As Barnes notes, Lipton's suggestion sounds plausible, but the similarity criterion seems vague. How much similarity is enough? And, does it matter in which respects they are similar? Without answers to these questions, the suggestion seems to be open to counter examples. Consider again the following contrast:

> Halonen won the Finnish 2000 presidential election on 6 February 2000
>
> [Halonen was depressed shortly after 6 February 2000]

Both have an extensive shared history, but the fact and the suggested foil do not seem to contrast appropriately (Barnes 1994: 45-46). According to Barnes, we need something more. His own suggestion is the following:

> 4) *f* and not-*c* should be culminating outcomes of *the same types of natural causal processes*. (Barnes 1994: 47)

What is wrong with the example is that the electoral process is the token of a different type of natural causal process than the process leading to psychological depression.

Now, I think that one can with justification ask whether Barnes' idea is much of an improvement over Lipton's. The concept of the 'natural kind' is in itself in need of clarification. The concept has been used in a variety of ways (see Hacking 1991), and its extension to the 'process kind' only adds to the confusion concerning its meaning. In which respects should two processes be similar to be members of the same natural kind? Barnes' own examples of process kinds are not very helpful. The examples he considers in his paper are the following: 'electoral process', 'deliberation process' 'psychological reaction to the perception of an apple' and 'the story of apple A'. It is very difficult to see what makes all of them examples of natural causal process kinds. What would not be a natural causal process kind? I don't want to say that there is something wrong with this proposal. To the contrary, as far I can see, it is true. What I want to point out is that it is not much of an improvement over Lipton's proposal.

The importance of how-questions

The above discussion seems to show that we cannot significantly improve Lipton's articulation of the presuppositions of the contrastive *explanandum*. Does it also follow that his difference condition is also beyond improvement? I don't think so. Lipton notes that the difference condition is quite analogical to Mill's method of difference. (Mill 1906: 253-259.) This has its advantages, but it also brings to light one serious problem: multiple differences. Just as Mill's methods cannot always pick

the right cause(s) if there are multiple differences, the difference condition seems to be in trouble in similar circumstances. Which of the differences are explanatory? Lipton himself suggests that we can reduce the list of differences by additional principles of causal selection. He mentions the requirements that a good explanation should tell us something new, that we prefer explanations in terms of sufficient causes and that we also like if the cause is necessary for the fact in the circumstances (Lipton 1990: 259-260). I don't have anything against these requirements, but I think Lipton fails to mention something essential from the point of view of causal explanation.

We make progress when we ask whether one single event or fact can really explain *f* [*c*]? Lipton and others seem to assume that the answer to this question is affirmative. I do not think that the answer is that obvious. I think that when explaining something we always also presume that we can also answer, or at least sketch an answer, to the question:

[M] Why *a* makes the causal process lead to *f* rather than *c*?

If we can't answer this question, we have the feeling that we have not fully explained *f* [*c*]. And sometimes we ask 'why *f* [*c*]?', even when we know *a*. In such cases we are looking for the answer to a how-question. This leads to the suggestion there has to be a mechanism which ensures that *f* occurs instead of *c*. If we cannot account for such a mechanism (or a story), we have not explained *f* [*c*].

Of course, this leads to following question: If this is such an important consideration, why do Lipton and others fail to see it? My answer is as follows. The usual examples of causal explanation are selected because of their relative simplicity and familiarity. The examples of explanation are familiar because they are very similar to the explanations we use in our everyday life or to the ones that are known from scientific folklore. We already know how the explanation works and by which mechanisms. This allows the examples to be relatively simple, as well. Since everybody already knows the mechanism, there is no need to spell it out in the paper. Using familiar and simple examples is an important pragmatic consideration when writing a paper; however as an unintended consequence, the menu of explanations also becomes systematically biased: It tends to include only examples in which we already know the answer to the question [M].

If we take the considerations related to the underlying how-questions seriously, we should formulate our criteria of explanatoriness in the following way:

a explains *f* [*c*] if and only if

(1) *a* belongs to a causal process leading to *f*, and

(2) there is a causal mechanism *m* that ensures that *f* occurs instead of *c* because of *a*.

I call (1) *the etiological requirement* and (2) *the mechanism requirement*. The etiological requirement is there to secure that we capture the facts

that have a true causal connection to *f*, and the mechanism requirement ensures that the explanation picks up the right factors from the causal history.[10]

The idea of a mechanism

As explained above, *m* is quite often left unstated because the mechanism is presumed to be already known by the audience of the explanation.[11] I do not require that the knowledge of *m* be very detailed. In practice *m* is often taken as a black box. We often have only a sketchy account of it. What is important is that the existence of *m* is presumed and that *m* is thought to be *in principle* specifiable. As a consequence, the detailed specification of the mechanism can be seen as auxiliary information. However, the nature of *m* cannot be taken as a complete black box. We have to know something substantial about *m*; otherwise supplying it would be very easy. In the vein of Molière's doctor, we could construe explanations just by describing things as having the power to bring about the effect we are interested in. But, although the reference to the *virtus dormitiva* does not give us a false explanation, it nevertheless also fails in giving us an informative explanation.

It might be claimed that my mechanism requirement suffers from similar problems of vagueness as Barnes' concept of the similar type of natural causal process. This claim is partly true. It is very difficult to give a characterization of a causal mechanism that is both general and informative. My model of explanation is intended to be applicable to all sciences, but the causal mechanisms in these sciences are so different that capturing their common essence seems futile. The only unity one can give to the idea of a causal mechanism is that it answers the question how *a* makes *f* happen instead of *c*?[12]

This does not mean that we should leave the issue here; after all, definitions are not the only way to characterize a concept. A more realistic way to approach a concept is to use paradigmatic examples. A float valve and a voltage switch (Glennan 1996) are examples of very simple, and literally mechanical, mechanisms. But the concept of a causal mechanism is not limited to the world of mechanical apparatus. DNA replication, protein synthesis and chemical neurotransmission (Machamer, Darden and Craver 2000) are examples of more complicated biochemical mechanisms. In Chapter 4 I will discuss evolutionary biology where the central, but not the only, mechanism is natural selection. In the human sciences folk psychology covers a wide array of explanatory mechanisms, starting from the process of practical deliberation. The mechanisms in the social sciences are not limited to those of individual psychology: a competitive market and a social filtering process both include structural components. Some examples of these mechanisms will be discussed in the second part of this study. I think

that with the help of these examples it is possible to capture the basic idea behind the requirement of mechanisms.[13]

None of these examples of mechanisms are from fundamental physics. There is a good reason for this. Once we reach the fundamental level, there are no underlying mechanisms to be found. This is true by definition. If gravity is one of the fundamental forces in the nature, we are not in a position to explain it. As a consequence, we cannot explain facts at the fundamental level. (Glennan 1996.) We can use them to explain facts at other levels, with the help of mechanisms, but they will always remain outside our explanatory reach. We just have to accept that these things happen as they do. This view has an interesting consequence. It might be that the original intuitions behind the covering-law theory of explanation fit only explanations at the level of these fundamental nomic attributions. At that level all that we can have for an explanation is the observation that 'it always happens this way'. This contrasts nicely with the special sciences, where we can always do better than just noticing law-like regularity: we can look for the mechanisms that make the things happen the way they do.

There is one more justification for the requirement of mechanism. Simple counterfactual theories of causation and explanation are in trouble when they try to explain why reliable symptomatic or diagnostic features are not explanatory, although they do fulfill the requirements made by these theories. The red shift in spectra of distant stars does not cause the expansion of the universe, although we can say that if there were no red shift the universe would not expand. (Miller 1987: 72-73.) The idea that we always presume some mechanism to support our explanatory counterfactuals makes this judgment natural. Once we start to consider how the red shift would make the universe expand, our knowledge of physics would lead us to the conclusion that there is no way in which the red shift could cause the universe to expand, but there are good reasons in favor of the opposite conclusion.

The formal properties of the explanatory relation

What are the formal properties of the relationship of explanation in my model? Explanatory relevance is based on causal relevance. Causation is generally accepted to be both irreflexive and asymmetric. My account of causal explanation inherits these formal properties. Consequently, nothing can explain itself, and the *explanandum* fact cannot explain the *explanans* fact.

Let us next consider transitivity. As causal chains are transitive, at least sometimes causation is transitive at token level. When we move to the type level, things are not that obvious. For example, employing birth control may be a type-level cause of sexual activity that is, in turn, a type-level cause of pregnancy. However, it is not sensible to claim that

employing birth control is a type level cause of pregnancy. So, at least in the case of probabilistic type level causal claims transitivity does not hold. What about causal explanation, is it transitive? The first problem in discussing this issue is that in my theory *explanans* and *explanandum* are different kinds of things. The *explanans* is a fact or a conjunction of facts, whereas the *explanandum* consists of a fact and its contrast. *a* explains *f* [*c*], but *f* [*c*] is not a kind of thing that could explain anything.

But let us consider the following situation: *a* explains *f* [*c*], and *f* explains *g* [*d*]. Does *a* explain *g* [*d*]? According to the etiological requirement, *a* has to belong to a causal process leading to *f*, and *f* has to belong to a causal process leading to *g*. It follows that *a* belongs to a causal process leading to *g*. Transitivity holds here, but what about the mechanism requirement? Let us suppose that a mechanism *m* ensures that *f* occurs instead of *c*, because of *a*, and that a mechanism *n* ensures that *g* occurs instead of *d*, because of *f*. Transitivity does not seem to hold here. It is not necessary that *a* and the mechanism *m* ensure that *g* occurs instead of *d*. The same applies to *a* and the combination of mechanisms *m* and *n*. Furthermore, the causal fields presupposed by explanatory questions are not similar: *f* [*c*] presupposes a different causal field from the one presupposed by *g* [*d*]. This context sensitivity of explanatory claims prevents their transitivity. However, it does not make it impossible that there are cases where *a* could explain *g* [*d*], and this is why the relation of causal explanation is not intransitive either. The fact that my theory shows that explanation is an irreflexive, asymmetric and non-transitive relation gives some support to my theory, since intuitively we think that explanation has these properties (cf. Ruben 1985: 149).

Causal pre-emption and overdetermination

There is one interesting consequence of my account of an adequate answer to a contrastive question, a kind of bonus. It justifies a natural way of treating the cases of causal pre-emption and overdetermination in explanatory context. To make these ideas clear, let us consider the famous case of an unfortunate desert journey (Hart & Honoré 1959: 219-220; Mackie 1974: 44-45).[14]

George sets out on a trip across the desert. He has two personal enemies. The first enemy puts a deadly dose of poison in his drinking water can. The second enemy, without knowing about the first plot, makes a hole in the bottom of the can. George starts his desert journey and all the poisoned water leaks out before he needs any water, and as a consequence he dies of thirst. The death by thirst pre-empts the death by the poisoning.

Now, we have different explanations of the death depending on how the *explanandum* is specified. If we want to explain why George died of thirst, the explanation is the hole in the can, and the poison has no ex-

planatory relevance. But if we want to explain why George died during his desert journey, we have to cite both the hole and the poison. This is a very natural judgment, but it creates a problem. The problem is that poison and the hole in the can are not causes in the same sense. Although the poison belongs to the causal history of George's death, the death is not by poisoning.

We think that a causal explanation should reflect the causal facts; however, as there is no causal fact of death by poisoning, why should the poison be mentioned? The requirement of mechanism spells out this justification. We want an explanation that ensures that *f* occurs instead of *c*. Pre-emptive (and overdetermining) causes can be treated as fail-safe causal set-ups. To be sure, *f* is actually produced by a_1 and mechanism m_1, but had they failed, a_2 and m_2 would have produced *f*. Since the absence of a_1 & m_1 would not have made the *c* happen instead of *f*, we also have to include a_2 & m_2 in our explanation. And because of the mechanism requirement, we should specify how our explanatory factors bring about the fact instead of its foil. This means that we should specify whether this is a case of joint causation, pre-emptive causation or causal overdetermination, since these are different ways in which the explanatory factors bring about *f* rather than *c*. The requirement that an adequate explanation specifies the mechanism by which the *explanandum* fact was produced makes our solution to the cases of causal pre-emption and overdetermination clearly something other than *ad hoc*. After all, this was the bonus, not the reason why the requirement was there in the first place.

Notes to Chapter 2

1 See Mackie (1974: Chapter 10) and Ruben *(*1990: Chapter 5) for extensive arguments for facts as *relata* of explanation.

2 Mellor (1995) argues that also the relata of causation are facts. The acceptance of Mellor's thesis would not change the essentials of my account of causal explanation. However, if one wishes, one can challenge Mellor's claim by distinguishing causal explanation and causation, and by arguing that Mellor confuses these two matters. My account of causal explanation could be useful in the development of such a challenge, but I will not attempt that here.

3 For example, van Fraassen 1980; Tuomela 1980; Sintonen 1984; Koura 1988; Cross 1991; Hintikka & Halonen 1995.

4 It also provides a nice way to tie explanation in a natural way to the rest of the process of inquiry, but I will not develop this theme in this work (cf. Jardine 1991; Sintonen 1993).

5 For example, Hart & Honoré 1959; Hansson 1975; van Fraassen 1980; Garfinkel 1981; Hesslow 1983; Woodward 1993 (originally published in 1984); Lewis 1986; Sober 1994 (originally published in 1986); Temple 1988; Kitcher 1989; Lipton 1990, 1991, 1993; Cross 1991; Grimes 1993; Henderson 1993; Barnes 1994; Hitchcock 1996, 1999; Carroll 1997, 1999; Glymour 1998;

Percival 2000; Risjord 2000.

6 For a review of recent empirical studies of human causal and explanatory reasoning that show support for the contrastive idea, see Hilton 1995.

7 See Glymour (1998) for an account of statistical explanations that use the contrastive idea.

8 Many writers seem to have accepted this thesis without reservations. See for example Barnes 1994, Carroll 1997, 1999, Hitchcock 1999, Percival 2000, and Risjord 2000.

9 In some explanations of differences the situation is symmetrical. We might want an explanation either for *f* [*c*] or for *f** [*c**]. But notice that here we have two separate explanations, not one explanation with compatible contrasts.

10 Note that this analysis involves elements that are not included in purely conditional analyses of causal explanation. For example, an analysis in terms of INUS-conditions (Mackie 1974: 62) requires neither the presence of a continuous causal process nor the elaboration of a mechanism.

11 For a review of empirical work related to the role of mechanisms in causal reasoning, see Ahn & Kalish 2000.

12 Although there are many valuable ideas that can be found in Jon Elster's discussions of mechanisms (Elster 1983, 1989, 1999), I do not want to be associated with his latest definition: "… mechanisms are *frequently occurring and easily recognizable causal patterns that are triggered under generally unknown conditions or with indeterminate consequences*" (Elster 1999: 1). Almost every part of this characterization is suspect. Why should a mechanism be by definition frequently occurring? There are no conceptual problems with rarely occurring mechanisms. Secondly, as finding underlying mechanisms is sometimes the hardest part of scientific work, it does not make sense to require that mechanisms are 'easily recognizable'. And thirdly, why should we suppose that mechanisms are always plagued by some sort of indeterminacy? It might be that Elster's characterization applies to some subset of folk-psychological explanatory patterns that are the topic of his discussion, but there is no justification for the generalization of these characteristics to all mechanisms.

13 For a discussion of mechanisms in the social sciences see Hedström and Swedberg 1998.

14 For a further discussion about causal pre-emption and overdetermination, see Mackie 1974, Bunzl 1979, Lewis 1986, 2000, McDermott 1995, Hausman 1998.

Chapter 3
Indeterministic causation and the role of laws

Developments in physics during the last century have forced us to give up the idea that the world is fully deterministic. This has led philosophers to consider the concept of indeterministic causation. It is no longer plausible to think that all probabilistic claims in science are only reflections of our ignorance. There are at least some contexts in which the probabilities can be plausibly taken as reflecting the truly indeterministic nature of the processes under consideration. In this chapter, I will take this situation for granted and proceed to consider its implications for theory of explanation. The central question is the following: are there good reasons to modify my account of causal explanation in response to the challenge of indeterministic causation?

I will first discuss Wesley Salmon's argument to the effect that there are indeed such reasons. I argue that Salmon's arguments against *the Leibniz principle*, that underlies the contrastive approach are not convincing. First, it is not obvious that Salmon can dispense with the principle. To the contrary, it seems to be a natural companion to Salmon's ontic approach to explanation. Then I proceed to discuss the problem explaining outcomes of singular indeterministic processes. I suggest that we should bite the bullet and accept the fact that we cannot in such cases explain why *f* occurred instead of *c*. However, I point out that this does not lead to the conclusion that such indeterministic process are completely outside the scope of explanation. For example, we can try to explain probabilities attached to the outcomes of an indeterministic process. I argue that the contrastive approach succeeds better than Salmon's approach in explicating what we can explain about indeterministic processes.

In the second section I will consider probabilistic versions of the contrastive theory of explanation. I argue that these theories confuse statistical relevance or expectability with explanation. I also argue that probabilistic accounts fail fatally in the case of singular causal claims,

and that they miss the distinction between a partial and a full answer to an explanation-seeking question.

Because the deductive-nomological theory of explanation (D-N model) has been extremely influential in philosophy of science and has without doubt helped philosophers form their intuitions about explanation, it is useful to discuss the role of laws and deduction in explanation. I will consider these issues in the second half of this chapter. In accordance with Chapter 2, my discussion concentrates solely on the case of singular causal explanation.

In the third section I will discuss Salmon's critique of the third dogma of empiricism. According to it, all scientific explanations are arguments. I argue that Salmon's thesis is ambiguous, and that his arguments suffice to give up stronger versions of the dogma, but not weaker ones. I suggest that the deductivist position which claims that explanations can be reconstructed as deductive arguments, but which denies that the deductive relationship has a constitutive role in explanation, is acceptable. I also argue that this position should be distinguished from the covering-law theory, since such a deductivism is not committed to idea that an explanation always needs a covering law among its premises.

Finally, the last section will discuss the role of covering laws in explanation. After going through several arguments to the effect that a full explanation includes a covering law, I conclude that no convincing arguments can be found to support the covering-law requirement. To the contrary, I argue, there are several reasons why we should think that singular causal explanations do not include an essential reference to a covering law. However, while I deny that laws have a constitutive role in singular causal explanation, I argue that laws, truisms and other generalizations can have an important heuristic role in the construction of explanations.

1. Explaining outcomes of indeterministic processes

Wesley Salmon has claimed that in some indeterministic contexts the *explanandum* is not contrastive. He argues on this basis that the contrastive theory does not apply to an important class of *explananda*. (Salmon 1998: 328-329.) He notes that contrastive explanation rests on the following principle:

> If *a* explains *f* in one case, and *c* and *f* are incompatible, then *a* cannot explain *c* in another (similar) case.

Salmon calls this *the Leibniz principle* (after Stegmüller). David Ruben calls this same principle *Plato's second principle* (Ruben 1990: 64-66). Since the above formulation is in non-contrastive form, let us reformulate it as follows:

> If *a* explains *f* [*c*] in one case, then *a* cannot explain *c* [*f*] in another (similar) case.

According to Salmon, this principle does not hold in the indeterministic cases that are familiar from modern physics (Salmon 1998: 154-158, 326-329). To understand Salmon's reasoning, let us consider an example. The example is not from a textbook of modern physics, but it allows us to see the problem clearly.

Example 1: A chance set-up

> We have a coin-tossing device, with two coins and a switch. When the switch is down, a biased coin is tossed, so that that the chance of heads will be .35 and the chance of tails will be .65. When the switch is up, an another biased coin is tossed, so that the chance of heads will be .65, and the chance of tails will be .35. We will suppose for the sake of simplicity that tossing a coin with this set-up always includes irreducibly random elements. We want two cases to be explained. In the case A, the switch is turned down and the result is heads. In the case B the switch is turned up and the result of a toss is also heads. Let the *explanandum* be *f* [*c*] in the case A, and *f** [*c**] in the case B.

According to Salmon, we cannot explain why the result of the throw in cases A and B was heads rather than tails. However, he claims that we can explain why the result was heads. So, we cannot explain why *f* [*c*] or *f** [*c**], but we can explain the plain *explananda f* and *f**. (For a similar account see Lewis 1986: 230-231.) This observation convinces Salmon that 1) the Leibniz principle is wrong in indeterministic contexts, 2) that not all *explananda* are contrastive, and 3) that we can also explain events with low probability.

My concern here is only with theses 1) and 2). According to Salmon, subscribing to the Leibniz principle and to the idea that all explanation is contrastive springs from an attachment to a deterministic worldview. If we give up this worldview, as modern physics suggests we should do, we should also give up the Leibniz principle and the contrastive account of explanation which rests on it. (Salmon 1998: 327-328.) I do not share Salmon's assessment of the situation. Remember that we are now speaking about explanation, not causation. It is quite uncontroversial that modern physics describes the world as including irreducibly indeterministic processes. However, I claim that the implications of this fact for the theory of explanation are not that obvious. Salmon needs to show that when we give up the deterministic picture of the world, we also need to give up these two fundamental ideas about explanation.

Salmon argues for his position by trying to show that we can do without the Leibniz principle. He claims that the main reason for keep-

ing the Leibniz principle sacrosanct is that it rules out some pseudo-explanations. He gives three examples of pseudo-explanations: 1) the explanation of the power of opium to produce sleep by its dormitive virtue; 2) the explanation of all happenings by the will of God; and 3) the explanation of behavior by a psychoanalytic theory that is compatible with all possible behavior. Salmon argues that we don't need the Leibniz principle to show that these explanations are not scientific explanations. According to him, the dormitive virtue theory is too *ad hoc* to be acceptable, the will of God is a scientifically unacceptable explanation because it refers to a supernatural agency, and that psychoanalytic theories which are compatible with all possible behavior cannot be scientifically confirmed. According to him, it suffices that we require that scientific explanations appeal to *bona fide* scientific laws and theories. (Salmon 1998: 327.)

Salmon's argument is interesting. He claims that the Leibniz principle is only used to rule out the kinds of explanations he mentions, and that he can show them illegitimate without it. This suggests that the Leibniz principle is only an *ad hoc* device to rule out pseudo-explanation, not a deep principle of explanation at all. Let us see how his argument fares, before considering whether the Leibniz principle has also some other uses.

As I see the situation, we don't need the Leibniz principle to rule out the dormitive virtue theory. However, Salmon's diagnosis of the situation is not accurate either. The explanation is not a pseudo-explanation because it is *ad hoc*. It is a pseudo-explanation because it is non-explanatory, and it is non-explanatory because it is uninformative. In contrast, some *ad hoc* explanations are both informative and explanatory. The trouble with the dormitive virtue explanation is that it simply describes the cause as the kind of cause that causes the effect we are interested in. This explanation does not require any empirical research, and it cannot fail. The *explanans* is basically deduced from the *explanandum*, and we do not want explanation to be that easy.[1]

What about God's will as an explanatory factor? Here I think Salmon's diagnosis is *ad hoc*. It is not enough to claim that the reference to a supernatural agency is not acceptable in science: there should also be some kind of a justification for this claim. We would like to have some principled reason why the God's will is not an acceptable explanatory resource in science. Salmon owes us a true diagnosis of this case that does not presuppose the Leibniz principle. Before that, I think we should stick with the standard justification, which is the following: 'It's God's will' is too easy an explanation. It can explain everything equally well, and it is compatible with everything. Consequently it does not tell us why *f* happened instead of *c*. Because of this, its explanatory power is only apparent. The competing (scientific) theories do not claim to explain everything, but they at least sometimes provide us reasons why

f had to occur instead of *c*. We are here basically resting on the Leibniz principle. So, when we appeal to the principle Salmon wants to dispense with, we can make sense of the situation.

I think that a similar solution is applicable to psychoanalytic theories that are compatible with all possible behaviors. (I don't know if there are such psychoanalytic theories, but let us suppose that there are.) If a theory really cannot explain any *f* [*c*], it cannot explain anything. The point of scientific theories is to track dependencies in the world, and if a theory does not find any, its explanatory power is zero. We do not need considerations concerning the confirmability of the theory to establish this, as Salmon thinks. And if we do so, we again confront the charge that the reply is *ad hoc* in nature. We want an account which shows why the theories that are too easy to confirm are not good explanations. And for this purpose we need more fundamental principles, like the Leibniz principle.

Of course, Salmon might just have been too hasty in his analysis of these examples. However, there is a more general reason why his approach to these pseudo-explanations does not work. He thinks that the criterion for ruling out these pseudo-explanations is their unscientific nature. But if the argument rests on this criterion, one needs a criterion of demarcation between science and non-science. Salmon has not provided us any indication as to how this perennial problem can be solved without the Leibniz principle.

Furthermore, it is unclear why we should raise the issue of demarcation at all. An appeal to *virtus dormitiva* is a bad explanation outside science as well. We should be analyzing what makes explanations good in general. There are good explanations that are not scientific, and there are bad scientific explanations. It would be inconceivable, at least to me, if scientific and everyday explanations did not share some basic ideas about good and bad explanations. After all, science has its origin in common sense cognition. There should be some kind of continuity. Of course, it now includes much that is apparently incompatible with common sense, but I seriously doubt that this incompatibility also extends to the general principles of explanatoriness.

It seems that Salmon is not very successful in dispensing with the Leibniz principle. It is needed to explain why some pseudo-explanations are not explanatory. I think the real motivation for accepting the principle is that it reflects our conviction that an explanation should trace objective relations of dependency as explained in the previous chapter. And since Salmon's ontic conception of explanation is based on this same conviction, it seems clear that he should also stick to something like the Leibniz principle.

Does Salmon have any other arguments for dispensing with contrastive explanations? Surprisingly he does not. His discussion is directed against the high probability theories of explanation. These theories ac-

cept that we can explain f^* [c^*] in the case B, but not f [c] in the case A. By referring to the arbitrariness of any probability threshold for explanation and by bringing in some symmetry considerations, he argues that the supporter of the high probability requirement should also accept explanations for the improbable outcomes. I take these arguments to be convincing. If we accept that we can explain f^*, we should also accept that we can explain f. And if we accept that we can explain f^*[c^*], we should also accept that we can explain f [c]. However, we need some reasons to accept the antecedent of the counterfactual.

Salmon's justification seems to be that if we don't accept the antecedent, we cannot explain indeterministic events at all, and this would be too heavy a price to pay. Fortunately, we do not have to pay this price. The central intuition behind the contrastive approach to explanation is that the proper understanding of the explanation requires thinking carefully about the *explanandum*. Salmon is right that we can explain *something* about cases A and B. But the question is: what about them we can explain? And there are a lot of things to explain. For example, in case A, we can try explain:

1) why the probability of heads was .35 rather than some other value; or

2) why the result of the experiment was that the coin landed heads or tails, rather than something else, for example, that it melted, was vaporized, etc.).

Question 1) directs us to the theoretically interesting case of explaining the propensity of the coin to land heads with some specific probability. Once we have an answer to this question, the lack of an explanation for the outcome of a single experiment is much less important. Question 2) shows that there are many other contrasts to be explained, some of which are interesting. The only thing we cannot explain is f [c], since it does not satisfy our requirements.

Is there anything else to explain? I don't think so. Contrary to Salmon, the contrastive approach is not a relic from the deterministic world that is only a hindrance to the proper understanding of explanation in the indeterministic contexts. Rather, it seems to be an indispensable tool for clarifying what we are capable of explaining in such contexts. Salmon agrees that we cannot completely explain f, but his theory does not allow him to explicate *what* about f can be explained. The contrastive theory can do this, which gives it a clear advantage.

And note how little we have to give up to be able to use this tool. We only have to give up the possibility of explaining singular outcomes of experiments (and actual frequencies that are the results of finite series of experiments), when both the fact and the foil have the possibility of occurring. And usually these are not the scientifically most interesting cases. For example, physicists working with quantum theory are not

interested in explaining the results of single experiments or actual frequencies of outcomes in particular series of experiments. These actual frequencies are only evidence for the real objects of explanation, which are the probability distributions for the possible outcomes. (Watkins 1984: 228-229; Kitcher 1989: 450-452; Woodward 1989: 367-368.) Scientists are interested in the general phenomenon and its probability distributions, not in the idiosyncrasies of some specific set of data. This emphasizes the fact that we can have a deep theoretical understanding of the phenomenon without being able to predict or explain singular outcomes. The belief that the explanation of singular probabilistic outcomes is somehow scientifically central is an artefact of the philosophical discourse on explanation. The problem of the single case is philosophically challenging, but scientifically it is not a big deal.

In a later article, Salmon seems ready to concede this point. He accepts that he cannot convince his opponents about the importance of singular outcomes to theoretical basic science (Salmon 1998: 158). Without giving up his position, he tries to show that there are some cases of single-case explanations in the applied sciences that the above strategy cannot handle. He has basically two examples of such situations.

The first example concerns the outbreak of Legionnaires' disease in 1976. All the (original) victims of the disease had stayed in the same hotel, but not all hotel visitors contracted the disease. Actually, only a small percentage did. Later the bacillus responsible for the disease was found in the cooling towers for the air-conditioning system. Salmon suggests that since quantum fluctuations may lead to large uncertainties in the future trajectories of molecules in the air, we might not be able to provide a strictly deterministic explanation for why particular persons contracted the disease and some other people did not contract it in similarly infested rooms. Consequently, we cannot have a deductive explanation why precisely these people contracted the disease. However, he argues that for the purposes of assigning responsibility and taking preventive steps in the future, we have an adequate explanation of the disease in this limited sample of people in 1976. (Salmon 1998: 158-159.)

The second example concerns eight soldiers (out of a group of 2235) who witnessed the detonation of an atomic bomb at close range during Operation Smoky in 1957, and who subsequently developed leukemia. According to Salmon, the high levels of radiation they were exposed to explain this incidence. Leukemia occurs with a non-zero frequency in the population at large, and it is possible, but not very likely, that these incidences of leukemia were due to a chance fluctuation rather than exposure to high levels of radiation. This makes the explanation probabilistic. Salmon argues that from a practical standpoint the fact to be explained is the high incidence of leukemia in this particular sample, not its incidence among all people who ever have been or will be ex-

posed to that amount of radiation. Consequently, he argues, this is not a deductive statistical explanation. He also argues that this focus on this particular sample was of importance in the deciding whether the soldiers should receive extra compensation from the federal government. (Salmon 1998: 134, 159.)

Salmon's examples, and his arguments, are not convincing. (See Hållsten 1999 for a similar assessment.) In the case of Legionnaires' disease, we do not need to explain why these particular hotel guests contracted the disease in order to assign responsibility or to take preventive steps. For the latter, it is sufficient to know that there is a risk of Legionnaires' disease with certain kinds of air conditioning machines. It does not matter which individuals contracted the disease or exactly how many instances there were. The central issue is to eliminate the possibility of an outbreak of the disease. Similarly, when attributing responsibility, the significant issue is whether someone is responsible for putting the hotel visitors at risk, not the exact identity or the number of people contracting the disease. We don't have to explain why one particular guest contracted the disease rather than another. All we need to know is that it came from the same source. Consequently, we do not have to explain single cases, we only need to recognize that there is a number of incidences. Also, we can know the cause of their disease without being able to explain why they in particular contracted it.

Similar points apply to Salmon's leukemia example. The basis of the government's responsibility is in having put its the soldiers at a high risk, not causing particular soldiers to have leukemia (Kitcher 1989: 459). We do not have to explain the individual cases, we need note only their higher risk of getting this particular form of leukemia. And if a court decides that the risk was high enough, it can decide that the government is responsible, and that it must compensate all soldiers who were put at risk and who developed later leukemia. In this case we cannot know the cause of leukemia in individual cases (as we knew with Legionnaires' disease), but we know what increased their risk of having leukemia. This is enough for attributing legal and moral responsibility.

These considerations show that the explanations of single outcomes of chance set-ups are not as important as a practical matter as Salmon thinks they are. The sky does not fall if we recognize the fact that we cannot explain them. It might be that my conservative position on explanation makes me a "deductive chauvinist pig" (Salmon 1998: 161), but if the alternative is to be a philosopher who gives up well-supported principles without good argument, then let it be so.

Let us review the position that I have adopted. The central idea is that a contrastive approach can be used to explicate what is explained and what is not. An approach that gives up the Leibniz principle cannot have this advantage. The only drawback of this approach is that we cannot explain some singular outcomes of indeterministic processes.

The Leibniz principle rules out cases where both *f* and *c* have a chance of occurrence after *a* occurs (within the intended causal field). However, we can explain *f* [*c*] in situations where *a* and the presumed causal field do not necessitate *f*, but make it impossible for *c* to occur. This was the case in our earlier example of paresis. So we need only rule out the possibility of the occurrence of *c*, while still retaining the possibility that *f* will occur. We cannot predict the occurrence of *f*, but we can explain *f* [*c*] after the fact.

We can use Mackie's famous example of three slot machines to illustrate how these ideas work in deterministic and indeterministic contexts. Mackie presents us with three different machines, *K*, *L* and *M*, which all profess to supply bars of chocolate and they all have a glass front that makes it possible to observe their internal mechanisms. Machine *K* is deterministic. That is what we expect of slot machines. It does not always produce a chocolate bar when a coin is inserted, but when it does not, we can always, in principle, find the fault in the machine or the outside interference that caused the failure to deliver. So, in normal circumstances, the insertion of a coin is both necessary and sufficient for the production of chocolate bar. Machine *L* is different. It is indeterministic. It will not, in normal circumstances, produce a chocolate bar when a coin is not inserted, but it may fail to produce a bar even when this is done. This failure is a matter of chance. We cannot explain the individual failures of *L* as we did in the case of *K*. So, in the normal circumstances, the insertion of a coin is only a necessary condition for the appearance of a chocolate bar. The third machine, *M*, is also indeterministic. In ordinary circumstances it will produce a bar of chocolate whenever a coin is inserted. However, it will also occasionally produce a chocolate bar even without insertion of a coin. In such cases the cause is not discoverable even in principle. So in a sense, *M* is the opposite of *L*, and for it the insertion of a coin is the sufficient condition of the production of a chocolate bar. (Mackie 1974: 40-41.)

We can explain why machine *K* produces a bar when a coin is inserted by describing the internal mechanisms that produce this result. Similarly we can explain why it does not produce a bar when a coin is not inserted. We can also explain the failures of the machine, because the glass front allows us to diagnose the failures. In the case of machine *K* we can always (in principle) explain the following *explananda* (among others):

1) the machine does not produce a bar when a coin is not inserted [the machine produces a bar when a coin is not inserted].

2) the machine does not produce a bar when a coin is inserted [the machine produces a bar when a coin is inserted].

3) the machine produces a bar when a coin is inserted [the machine does not produce a bar when a coin is inserted].

4) the machine produces a bar when a coin is not inserted
[the machine does not produce a bar when a coin is not inserted].

With *L* and *M*, explaining all these *explananda* is not possible. In the case of *L*, we can explain 1) and 4), but we cannot explain 2) and 3). Note that we can explain 4) because the production of a bar would be a case of a deterministic mechanical failure observable through the glass front of the machine. Notice also that with machine *L* we can *ex post facto* know the cause of every production of a chocolate bar. It is either the insertion of a coin or a mechanical failure (or their combination). It is just that we cannot explain why a particular insertion of coin produces a bar and why some other insertion of a coin does not.

In the case of *M*, we can explain 2) and 3). Case 2) would again be a deterministic mechanical failure. However, in 3) there are two alternative mechanisms: the appearance of a bar might be a 'spontaneous' random event or it might be the result of the proper coin mechanism. So, in some cases we have an example of a causal preemption: the spontaneous mechanism preempts the production of the chocolate bar by the 'proper' mechanism. According to the account developed in the previous chapter, the explanation would have to mention this possible preemption, although the insertion of a coin ensures that a bar will be produced. But again, we have only two *explananda*, and we cannot explain 1) and 4).

Mackie's example does not mention any specific probability values for the indeterministic processes in *L* and *M*. Maybe there are no such things. But if there were, would it change the situation? My answer is that in the explanation of single occurrence they would not, no matter what the probability values are. Of course, the probabilities would help us in some other ways. They would help us in the theoretical description of the propensities inherent in the machines. It would also make it possible to predict outcomes more accurately. But they would not affect the explanation.

Note that the probabilities discussed above reflect truly indeterministic processes. However, of the most probabilities used in the sciences and in everyday life are not like this. These probabilities reflect our ignorance, and my points do not apply to them. In these cases we have only *partial* explanations. One advantage of my position is that it does not make these explanations look better than they are. The existence of indeterministic phenomena should not lead us to be more satisfied with our usual statistical explanations as they deserve. Some accounts of probabilistic explanation do have this vice, as they accept a partial explanation as a fully adequate explanation.

I conclude that indeterministic causal processes do not produce any fatal problems for my account of causal explanation. I hope that I have also satisfied Salmon, who urges deductivists to consider cases like Mackie's chocolate machines (Salmon 1989: 176-177).

2. Probabilistic theories of contrastive explanation

One approach that I have not considered so far is the suggestion that we should understand contrastive explanation in probabilistic terms. There are slightly different versions of this idea in the published literature. Although they differ in the implementation of this basic idea, they all propose that the explanatory cause makes *f* more probable than *c*.

The first version is van Fraassen's suggestion that *A* should favor *F* against *C*. He interprets this favoring to mean that the *A* raises the probability of *F* in comparison to *C* (van Fraassen 1980: 146-147).[2] Twenty years later Philip Percival suggested that we should require that *A* strongly favors *F* against *C*. Although his proposal was not fully developed, it seems that *A* strongly favors *F* against *C* when it makes the probability of *F* near one and the probability of *c* near zero (Percival 2000: 63). Percival's position is consequently more demanding than van Fraassen's. Elliot Sober, on the other hand, has taken a more liberal stance. His suggestion is that a common cause *discriminates* between *F* and *C*. For him, this means that the common cause makes *F* more probable than *C*. (Sober 1994: 178.)

There is also an interesting suggestion by Christopher Hitchcock (1999) that does not require that the explanatory cause makes *F* probable or that the explanatory factor raises the probability of *F*. In this account it is only required that the explanatory factor make a difference in the probability of *F* against the probability of *C*. I will return to Hitchcock's model after considering the ideas presented in earlier theories.

I think that there are some weighty general arguments against the whole idea of a probabilistic approach to contrastive explanation. I will not go to the details of the suggestions made by van Fraassen, Sober and Percival. Instead I will go directly to the arguments against the underlying idea.

The first argument is actually a question. These probabilistic theories of contrastive explanation start from the idea that an explanation makes the *explanandum* expected. This is the old Hempelian intuition about explanation. Despite this influential source, I think we should ask what explanation has to do with expectedness. Once we have given up the idea about the symmetry between explanation and prediction (or rational expectability) there is no positive reason to entertain this connection. If explanation is really about answering why- and how-questions, the connection between explanation and high probabilities seems arbitrary. Explanation deals principally with mechanisms and processes, not with the probabilities of the various outcomes. It might be that some philosophers have a contrary intuition, but we should ask whether their intuitions are simply products of too much exposure to

the D-N model? Of course, the causal origin of the idea does not invalidate it, but if we have no other justification for it, we should consider whether we should still entertain it. What we need is some positive motive to connect statistical relevance with explanation. The authors under consideration here have not offered any.

The second argument is by Wesley Salmon. Let us call it *the symmetry argument*. Recall our experimental set-up from example 1. The theories we are now considering suggest that we can explain the coin's landing heads in case B, where the switch was turned up, but we cannot explain its landing heads in case A, where the switch was turned down, whereas if the coin had turned tails, we could have explained it in case A, but not in case B. Now Salmon's symmetry argument asks why we could not explain the heads in case A, if we can explain the tails if it were the result of the experiment? After all, the causal facts are the same. The components and mechanisms of the causal process are exactly the same in both cases, only the outcome is different. If the explanation is about understanding causal processes and mechanisms, as suggested in the previous argument, there should not be any difference in the explanations of f [c] and c [f]. We understand the causal process equally well in both processes, so where is the explanatory difference? (Salmon 1998: 152-153.)

The third argument is also by Wesley Salmon. Let us call it *the arbitrariness argument*. Salmon notes that a supporter of the high probability requirement for explanation has trouble with justifying any specific probability threshold for explanatoriness. Suppose that we require that A must make F more probable than C and that we are comparing the following two cases. In the first case the probability of F is .501 and in the second case its probability is .499. According to the view under consideration we should judge that A is explanatory in the first case, but not in the second. This judgment feels arbitrary, since the difference between the probabilities is so small. We can develop a similar argument for any possible probability threshold (Salmon 1998: 151-152). The differences among the suggestions of van Fraassen, Sober, and Percival reflect the problem: they all characterize probability threshold slightly differently. Why? No justification is given. Maybe there is no such thing, and the whole idea of the high probability requirement for explanation is misconceived.

We have still the fourth and, to my mind, the strongest argument. The idea for this argument comes from Woodward (1990), who argues that there are some difficult problems for current theories of causation in the context of applying probabilities to singular causal statements. The basic idea of this argument is simply that probabilities do not reflect facts about singular causal processes, as they should. Consider the following example:

Example 2: A recovery with multiple causes

> Suppose there is a disease that has a spontaneous recovery rate of .05 within two weeks. There are two alternative treatments, which work by completely different causal mechanisms from each other and from spontaneous recovery. There are also no interaction effects between the treatments. Treatment A raises the probability of recovery within two weeks by .47 and treatment B raises the same probability by .28. Both treatments are used to treat a patient who recovers within two weeks. Our explanation-seeking question is the following: Why did the patient recover within two weeks rather than later?

Let us now consider the causal facts of the case. We have at least seven alternative causal histories for the recovery of our patient:

1) the recovery was caused by *a* (spontaneous recovery)

2) the recovery was caused by *a** (treatment A)

3) the recovery was caused by *a*** (treatment B)

4) the recovery was caused by the combination *a* & *a**

5) the recovery was caused by the combination *a* & *a***

6) the recovery was caused by the combination *a** & *a***

7) the recovery was caused by the combination *a* & *a** & *a***.

There are alternative ways of interpreting the combination of causes in cases 4) - 7). The combination of causes can be understood either as causal overdetermination, causal pre-emption, or joint causation. Consequently, we can have up to 20 possible alternative causal histories.

However, from the point of view of my argument, the exact number of alternatives does not matter. What is important is that *the probabilistic theories of explanation rule out most of these causal possibilities on purely conceptual grounds*. Intuitively we think that the explanation should reflect the causal facts of the case. So, if the recovery was caused by treatment B, then the explanation of the recovery is the successful use of treatment B. Similarly, if the recovery was spontaneous, we should not claim that the treatments were effective in this particular recovery. Probabilistic theories do not allow us this kind of sensitivity to the facts. We have to accept 2) (or alternatively 7)) as the right explanation since it makes the recovery most probable. This judgment is not affected by what happens in this particular case. It might be that different probabilistic theories will pick different causal factors (or their combinations) as explanatory, but none of them allows all seven (or twenty) explanatory scenarios. The theories accept that the events might have unfolded according to any of the above causal scenarios, but only one of them can be presented as an explanation, independently of its truth in this

particular case. This is clearly unacceptable: why should statistics about other cases determine what happens in one particular case?

I think that the correct judgment is that with the information provided we cannot yet explain why the patient recovered within two weeks. All causal scenarios should be kept open. What is needed is a closer examination of the patient's recovery process in order to find out how recovery occurred. It might be possible to rule out some of the alternative causes. Of course, it is also possible that we would never find out how recovery occurred. In such a case, we should notice the causal possibilities (and their probabilities) and remain uncommitted to any claims about the true explanation. There is a fact of the matter, but we just do not know it. Theories of explanation should respect this possibility rather than prejudging the issue in the manner of the probabilistic theories under consideration.

A similar lesson can be drawn from other examples, where the actual cause either diminishes the probability of the *explanandum* or it does not make any difference in its probability. (For a wide array of such examples, see Eells 1991.) The crucial point is that the probabilistic accounts do not treat causal facts about the causal history correctly, and they should fail because of this shortcoming. The singular causal explanation should reflect the facts of the particular case, not what usually happens. Similarly, if the causal process producing the event to be explained includes complications like overdetermination, joint causes or causes that diminish the probability of the effect, the explanation should spell out these complexities instead of covering them up with probability statements.

My own suggestion is that we treat probabilistic cases as partial explanations. They provide us part of the explanation without fully explaining the intended contrast. This would allow us to treat probabilistic information as contributing to explanations without compromising our standards for them. This is important because most of our probabilistic explanations reflect more of our ignorance than the causal indeterminacy of the world. In lowering standards by adopting a probabilistic account of explanation, one only fools oneself.

Hitchcock's explanatory relevance model

Christopher Hitchcock (1999) has presented a theory that avoids the first three critiques presented above. The fourth argument applies to his theory, and I consider it to be enough. However, since Hitchcock's paper adds some new elements to the literature on contrastive explanation, I think it is appropriate to comment on some of his ideas and to clarify my own position against them.

Hitchcock argues that we can explain probabilistically why F rather than E happened. His basic idea is as follows (B refers to the background conditions):

> … we should view A as explanatorily relevant to the contrastive question 'why E rather than F' if A continues to be relevant to E when the (exclusive) disjunction $E \vee F$ is held fixed. […] this means that A is explanatorily relevant to E rather than F when $P(E \mid (A \,\&\, B) \,\&\, (E \vee F)) \neq P(E \mid B \,\&\, (E \vee F))$ (Hitchcock 1999: 587).

Now, the most important change that Hitchcock makes to the settings of the earlier discussion is that he speaks of *explanatory relevance* instead of *explanation*. This change is very significant. According to the Leibniz principle, *a* cannot *explain* (in similar circumstances) both *f* and *c* if they are incompatible with each other. However, the principle says nothing about *a*'s being explanatorily relevant.

Let us have a look at Hitchcock's central example to see how his idea works. Suppose a vertically polarized photon approaches a polarizer that is tilted at an angle from the vertical. Now this is, according to our current knowledge, a chance set-up. There is a certain chance the photon will pass through the polarizer and a certain chance that it will be absorbed by it. (Hitchcock 1999: 585-586.) Let us suppose that that the photon was transmitted. Now the Leibniz principle would say that we cannot explain why the photon was transmitted rather than absorbed. Hitchcock wishes to disagree. His motivation becomes understandable when we add more information to the example. The fact to be added is that the probability of transmission varies with the relative orientation of the photon and the polarizer. Depending on the angle of contact the probability varies between zero and one. Hitchcock's point is that the relative orientation of photon and polarizer is explanatorily relevant because it affects the probability of the transmission. (Hitchcock 1999: 596.)

Now, I agree with Hitchcock that the relative orientation is explanatorily relevant. However, the big question is: *for which explanandum it is explanatorily relevant?* When we ask why the transmission had a certain probability rather than some other probability, the reference to the relative orientations of the photon and the polarizer is certainly appropriate. After all, the angle makes the difference between probabilities of transmission. In other words, it *explains* the probability in question (in contrast to some alternatives). The possible controversy concerns the explanation why the photon was transmitted rather than absorbed. Again, nobody would deny that the relative orientation of the photon and the polarizer was a *causally* relevant factor in the process of the transmission of the photon. There is no disagreement about facts of the example. The issue concerns the explanation of a singular event.

Now, let us suppose that someone asks why a particular photon was transmitted rather than absorbed? Would there be some differences between the explanatory answers given by Hitchcock and by more conservative theorists? The answer Hitchcock would give to this question

would refer to the probabilistic nature of the process and to the role of the relative orientation of the photon and the polarizer in this process. After all, there is some causally relevant information to be provided. How would Hitchcock's opponent answer this question? I do not think that she would provide 'nothing' as her first reply to this question, as Hitchcock seems to imply. She would probably give exactly the same answer to the question, but she would add that these facts are not enough to explain the contrast. So with this explanation-seeking question we cannot establish the real difference between the positions. Now let us move to the follow-up question: *given that the transmission had a certain probability to occur, why was the photon transmitted?* I think that the both parties would give exactly the same answer here: 'we can't explain that, because it was due to chance'. These considerations seem to suggest that there is no real disagreement between Hitchcock and me as to how to handle this example.

Since Hitchcock cannot show that any real advantage derives from his proposal, I think we should not give up Leibniz's rather intuitive principle. Rejecting it might lead us to situations where explanation comes too easy and is intuitively non-satisfactory. Hitchcock might reply to this by claiming that his opponents cannot really do justice to explanations that do not fit their ideal standards. By this he would point to the fact that the intuitive answer to the above example is not really permitted by his opponents' standards of explanation. I think this argument is a non-starter. Hitchcock's points can be easily accommodated to the more traditional account without giving up the Leibniz principle.

First, we should distinguish between an explanation of an event and an explanation of that event's probability. Hitchcock might be confusing two different *explananda*. When explaining the probability of the individual event, Hitchcock's considerations are relevant. But they are not relevant when we are explaining a singular occurrence. And this is the only case where his considerations would contradict the received view. When we are explaining a singular occurrence, which is, as I have already noted, of peripheral scientific interest, the thing either happens or it does not. This is all there is to it. I cannot see how the probability of this occurrence could have any causal influence here. Suggesting otherwise would imply that chance is one causal factor among others, and this would be a category mistake (Humphreys 1989).

Second, despite the fact that I set high standards for explanatory adequacy, it does not follow that I do not see the value of partial explanation. Sometimes a more detailed description of the process leading to the event to be explained is valuable without being able to explain the contrast we have in mind. Sometimes this description can also work as a correction to the question with an inadequate assumption. A partial explanation is better than no explanation at all, so I think Hitchcock's worries can be handled.

3. The third dogma of empiricism

The main motivation for the idea that a complete explanation includes or refers to covering laws is a thesis that Wesley Salmon calls 'the third dogma of empiricism'. This 'dogma' says that all scientific explanations are arguments (Salmon 1998: 95). The dogma regards scientific explanation as an inference from the statement of the *explanans* to the statement of the *explanandum*. The explanation has a deductive structure in which the *explanans* logically entails the *explanandum*. Salmon's definition of the third dogma looks simple, but when one considers his discussion as a whole, the issue becomes more complicated. The dogma can be given various interpretations. Consider the following theses:

1) All explanations can be *reconstructed* as deductive arguments

2) Deduction has *a constitutive role* in explanation

3) An explanation *involves* a deduction from *a covering law*

4) An explanation is nothing but a deduction from one or more covering laws (and the statement of the initial conditions)

Salmon's official definition can be interpreted as claiming 1) or 2), but his arguments against the dogma are directed mainly against 4). This leaves open the exact content of the dogma. To get started, let us first consider the three arguments Salmon himself presents against the dogma, and let us see where they lead.

Salmon's first argument is based on the observation that the addition of irrelevant information is harmless to an argument, but fatal to an explanation. In deductive logic irrelevant premises do not undermine the validity of the argument. Of course, there is no point in adding such premises, but importantly, they do not affect the validity of the argument. With explanation things are different. It seems that in explanation only relevant information should be included. If one adds to the explanation of the dissolving of a piece of sugar to the water the information that the water in question is holy water, one ruins or worsens the explanation. (Salmon 1998: 96-97.)

I think we should judge that this argument is less than fatal. It does not show that explanations are not arguments; it only suggest that explanation involves something more than a simple deductive structure. Salmon's argument works against the position that regards deductive structure as a sufficient condition for explanation and against Hempel's position in particular. But if the supporter of the dogma has some extra conditions for explanatoriness in his analysis, the argument does not work. In such circumstances one can claim that the added irrelevant premises are not a part of the explanation. And as they are not parts of the explanation, they do not worsen or ruin it, as Salmon claims.

There is something right in Salmon's observation, but it is not re-

lated to explanatoriness *per se*. Rather, it is related to communication in general. In the delivery of explanatory information, as in all communication of information, the addition of irrelevant information does damage to the message. The recipient of the explanation may not be able to distinguish the information that is intended to be explanatory from the background noise. This general observation has some important consequences for the pragmatics of explanation, but it does not concern the analysis of what makes an explanation explanatory.

Salmon's second argument concerns the possibility of explaining random events. When explaining outcomes of indeterministic processes, the *explanandum* is not logically entailed by the *explanans*.[3] According to Salmon, the deductive idea connects explanation too closely to the validity of the doctrine of determinism. Developments in physics have shown that the idea of indeterminism should be taken under serious consideration. In this situation a precommitment to the idea of determinism is a clear disadvantage for a theory of explanation. It is more preferable that a theory of explanation remains neutral on this issue and is compatible with the possibility of indeterministic causation. As the scientific evidence seems to strongly support the idea that there are truly indeterministic processes in the world, the deductive ideal of explanation seems to be in trouble. (Salmon 1998: 97-101.)

Again, this is not a knockdown argument against the deductive ideal. A deductivist can reply to Salmon in various ways. I developed one such a reply earlier in this chapter. There is no essential connection, or even tension, between the deductive ideal of explanation and the issue of indeterministic causation. One can easily accept both. This makes Salmon's second argument a non-starter. In line with my earlier analysis, one can accept that *some* explanations in indeterministic contexts are doomed to be partial or that we cannot explain singular outcomes of indeterministic processes. As my analysis shows, this does not apply to all *explananda*, so there is much space for Salmon's intuition that we can explain *something* in the indeterministic contexts. My discussion also shows that the position that is in more trouble is Salmon's own account. After giving up the deductive idea, he really has no resources to explain what it is about the random event that a given explanation explains The requirement that the statement of the *explanandum* should be logically entailed by the *explanans* gives a neat way to show the explanatory scope of the *explanans*. The aspects of the *explanandum* event that are accessible by deductive reasoning are inside the scope of explanation, and the rest is beyond it.

Salmon's third argument is based on the idea that explanation requires temporal asymmetry, whereas arguments are not subject to similar constraints. Salmon points out that in common counterexamples to the D-N model, such as in the explanation of the length of the flagpole by the length of its shadow, the requirements of the D-N model do not

respect the intuition concerning the asymmetry of the causal explanation. (Salmon 1998: 101-104.)

In principle, this argument works against Hempel's version of the D-N model. However, although it shows that deductive structure is not sufficient for the explanation, it does not show that it is unnecessary. If the supporter of the D-N model adds to her model a requirement concerning causality, as a number of authors have done, she escapes Salmon's critique. Clearly, this argument is only compelling against thesis 4).

Summing up the arguments so far, it seems that Salmon is not successful in providing good reasons for giving up the third dogma, at least as he himself formulates it. He only gives us some reasons to give up, or improve, Hempel's theory of explanation. Do we have some other arguments against the dogma? What about van Fraassen's (1980: 134) thesis that explanations are not arguments, but answers to explanation-seeking questions? This avenue of argumentation is another non-starter. One can accept both ideas and take them to be complementary ideas. An explanation can be an argument and still an answer to an explanation-seeking question, as Hempel saw. Similarly, I think that Achinstein's (1983: 81-83) worries that the D-N account cannot handle the role of emphasis in explanatory statements can be shown to be less than fatal to the deductive ideal of explanation.

But are there any positive motivations for remaining attached to the dogma? After all, if there are no positive reasons for holding the view, even very weak arguments can be enough for giving it up. In my view there is one positive argument for holding on to the dogma. The deductive ideal provides a useful test for the explanatoriness of an answer to an explanation-seeking question. One should be able to deduce the statement of the *explanandum* from premises that include both the statement of the *explanans* and the background assumptions of the explanation-seeking question. If this is not achieved, the explanation remains partial, no matter what its other merits are. On the other hand, if this test is passed, it is *possible* that we have an adequate explanation in our hands.

My account diverges a bit from the standard analysis behind the D-N model. I do not require that one must be able to deduce the statement of an *explanandum* from the *explanans* alone. One needs help from the presuppositions of the question. This way of identifying the *explanans* is different from the standard D-N account. For a D-N theorist the *explanans* is the set of premises from which the statement of the *explanandum* can be deduced. In my view, the *explanans*, that is the explanation, is the (adequate) answer to the explanation-seeking question. My position is not unique, it is supported by most pragmatic theorists of explanation (cf. van Fraassen 1980; Sintonen 1984).

The difference is not purely terminological. My account fits every-

day explanatory practice better. First, in the D-N account, practically all answers to explanation-seeking questions are deemed to be incomplete explanations, since they do not state all the relevant premises. This gives the wrong impression that these explanations, or explanation sketches, are somehow inadequate. My account does not have this unintuitive consequence. In it the *explanans* is the answer actually given. The deductive structure is only achieved by adding the presuppositions of the question as extra premises. Secondly, my account gives a more central role to explanation-seeking questions than the D-N model. A question is essential to the identification of the right *explanans*, whereas in the D-N model the same *explanans* can in principle answer many different questions.

These considerations show that if the deductive ideal is understood in the sense of thesis 1), it cannot be readily dismissed. Notice also that nothing in the above discussion shows that the acceptance of the deductive ideal commits one to the idea that a singular causal explanation necessarily includes a reference to a covering law or laws. These two ideas should be kept distinct. To see why this is important, let us consider an argument presented by David-Hillel Ruben. The argument shows that the only way to save the motivation for the D-N model is to introduce considerations that destroy the anticipated fruits of this approach.

Ruben's argument

Ruben starts by noticing that in order to cope with the standard counterexamples based on irrelevant information, causal preemption and considerations of explanatory asymmetry, the covering law account must be strengthened by including considerations of causality (Ruben 1990: 183-194). But is this enough, and how it should be done? Ruben considers Timothy McCarthy's (1993) argument to the effect that it is not enough to add to Hempel's standard conditions the requirement that there be a premise which mentions the actual cause of the event to be explained. McCarthy argues that it is always possible to construct derivations that meet these extended requirements, but which are still not explanatory. To follow Ruben's discussion, let 'D(*e*)' be a sentence describing the *explanandum*, 'C(*e*)' a sentence describing the actual explanatory cause (*c*) of *e*, *o* any object such that A*o*, and let '(*x*) (A*x* -> B*x*)' be any (causal or non-causal) law utterly irrelevant to the occurrence of *e*. Let us now consider the following derivation:

(1) (*x*) (A*x* -> B*x*)

(2) C(*e*) & A(*o*)

(3) ¬ B(*o*) ∨ ¬ C (*e*) ∨ D (*e*)

(C) D(*e*)

This derivation meets all of Hempel's own conditions and also the added requirement that the explanation should mention the cause of the *explanandum*. Furthermore, as 'C(*e*)' has an essential role in the derivation, the example is not arbitrary in this sense. But, as everybody would agree, this argument does not explain D(*e*). The derivation mentions cause *c*, but it gets us from the cause to the *explanandum* via a law irrelevant to *c*'s causing *e*, rather than by asserting that *c* is the cause of *e* (McCarthy 1993: 130-131). The problem with the example, and others like it, is that the deductive relation between *c* and *e* is causally and explanatorily irrelevant to the occurrence of *e*. There is something to the notion of explanatory relevance that even the causally strengthened deductive model cannot capture. (McCarthy 1993; Ruben 1990: 194-196, 249-251.)

Now, according to Ruben, there is a very simple way to bring the cause and the *explanandum* together in the right and relevant way. We only have to require that the derivation include as a premise a singular statement which asserts with respect to cause *c* that it is the cause of *e*. We capture the explanatory dependence by an explicit statement of the causal dependence: '*c* caused *e*'. The premise does not need to use the word 'cause'; it could as well assert that *e* occurs because *c* occurs, or that *c* is the reason for the occurrence of *e*. What could be more simple? (Ruben 1990: 196, see also Achinstein 1983: chapter 5.)

This suggestion has an important consequence. Hempel's requirement that there be a lawlike generalization in the premises that is essential for the derivation becomes unnecessary. Following Ruben's suggestion we would have a premise that says that the cause of the *explanandum* event is such-and-such. This premise would entail the statement of the *explanandum* without any help from the other premises. If the only motivation for the inclusion of laws in the explanation is the deductive ideal, Ruben's suggestion makes them redundant. An explanation of a singular fact does not require a generalization as a premise; one can deduce '*e*' from the '*c* caused *e*'.

For Ruben, this observation has also another interesting consequence. The derivation of '*e*' from '*c* caused *e*' is trivial, and all the other parts of the derivation are redundant. In the light of this, it is much simpler to think the explanation as consisting of a singular sentence, '*c* is the cause of *e*' or '*e* because of *c*'. The whole idea of explanations as arguments becomes unmotivated. Ruben concludes that explanations are not arguments, but simply sentences (Ruben 1990: 197).

When considering first the destiny of the D-N model, it seems that the required cure really kills the patient. The only way to save the covering-law idea leads to the considerations that are its demise. After all, the original point of the approach was to make sense of sentences like '*e* because of *c*'. If the *analysans* includes the statement of *analysandum*, the analysis has gone in circles, and the appropriate question is not

whether we have advanced, but whether we have been lucky enough to avoid more confusion.

But what are the consequences of Ruben's first argument for the supporter of the deductive ideal who does not require the presence of covering laws in the argument? The situation is less clear. Ruben's first argument clearly supports the idea that the deductivist position is distinct from the D-N account. The acceptance of the deductive ideal does not automatically lead to the acceptance of the D-N account of explanation. There is no compulsory move from thesis 1) to thesis 3).

What can be said about Ruben's second observation from this point of view? First, it helps us to make difference between the positions subscribing to theses 1) and 2). They are not necessarily the same things. Ruben's sentence explanations can be reconstructed as deductive arguments, but deduction is not constitutive of their being explanations. Their explanatoriness derives from some other source. One can be even more liberal than Ruben. One can allow that we can explain also by pictures, diagrams, gestures and physical models, and still hold that these explanations can, in principle, be reconstructed as deductive arguments.

Second, Ruben's view does not show that thesis 1) is false. One can accept that in his example deduction is a trivial exercise. This might be true also of more realistic, or less abstract, examples. But note, this only applies when we actually have an explanation. If our explanatory information is not really explanatory, the deduction does not go through. This supports the idea that the deductive ideal can still serve as a test for explanatoriness. The deductive ideal is not necessary from the point of view of the theory of explanation developed in the previous chapter, but I think it is a useful addition to it, and consequently worth preserving.

The above discussion shows that there are good reasons to think that the deductive ideal does not directly support the idea that singular causal explanation essentially involves reference to a covering-law or generalization. Deductivism and commitment to covering laws in explanation are separate issues. The acceptance of the former does not commit one to the latter. But the requirement of covering laws has also been defended on other grounds than the deductive ideal. Now it is time to turn to these arguments.

4. The role of laws in explanation

We can distinguish three different arguments for the necessity of laws in singular causal explanation. Let us go briefly through these arguments before considering the arguments against this requirement.

The first argument is epistemological. It says that unless we suppose that there are certain uniformities in the nature, we lack justifica-

tion for our singular causal claims (Hempel 1965: 360). The idea is that without the knowledge of the appropriate laws, we would have no grounds for distinguishing between true and false causal claims. Our explanations would lack justification.

Already Michael Scriven has shown that this argument is not viable. As Scriven insists, we should distinguish between explanatory claims and our grounds for accepting these claims (Scriven 1959: 446-450; 1962: 196-200). It is clear that we can use laws and other generalizations to justify our causal claims. However, this does not imply that we should include these justifications in our explanations. Hempel does not require that we include justification for non-causal claims, so in the name of consistency, why should we include justification for causal claims in the explanation? This asymmetry would require some justification, and no such thing is provided. To the contrary, there are good reasons not to require that the justifications be included in the explanations. First, their inclusion would call for an infinite regress: if we include justifications for the explanatory facts, should we also include justifications for the justifications? There is no natural boundary to stop the regress if it gets started. Secondly, in explanation, we are not looking for reasons for our belief in the *explanandum*: we presume that it is the case, and we are asking *why* it is so. Similarly, explanation presumes that the facts it cites are true, it does not look for justification for our belief in them. Our explanation explains if it is true; the grounds for believing it are a separate question.

The second argument for the necessity of laws in causal explanation is semantical. It states that causal explanatory claims tacitly imply that there are appropriate laws that cover them (Hempel 1965: 349). Again, I think that there are good reasons for not going along with this argument. The argument seems to presuppose that some kind of regularity account of causation is correct. But this presupposition is clearly unjustified: it is not the case that a singularist account of causation is wrong on purely semantical grounds. To the contrary, it seems that most philosophers of causation nowadays regard regularity accounts of causation as highly problematic and suspect. It would be unfortunate to bind the account of explanation to such a weakly supported metaphysical position. And if this is to be done, it should not be done on purely semantical grounds. Secondly, even if a causal explanation would imply that there is an appropriate covering law, this would not necessarily imply that this covering law has *a constitutive role* in the explanation. The supporter of Hempel's position would have to present some argument to the effect that covering laws are not just epiphenomenal on explanatory causes. The only way to do this would be to claim that the regularity constitutes the causality, which would bring us back to the problematic empiricist account of causation. Clearly, the semantical argument is as unconvincing as the epistemological one.

The third, and the final, positive argument is based on the ideal form of scientific knowledge. This is the primary motivation for Peter Railton's DNP model of explanation. According to Railton, "the fundamental insight of the covering-law approach to scientific explanation is that science seeks theoretical understanding, guided by a nomothetical ideal" (Railton 1980: 117). The basic goal of science in this account is the subsumption of particular facts and regularities to the 'nomic nexus'. This is achieved by fitting the world's phenomena into a fully general and comprehensive theory (Railton 1980: 117-119.) In Railton's vision, the ideal for which science strives is the 'ideal explanatory text', that would be able to explain every aspect of the phenomenon under consideration and would that also have a deductive-nomological structure. He notes, that "… plainly there is no question of ever setting such an ideal text on paper" (Railton 1981: 247), but he wants to underline the regulative role of such ideal. The aim of science is to develop a *capacity* to provide material for such ideal explanatory texts.

I do not want discuss the merits of Railton's ideal as a description of the goal of scientific knowledge. Clearly, he describes a picture of science that is accepted by many philosophers of science. Supposedly the ideal has also motivated quite a few covering-law theorists. However, this picture does not in any way show that laws have a constitutive role in actual explanations. One could agree with everything Railton says and hold that the fundamental aim of science is to create a general theoretical conception that reveals the nomological structure of the world, but still maintain that laws do not have a constitutive role in the singular causal explanation. Actually, this seems to be even Railton's own position, since he holds that we can perfectly well explain without recourse to covering laws (Railton 1981: 249).

Clearly, the question concerning the ideal form of scientific knowledge is a separate issue from the question 'what makes a given explanation explanatory?'. The idea of the ideal explanatory text does not give us any hint about the principles that govern the choice of explanatory information. It requires an addition from some other theory of explanation to be able to handle this central problem. On the other hand, as I argued in Chapter 2, the contrastive approach to explanation can be used to characterize the concept of the ideal explanatory text. The elements of the ideal explanatory text for event e are explanatory because they are parts of adequate answers to $some$ contrastive questions about e. This asymmetry between the two accounts of explanation suggests that the contrastive approach is conceptually more fundamental.

Careful consideration of these arguments shows that the positive grounds for the belief that laws have a constitutive role in singular causal explanation have always been quite shaky. The position should have somewhat stronger grounds for it to be taken seriously. Contrary to a common presumption, the burden of proof is with the supporters of

the covering-law account, not with its opponent. Furthermore, there are weighty arguments against the idea. Let us next consider these arguments briefly.[4]

The arguments against laws

First, consider the fact that our everyday explanations do not mention laws. This goes also for most explanations in the sciences. Furthermore, laws are not even tacitly implied by these explanations, since often people do not know the appropriate laws. Either the laws are totally unknown, or we do not know them in the form that would allow us to make appropriate derivations. These facts are something that the supporters of the D-N model grant to their critics, but I think that they have not considered all the implications of this observation.

Consider next James Woodward's plausible *desideratum* for a theory of explanation:

> A theory of singular causal explanation ought to identify the structural features of such explanations which function so as to produce understanding in the ordinary user. Features which are entirely unknown to the ordinary user are not plausible candidates for what produces such understanding. (Woodward 1993: 249)

According to this criterion, the features that produce explanatory understanding should be *epistemically accessible*. (Woodward 1986: 269; 1995: 249.) The problem for the covering-law account is that it requires that the *explanandum* be deduced from general laws and a description of the initial conditions. Now, as both the law and the deductive argument are ordinarily unknown, the question is how does the ordinary explainer identify the explanatory factors and separate them from the non-explanatory factors? Does she have some sort of intuition that recognizes the existence of these unknown factors, or is she just guessing? The first sounds mystical and the second implausible. The supporter of the D-N account owes us 1) an explanation for how we can judge the explanatoriness of causal claims and 2) an account of the relation between the principles we use in these judgments and the conditions spelled out by her model. I don't claim that this is impossible, I just note that no one has done it. As long as these questions remain without answers, the covering-law account cannot be taken as a serious explication of the ordinary notion of (scientific) explanation. Of course, it is always possible to take the stipulative strategy and treat the covering-law account as a definition of explanation, but what is the point of that exercise?

There are some further difficulties for the covering-law account. One clear problem is that apparently there are no covering laws for most

explananda. For example, Hempel owes us a justification for the assumption that there are enough *covering* laws for all acceptable explanations. Recall that he says that explanation is always under a certain description, but he does *not* say that we can only explain things that are described by using the vocabulary of the theoretical sciences. Even his examples of explanation make use of concepts and descriptions that come from our everyday affairs rather than from the theoretical sciences. The problem is that the sciences do not contain laws that cover all possible descriptions that we might wish to use in our explanatory questions. It is controversial whether so-called special sciences have laws, and if they have, what kind of laws they are, but most people agree that there are no laws for every description.[5] We do not have laws about shattering windows, lighting matches, or failing carburetors, as described. Even when we do have some 'laws', they are not of the right kind. These laws use descriptions that are alien to our *explananda*, and consequently, they are not capable of functioning in logical derivations as Hempel wishes they would.[6]

A scientific explanation that intends to make use of laws, or that aims to give an informative theoretical account of the *explanandum*, must redescribe, or reconstruct, the *explanandum* in its special theoretical vocabulary (Woodward 1979: 62). If we want derivations from scientific laws, we have to describe the *explanandum* appropriately. And this reconstruction will, at least sometimes, change the *explanandum*, so that we are not necessarily explaining the same thing anymore. Hempel faces a dilemma here: either he has to restrict the scope of explanations to the *explananda* that are using the vocabulary of our theoretical sciences, or he has to provide us with an account of these other covering 'laws'. This account would have to justify the scientific status of these 'laws', to show that they are available, and finally to show that there is a point in calling these generalizations laws.[7] All this work still remains to be done. There is a very natural reason for this: there is still no consensus about the whole notion of law of nature. As long as this issue remains open, the most substantial part of the D-N account remains open.

To add one more item to the list of things that covering-law theorist should explain, consider our everyday intuition that causation is a local process. The Titanic sank because it collided with an iceberg. The covering-law theorist would like us to explain this unfortunate incident by showing that there is a law to the effect that all the boats similar to Titanic sink when they collide with icebergs (that are similar to the iceberg that Titanic collided with) in an appropriate way. My concern now is not with the form of this law, but with its function in the explanation. The question to ask is *how the knowledge that all similar boats sink under these circumstances helps us to understand why the Titanic sank*.[8] Our original question concerns one particular local causal process, and the

explanation of the covering-law theorist refers to other similar causal processes that have no connection to the process we are interested in. How can noting that there are (or can be) similar incidents make us understand this particular incident? I have to confess that I do not understand. If you accept an empiricist account of laws, then generalizations are just summaries of particulars. What emergent explanatory resources does the generalization have that its instances lack? (See also Ruben 1990: 204.) As I see the situation, we do not need laws to explain why the Titanic sank. We only need to know the particular causes at work in this particular case.[9]

Of course, we can further ask why, or how, ships like the Titanic sink when they collide with an iceberg. Indeed, this seems to be the kind of question that pure basic science tries to answer. My claim that we do not need laws to explain singular events does not imply that laws are completely uninteresting or that they cannot be used to explain anything. Apart from explaining singular occurrences, science also aims to uncover regularities in nature and to explain them. Notice that when we explain why ships like the Titanic sink when they collide with an iceberg, the covering-law theorist's generalization reappears. However, it is important to recognize that its not a part of the *explanans*: it works as the *explanandum*. This is quite typical in special sciences: the 'covering laws' are things to be explained, not things providing explanatory understanding (cf. Cummins 2000: 119-122).

I think that these considerations are enough to justify the position that does not take laws to have a constitutive role in the singular causal explanation. Laws (or generalizations) have a more essential role in other forms of explanation, but these cases have not been the topic of my discussion. I am not claiming that there are no scientific explanations that are derivations from laws. Evidently there are such things. Neither am I claiming that scientific explanations cannot be represented as arguments. To the contrary, I accept this thesis. For example, there might be some good pedagogical reasons to represent an explanation as an argument and in this way display its tacit presuppositions. It is also true that when one's science is privileged enough to have quantitative theories, a derivation is a natural way of demonstrating the dependence between the explaining factors and the factors to be explained. This is the only way to represent precisely the quantitative relations between the variables. This rationale is missing from the usual examples of covering-law explanation. In the first place, the laws in these explanations are non-quantitative. Second, they are singular explanations, whereas scientific derivations concern law-like generalizations. Furthermore, the laws in Hempel's examples are usually mere generalizations, not statements of functional interdependence. They do not allow the derivation of other *explananda* that are appropriately different from the actual *explanandum*. (Woodward 1979: 46-49.)

The relation of my account of explanation developed in the previous chapter to the requirement that covering laws should have a role in a singular explanation is analogous to the situation with the third dogma. My account can have it both ways. Even if covering laws had an essential role in all explanations, my account would not suffer. However, on the basis of the above discussion, it is prudent not to include this requirement in a theory of explanation. I cannot see any positive motives for doing so, but I can see a lot of problems. The question to be asked is not why I am not a covering-law theorist, but why anyone should be one.

An alternative role for laws in singular explanation

In the above discussion, laws play a minimal role in singular causal explanation. However, there are roles for laws in explanation other than being parts of it. First, we have already mentioned their role in justifying singular causal claims. But there is also another important role. This role concerns the discovery of explanatory causes. The central function of theories in the sciences is not to provide generalizations or 'laws' that cover individual types of separate events. Rather, it is to serve as general schemes for constructing explanations. Theories are not just collections of covering laws, but accounts of systematic dependencies between the variables. As far as singular causal explanations are concerned, the utility of theories is that they allow an easy and convenient way to construct explanations. With a good theory, all one needs to do is to check the initial conditions and parameters, and then ensure that the provisos of the theory are satisfied; and the theory will then lead one to the explanatory causes. The generality of theories does not make them explanatory, rather it makes them very useful in constructing explanations.

This observation suggests a positive role for various truisms, rules of thumb, tendency statements and other generalizations that are found in the social sciences. Their central role is not to work as premises of an explanatory argument, but rather to work as heuristic rules for the search of explanations. They tell us where to look for the explanatory factors. These generalizations apply only to limited domains and can include numerous exceptions, which makes them less than laws, but they can still be useful. They tell us what kinds of causes are worth looking for. And if we can spell out the exceptions and the intervening factors, we know what kind of factors to look for when the truism does not seem to apply. Consequently, the extreme *ceteris paribus* nature of these generalizations is not a problem in this view. As their role is heuristic, and not explanatory, the fact that they are not really laws does not matter.

Notes to Chapter 3

1 There are some contexts where a reference to dormitive power can be informative and explanatory. If we do not know that opium has such a power and we want know why a person, who has had a dose of opium, fell asleep, the reference to opium's dormitive power is appropriate. However, a reference to a dormitive virtue as an answer to question 'why does opium cause sleep?' is completely unexplanatory. This observation concerns explanation by a dispositional property both in scientific and everyday reasoning. I will return to this issue in Chapter 4.

2 This can be understood for example as:
$P(F|A) > P(F)$ & $P(C|A) \leq P(C)$ & $P(F|A) > P(C|A)$. However, other reconstructions are also possible. In the following, letters refer to the types, small letters to the tokens.

3 Salmon also discusses Hempel's inductive-statistical model of explanation. I will not consider it here since I take the very idea of valid inductive argument to be absurd. By accepting inductive-statistical explanations one gives up the deductive ideal of explanation. If one wishes to remain faithful to the original ideas behind the D-N model, one should use Railton's DNP model of statistical explanation (Railton 1978, 1980, 1981).

4 My critique of covering laws is not a critique of the assumption that there are laws about causal processes. This is a background assumption of any causal inquiry. My critique is aimed against covering laws and their role in explanation, not against the ontology of 'backing laws'.

5 For recent discussion see, Schiffer 1991, Henderson 1993, Kincaid 1996, McIntyre 1996, Woodward 2000.

6 Hintikka and Halonen (2001) present a metatheorem, called the covering-law theorem, that is an application of Craig's interpolation theorem. It shows that if the explanandum can be deduced from the background theory and initial conditions, and these premises can be formulated in first-order logic, a covering law always exists. I find the premises of this argument problematic, whereas the argument itself is rather trivial. First, it is quite ambitious to assume that the *explanandum* can be deduced from the background theory (and initial conditions). This presupposes some way to connect the terms used in scientific theory and the terms used in the description of the *explanandum*. And this is precisely the point I am making: scientific theories do not use kind of vocabulary normally used in the description of the *explanandum*. Second, the assumption that a scientific theory can be formulated in first-order logic seems very strong. However, the most important point is that the covering-law model requires laws that are explanatory, not just any logical constructions that can work as general premises of a deduction. As Hintikka and Halonen themselves note, their argument does not save the covering-law account. For them, a covering law is not a vehicle of explanation, but rather a summary of an explanation. The explanatory work is done by the background theory.

7 Scriven's (1959) discussion of truisms and the problems in formulating appropriate covering laws is relevant here.

8 Note the following difference between a covering-law theorist and a supporter of contrastive approach. The covering-law theorist is interested in other sinking ships because they provide confirmation for his law claim.

The supporter of contrastive approach is also interested in other ships, but she does not want information of the exactly similar ships in the precisely same circumstances as the covering-law theorist does. She would like to know about slightly different ships in the same circumstances or about similar ships in different circumstances in order to find interesting differences to be explained.

9 This claim makes an assumption about causation. I assume that the pure regularity view of causation is false, and that singular causes are primary to causal regularities. The regularity theory asserts that *a* causes *b*, because there is a constant conjunction between *A* and *B*. Common sense claims that there is a constant conjunction between *A* and *B*, because singular *a*'s cause singular *b*'s. I side with common sense; token causes are ontologically more primary than type-level regularities. (cf. Hausman 1998) The regularity view seems to spring from a sloppy reading of Hume (cf. Strawson 1989), and from a confusion between epistemology and ontology of causation.

Chapter 4
The problem of macro explanation

In this chapter, I will discuss the problem of macro explanation in order to illustrate my account of causal explanation. I will first give a general characterization of the problem of explanatory epiphenomenalism of macro properties in section 1. This problem arises from widely accepted premises, and it is discussed especially in philosophy of mind. The problem concerning mental states is the following: either mental states are identical with physical states, or they are epiphenomenal from the point of view of causal explanation. However, I argue that the problem is more general. The main argument of this chapter is that this problem does not arise when one accepts the account of causal explanation developed in previous chapters.

In the second section, I will illustrate the problem of macro explanation by using the concept of fitness from the theory of evolution. This example has certain advantages over more standard illustrations derived from philosophy of mind. In the case of fitness we have a developed account of both and micro and macro levels, whereas in the case of mental causation we have neither. This gives the example some respectability, and it also shows that the problem of macro explanation is not a product of pure science fiction. My discussion of evolutionary explanations will also be relevant to Chapters 6 and 7, where I discuss non-intentional filtering explanations.

The third section will discuss the program model of explanation developed by Philip Pettit and Frank Jackson. After introducing the basic idea, I will show how the model solves the problem of macro explanation in the case of fitness. By assessing Pettit and Jackson's model, I will argue that it has some philosophical problems that can be solved by accepting my account of causal explanation.

In the final section, I will briefly discuss the application of my model to intentional explanation. I argue that the model justifies the common sense attitude which regards intentional explanations as causal expla-

nations. I also argue that in common sense explanations the agent's practical reasoning works as an explanatory mechanism that is used to answer the underlying how question. The account developed in this section will serve as a starting point for the discussion of interest explanations in Part II.

1. The problem of explanatory epiphenomenalism

Recently the problem of macro explanation has been discussed intensively in philosophy of mind. The issue has been the causal efficacy and explanatory relevance of mental states and properties. In a physicalist picture of the world, all causal processes are considered to be physical processes. This implies that all events have a causal explanation in terms of the physical processes that produced them. The problem concerning mental states is the following: either the mental states are identical with the physical, or they are epiphenomenal from the point of view of causation. Both options have their problems. There are good arguments against the identity theory, so it is not very eagerly accepted. On the other hand, we have very strong intuitions suggesting that mental states are causally effective and that they are needed in explanations of human behavior. (See articles in Heil and Mele 1993; Kim 1993.)

It is easy to see that the problem of explanatory epiphenomenalism is more general if one considers the terms under which it is discussed. The concepts of causation, explanation, supervenience, identity, realization, etc. work similarly, irrespective of the context. The discussion does not give any essential role to the concepts that are especially related to mental phenomena. The same problem can be created every time we use causal descriptions and explanations at different levels of description.

To get a hold on the problem, let us start with three generally accepted presuppositions that constitute the problem. The acceptance of these theses is not the only way to create the problem of macro-causal explanations, but it is very typical.

The first thesis goes as follows:

1) Reality (or parts of it) can be described at different levels of description, and these levels of description form a hierarchy of micro and macro levels.

The first part of the thesis seems quite intuitive. We know from our experience that the same phenomenon can be described by different vocabularies. Different scientific disciplines can provide these alternative vocabularies. For example, if we are describing a person running, we can describe her activity by using microphysical, chemical, physiological, psychological, and maybe even sociological theories and concepts.

The idea of hierarchy requires a bit more immersion into philoso-

phy of science. Philosophers commonly think that various sciences form a hierarchy of levels, where microphysics is at the bottom, and macrophysics, chemistry, biology, psychology, and social sciences build upon each other. We have, for example, a continuum of elementary particles, atoms, molecules, cells, organisms, persons, groups, and societies. The latter-mentioned sciences describe entities that are made of the entities of the previous level. I do not claim this picture represents the relations between the sciences correctly. After all, the relation between neurophysiology and psychology is not quite analogical to the relation between elementary particle physics and physics of macro objects. However, it can still be accepted that our scientific worldview includes a hierarchy of micro and macro levels.

We can formulate the next thesis as follows:

2) Things at the macro level are the way they are *in virtue* of how things are at the micro level.

This thesis states that there is an asymmetric relation between the levels. Macro-level properties facts or regularities depend on properties, facts or regularities at the lower level. This relation of dependency has been variously characterized as determination, constitution, supervenience, and superdupervenience (Horgan 1993). I will not go to these various concepts in detail. The essential thing here is that all these concepts try to characterize the fact that all macro-level causal powers are *inherited* from the micro level in some non-causal manner. The entities at the macro level have the causal powers they have in virtue of the micro-level entities that constitute them and the causal powers they have. The acceptance of this principle commits one to a monistic ontology, since it denies that entities at different levels are separate and wholly distinct. However, it does not say that macro-level entities or properties are reducible or identical to those at the micro level.

If we follow this hierarchical picture, we must at some point get to the bottom, that is, to the fundamental level. This idea is stated in the third thesis:

3) 'Microphysics' is the ultimate micro level.

This thesis claims that microphysics is causally complete, and that entities at all other levels derive their causal powers ultimately from microphysics. The precise definition of 'microphysics' may be difficult, but that is not the problem here. Let us just assume that there is an ultimate micro level, and call this level microphysics. We can also assume that it has some connection to things done in physics departments, but microphysics is still philosophical science fiction. We do not have a microphysical theory yet, and we may never have. Furthermore, it is an open question whether there can be both a plausible and a non-trivial philosophical definition of 'physicalism' (Crane and Mellor 1990; Pettit 1993b). But these issues are not my concern here. I am not claiming

that microphysicalism or physicalism is a philosophically interesting doctrine. I am only saying is that most of us accept the idea that there is a fundamental level. And furthermore, many of us also accept the idea that the facts at the fundamental level fix all the facts at all other levels. One way to put this is to claim that there is always a microphysical realization for every instance of a macro property. The macro entities do not have causal powers over and above those that they have inherited from their microphysical constitution. In this sense, macro properties are *causally impotent*.

With these background assumptions, we can proceed to the problems. Suppose that we are trying to causally account for event *e* and suppose, for the sake of argument, that we accept the Lewisian idea that to causally explain an event we have to describe its causal history (Lewis 1986: 217).[1] Let us assume that we have an acceptable causal explanation at level of description L_1 and let us further assume that this level is not the 'microphysical' level. The explanation traces the relevant causal history of *e* by using the vocabulary of L_1 and by making references to causally relevant properties, events and processes at that level.

Since we have accepted 1) and 2), we must assume that there exists a micro level L_2 with its characteristic causal properties, in virtue of which entities at level L_1 have their causal powers. Accordingly, we can describe the causal history of *e* also at level L_2. Furthermore, as we have accepted 3) we have to infer that both L_1 and L_2 (and all the other possible intervening levels) get their causal capacities from microphysics (L_p) and that we can describe the causal history of *e* at level L_p.[2]

These inferences create the problem of macro explanation. To put it simply, the problem is that we seem to have competing causal explanations for the same exact *explanandum*. Moreover, L_p seems to render explanations at levels L_1 and L_2 epiphenomenal. L_p involves everything that L_1 and L_2 tell us about the causal history, since it includes (among all other things) descriptions of physical realizers of all properties and entities at levels L_1 and L_2. With the story at level L_p we have already fixed all the causal facts, and there is nothing to add. It is the most complete story about the causal history of *e*.

We cannot consider L_1 and L_p as complementing causes of *e*. Once factors mentioned in L_p have done their work, we do not need entities from levels L_1 or L_2 to complete the process. However, this is not a case of causal overdetermination either. L_1 and L_p do not describe alternative causes. After all, we assumed that the levels are not totally independent of each other. Thirdly, L_1 and L_p are not temporally different parts of the same causal chain. On the contrary, they are descriptions of exactly the same parts of the very same causal chain. It seems that there is no way in which descriptions at levels L_1 or L_2 could stand as fully respectable and non-redundant causal explanations. (Jackson and Pettit 1990a: 108.)

The problem cannot be solved by saying that the properties at different levels are identical. An example given by Peter Menzies and David Lewis shows this. Consider a metal bar. It is characterized by the macro properties of electrical conductivity, ductility, distinctive luster and opacity, and all these properties are due to its microphysical nature. A cloud of free electrons permeates the metal and holds it in a solid state, or so we assume. This same micro state works as the realization (or causal basis) for all four macro properties (Menzies 1988: 567). This all seems very plausible. The problem is that if we identify opacity and the other macro states with this microphysical state, we get absurd results. If conductivity is identical with having a certain microphysical state, and if ductility is identical with this very same microphysical state, conductivity and ductility would be identical also. But this is false, since the two properties are clearly distinct. After all, we can observe these properties separately. There are things that are ductile but not conductive, and vice versa. The same problem occurs in explanatory contexts. The fact that a current can pass through a piece of copper is due to its conductivity, not to its opacity or its distinct luster. We want conductivity to be explanatory *qua* its conductivity, not its being identical with a certain microphysical state. The identity solution does not allow this.

When we contrast microphysical and macro-level explanations, only the former seem to be really explanatory. The impression is that L_1 and L_2 are only useful placeholders for explanations in terms of microphysics. Only microphysical explanations seem to be proper explanations, whereas the others have value only as substitutes. Explanations at the macro levels would always be replaceable in principle by microphysical accounts. In this situation, the special sciences would not have any 'real' causal explanations. They could be explanatory only in some derivative sense. Of course, it is not probable that we are will ever be in a position to give explanations in terms of microphysics, but this observation does not eliminate the philosophically worrisome point. The point is that our argument seems to make all macro explanations second-class citizens in the republic of explanations. This is *the problem of explanatory epiphenomenalism*. The trouble is to show how macro levels can be genuinely causally explanatory in the face of the 'borrowed' nature of their causal powers.

This is not just a hypothetical position. Several philosophers think that since macro-level explanations describe causal entities that are causally impotent (when contrasted with microphysical entities), they do not provide us anything but surrogates for real explanations in terms of microphysics. (cf. for example, Kim 1993, 1998; Rosenberg 1994.) Since the microphysical properties and entities do all the causal work, they should also get all the credit for it (in our explanations).

The conclusions of these arguments seem quite counter-intuitive. A great number of philosophers have proposed analyses that could save

our intuitions. We have, for example, models of quasation (Horgan 1989), second-order properties (Block 1990), supervenient causation (Kim 1993), program explanation (Jackson and Pettit 1990a), counterfactual theories of causal explanation (Ruben 1994, Baker 1995), and *qua*-causation (Tuomela 1998). The literature is enormous, and space does not allow me to explore it exhaustively. In the following I will pick up one intuitively plausible suggestion and show how my account of causal explanation can explain why that suggestion works (to the extent that it works at all). I have chosen the model of program explanation developed by Frank Jackson and Philip Pettit. I will argue that my account of causal explanation does away with the problem of explanatory epiphenomenalism and that my account is an improvement over the program model since I can explain why it works.

To make the abstract discussion about micro and macro properties more understandable, it is good to have an illustration. Usually the discussion proceeds by taking mental states as exemplary macro states. I will not follow this practice. We lack a proper scientific account of mental phenomena. Similarly we lack a scientific account of the states that realize mental states at the underlying level. Some of the problems in the discussion derive from the conceptual indeterminacy that follows from these facts. It is better to have an example that is more scientifically developed. One such generally accepted exemplar is the biological concept of fitness. It has two advantages as an example in comparison with mental states. First, it is a respectable scientific concept that has an irreducible role in biological theories. Second, there is no scientific disagreement about the physical basis of fitness. Everybody agrees that fitness derives its causal powers from below.

2. The causal powers of fitness[3]

We may follow Elliot Sober (1984) in interpreting the theory of evolution as a theory of evolutionary forces. In addition to natural selection, the forces in this theory include such things as random drift, migration, mutation, and breeding structure. In this view, the theory of natural selection is not the same thing as the theory of evolution, but its proper part. Evolution is understood as changes in distribution and variation of characters (or traits) in a population. This is nowadays the standard way to understand evolution, at least in population genetics. (Sober 1984: 31-59.)

Ideally, all forces of evolution are reflected in equations of population genetics, although in practice, biologists are forced to build models that are far less ambitious. In this account, population genetics works as a bookkeeping discipline: it draws together considerations from different sub-fields of evolutionary theory and shows how they combine to reflect the dynamics of the population.[4]

Fitness is represented in the models of population genetics as a *selection co-efficient* that shows the force of selection on a given character (-type) or organism (-type). Basically, it summarizes the effects of selection in a way that makes it possible to include selection as a component of an equation of population dynamics. This makes a mathematical treatment of population dynamics possible. Without the concept of fitness, the effects of natural selection could not be included in models of population genetics, and theorizing in evolutionary theory would be purely qualitative.

Fitness reflects the force of natural selection. Natural selection is a process that occurs when population has

a) variation among individuals in some attribute or trait (*variation*);

b) a consistent relationship between that trait and mating ability, fertilizing ability, fertility, fecundity and/or survivorship (*fitness differences*);

c) a consistent relationship, for that trait, between parents and their offspring, which is at least partially independent of common environmental effects (*inheritance*) (Endler 1984: 4).

Differences in the fitnesses of organisms[5] reflect differences in their capacities to produce procreation-capable offspring. In this way they inform about differences in adaptedness.[6] The difference between these concepts is that adaptedness is qualitative, where as fitness is a quantitative concept. The difference reflects their conceptual roles. Adaptedness is related to an analysis of interactions between an organism and its environment, whereas fitness is used to compare different organisms in the same environment, or similar types of organisms in different environments. There is one further difference: fitness is a concept of population genetics whereas adaptedness is used in ecology to reflect how well an organism fulfills certain "design" criteria (Beatty 1980). Apart from these differences, the concepts of adaptedness and fitness are really different sides of the same coin: adaptedness determines fitness, the latter being the former's quantitative measure. It is no wonder they are so often identified with each other.[7]

This discussion already gives the impression that fitness and adaptedness are dispositional concepts. Both refer to capacities or propensities of organisms in a given type of environment. Although these notions have been used in this way since their introduction, their explicit analysis as dispositions is only some 20 years old. This analysis is known as *the propensity interpretation of fitness*.

The propensity interpretation was proposed independently by Mills and Beatty (1994 - originally published in 1979) and Brandon (1978). It was developed mainly against various arguments to the effect that the principle of natural selection is a tautology or that it is untestable. These

problems were thought to be acute since many biologists defined fitness as actual reproductive success while, at the same time, fitness was still being treated as an explanatory factor. The problem is easy to see: if we define fitness operationally as actual reproductive success and at the same time explain reproductive success by fitness, we are going round in circles. This type of circularity is clearly fatal. A thing cannot explain itself. To be explanatory, the concept of fitness had to be distinguished from actual reproductive success.

There is a conceptual difference between fitness and actual reproductive success. This difference is easy to see when one considers Scriven's twins. Consider two both genetically and phenotypically identical twins standing together in a forest. As it happens, the one on the left is struck dead by lightning, and the other is spared. After the incident the one that stood on the right side continues its life and reproduces, while the other obviously leaves no offspring. Surely, there is no difference between these two organisms that could explain the difference in their reproductive success. Their fitnesses were equal. The lightning strike was an external random incident and it should not be reflected in the theoretical account of the reproductive capacities of twins. In explanatory uses of the notion of fitness, the differences in the organisms should explain the differences in their reproductive success, not some external random facts. However, if we define fitness operationally by actual reproductive success, we get fitness differences even in the cases where there is no difference among organisms. (Mills and Beatty 1994: 7.)

One cannot get out of this circle by defining fitness independently of reproductive success. The features that make organisms fit are simply too varied and context-bound to make this kind of definition possible. (Mills and Beatty 1994: 7; Brandon 1990: 13.)

One plausible solution is to treat fitness as a (probabilistic) dispositional concept. In this account fitness is an ability (or capacity) to survive and to produce a certain number of offspring. Scriven's twins are equal in their abilities to produce offspring, but these abilities were actualized only in the case of one of them. Since it is contingent whether the triggering conditions of the disposition are realized, actual reproductive success and ability to reproduce do not go hand in hand.

Dispositional analysis makes a clear conceptual difference between fitness and actual reproductive success. This does not make the concept of fitness inapplicable. If the conditions are right, one can (fallibly) measure fitness by counting actual offspring. However, there are some limitations. One can predict reproductive success on the basis of fitness considerations only if relevant background conditions are fulfilled. These background conditions include the absence (or constancy) of other evolutionary forces.

Dispositional analysis has some additional advantages. It opens up

other ways to measure fitness besides that of counting actual offspring, for example by starting from the causal basis of the disposition. By analyzing how the organism faces some relevant design criteria one can estimate its capacity to reproduce. As these features of dispositional analysis are in agreement with biological practice (fitness is being measured in both ways), the propensity interpretation has received rather wide acceptance.

Fitness is characterized as a propensity rather than an ordinary disposition because evolutionary explanations are statistical. There are several reasons for this. The first is that theory of evolution includes essentially statistical elements, such as random genetic drift. Secondly, the selection environment that organisms face often includes statistical elements. For example, there is a certain chance that a predator will be encountered, that an epidemic will sweep through the population, etc. Furthermore, there are random events that are not considered to be parts of the selection environment, but which still can affect the population. Recall the example of Scriven's twins. For all these reasons and without doubt a host of others also, evolutionary explanations are destined to be statistical.[8]

Thus, we cannot explain why a certain variant became dominant in a population. We can only explain why it had a better chance of becoming dominant. Being the fittest does not guarantee survival. In theoretical models, we can sometimes ignore these problems by assuming that the population is infinite, that there are no hazardous external events, etc. When explaining events in real populations, we cannot make these idealizations, and our explanations are guaranteed to be statistical. Thus, *we do not explain individual occurrences; we explain only probabilities of these occurrences*.

A central feature of the concept of fitness is that it measures capacities of organisms *within a certain environment*. This is reflected in the fact that the concept of fitness is environment-indexed. The environment belongs to relevant triggering conditions of fitness disposition. The same organism has different fitnesses in different environments. And two (non-identical) organisms that have the same level of fitness within a given environment very likely have differing fitnesses in some other environment.

As the notion of fitness is always used in a comparative context and is always environment-indexed, the theory of natural selection is *idealizing* in a quite interesting way. When constructing models of particular selection processes, the biologist must always set the competing variants within the same environment in order to make comparisons of fitness possible. Sometimes this homogeneous environment is achieved by randomizing some of the environmental factors. The models always include other idealizations as well, since actual environments are too complicated and heterogeneous to be included in the models.

(Sintonen and Kiikeri 1994.)

Fitness is a classical example of the use of the concept of supervenience in philosophy of science (Rosenberg 1978). The physical properties of an organism and of its environment *determine* its fitness. However, by knowing its fitness we cannot determine the organism's physical properties. This is because the same level of fitness can in principle be realized by an infinite number of different physical configurations. Thus, we can say that if two organisms are physically identical, their fitnesses must be identical (in the same environment), but we cannot make the reverse inference. Having identical fitnesses does not make two organisms physically identical. To put it simply: there is no difference in fitness without a difference in the physical properties of the organism. Because of these facts, the relationship between of fitness and its physical basis is that of supervenience. (Rosenberg 1978; 1985: 164-169; Sober 1993: 73-75.)[9]

This supervenience relationship rules out any meaningful relationship of reduction between those theories that employ the concept of fitness and those physiological, behavioral etc. theories that are used to *explain* levels of fitness. Although it is conceptually possible to collect an infinitely long disjunction of possible bases of a given level of fitness within a given environment (and to repeat this at all fitness levels), the operation would not have any cognitive significance. The relationship between the basis and the fitness is very complicated because fitness is a relational property, whereas its basis properties are not. This can be seen from the fact that the same basis properties support as many levels of fitness as there are types of environments within which the organism can be placed. These facts guarantee epistemic autonomy to evolutionary theories. There is no property or complex of properties that can be inserted into equations of population genetics to replace fitness. Thus, references to differences in fitness will remain an essential part of evolutionary explanations.

Ontologically, the situation is similar to that considered at the beginning of this chapter. The facts that determine that an organism has a given level of fitness include physiological, anatomical, and behavioral traits that underlie its viability and fertility. In fitness there is nothing over and above these facts; its causal powers are exhausted by the combined causal powers of its constituents. Fitness does not reflect any emergent causal powers. Thus, when we contrast fitness with its basis properties, we can see that it is causally inert or impotent. Its explanatory value, if it has any, is not based on its unique causal powers. (Sober 1984: 88-96.)

I have one more point to make concerning the explanatory use of fitness attributions. Some explanations by dispositional properties raise the problem of circularity. Let us call this the *virtus dormitiva* challenge. Consider the explanation of the breaking of a glass by its fragility. It seems

quite trivial to explain the actualization of a disposition by the very same disposition. The connection seems to be too close. There is no falsity in the explanation, but it is not very informative either. The same criticism can also be raised against fitness explanations. One seems to be explaining the given level of offspring production by the capacity to produce that much offspring. The propensity interpretation might have a conceptual difference between these two things, but it seems that the distance is not big enough to be explanatory. Does it mean that the references to fitness and dispositions are completely unexplanatory? If they are, that would unfortunately imply that most of our explanations are not explanations at all, as all causal properties (or at least causal powers) are dispositional by their nature. We need not draw this conclusion. Fitness really is an explanatory property, one only needs to put it into the right explanatory context.

First, explanations in terms of fitness are not simple *ad hoc* explanations. One can identify the level of fitness beforehand, either by design analysis or by comparative data. One does not infer the *explanans* from the statement of *explanandum*.

Second, the role of fitness in population genetics models is to work as a placeholder for the effects of natural selection. It is of the form: 'there is something that …'. Its contribution is irreducible because it cannot be replaced by the properties realizing it. Otherwise the explanation would lose some of its generality and its counterfactual force.[10] It is important to understand that in models of population genetics one is not explaining the contribution of selection, one is explaining the whole dynamics of the population. Natural selection is just one component in the process of evolution. There are always other causal contributions besides the force of selection. For example, since all actual populations are finite, the effects of random drift are always present. Thus, a trait's fitness and its initial distribution in a population do not alone determine its frequency in later generations. Because of this distance between the *explanans* and the *explanandum*, a reference to fitness is not explanatorily trivial or circular. The causal contribution of the natural selection is a contingent fact that cannot be derived from the *explanandum*. It is a theoretical hypothesis that is fallible and in practice always approximate.

This point can be generalized to other dispositional facts. Consider the case of a breaking glass. It is of course quite uninformative to explain the breaking of a glass by its fragility. One should not refer to dispositions to answer *explananda* that are too close. But, if the *explanandum* is the fact that the table got wet, the reference to the fragility of the glass might by a very valuable and informative part of the explanation. The fragility of the glass makes the difference between the table getting wet or remaining dry when the causal background includes water in the glass and a blow with a hammer.

3. Program explanation explained

With the help of the above exemplar of a macro-level property with a clear explanatory role, we can proceed to consider the program model of explanation. Jackson and Pettit developed this model of explanation in a series of articles (Jackson and Pettit 1988, 1990a, b, 1992a, b; Pettit 1992, 1993a, 1995b) and have applied it to various cases in both philosophy of mind and in philosophy of the social sciences. I will not discuss these applications. Instead, I will first go through the general account of their model and then apply it to the case of fitness.

Jackson and Pettit accept a position they call causal fundamentalism. According to this view, all higher-level causal powers derive from the causal powers of the entities at the fundamental level. Consequently, the fundamental level is the only level that has properties, states or entities that are *causally effective*, whatever this is taken to mean.[11] For them, causally effective properties are obviously causally relevant from the point of view of causal explanation. One way to explain an event causally is to describe its causal history in terms of causally effective properties. However, they also want to defend the claim that macro-level properties are *causally relevant* from the point of view of causal explanation as well. The aim of the model of program explanation is to do justice to the intuition of the causal relevance of macro properties, and its main challenge is to give a substantial account of causal relevance that is not based on causal effectiveness.

In their ontology, Jackson and Pettit regard macro-level properties and states as higher-order states. For example, the elasticity of an eraser is a second-order property in contrast to the molecular structure that constitutes it. To be in a higher-order state of being elastic, is to be in whatever lower-order state that fulfills (that is, causally realizes) the conditions of elastic behavior. In other words, to be elastic is to have such a molecular structure that makes the item behave elastically in appropriate conditions. Elasticity and other higher-level properties are characterized functionally. The functional identification of higher-order states allows for the possibility of their unlimited multiple realization. (Pettit 1993a: 28.)

With this basic ontology at hand we can proceed to the characterization of the model itself. The idea of program explanation is easiest to understand by considering cases in which the higher-order state is picked out by existential quantification, to be more precise, by the suitable use of the word 'some'. Consider a case where we explain the strange noise that a clock makes by noting that some of its components are loose. This explanation seems to be plausible and, at least in some contexts, satisfactory. We can explain the noise by this higher-order state without knowing which particular components are responsible for the noise. The explanation abstracts from details and does not specify which

parts of the clock are loose. The some-state the explanation mentions is a higher-order state in relation to the realizer state that such-and-such particular parts are loose. The some-state presupposes that there is an appropriate realizer state, but it does not specify which one it is from the set of possible alternatives. (Jackson and Pettit 1990a: 112; Pettit 1993a: 36.) Pettit puts it as follows:

> The some-state is causally relevant in a way that presupposes the causal relevance of the realizer state: if that realizer is not relevant then neither is the some-state. More specifically, the some-state is relevant so far as its realization more or less ensures the presence and effectiveness of such realizer state (Pettit 1993a: 36).

Another of Jackson and Pettit's favorite examples works similarly. A closed flask containing water is heated and it boils – the molecules in the water reach a certain level of mean motion - and at a certain point the flask cracks. Here, the temperature of water is the macro-level state, and the movement by the molecules is the respective micro-level state. If we wish to explain why the flask cracked, we have two different accounts. The macro account would explain the breaking by referring to the rise in temperature, and the micro account would explain by it describing the movements of individual molecules and their collisions with the molecular bonds in flask's surface. Clearly, the micro account is more basic: the movement of molecules constitutes the temperature. And yet we think that explanation in terms of temperature is useful and informative. In Jackson and Pettit's view this is to be explained by the fact that the rise in temperature *ensures* that there will be a molecular collision of a kind sufficient to produce the cracking. This is informative since there is almost an infinite number of possible molecular collisions that could cause the crack. The temperature "arranges things non-causally" (Jackson and Pettit 1992a: 118) so that there will almost certainly be a collision that will produce the breaking (Jackson and Pettit 1990a: 110, 1992a: 117-118, 1992b: 5-6).

Basically the same idea can also be applied to other kinds of higher-order states. The fragility of a glass 'ensures' that the glass has one of the possible molecular structures that make it behave according to the regularities characterizing fragile objects. To further illustrate the idea, Jackson and Pettit adopt a metaphor from the world of computing. In their parlance, the higher-order state *programs* for a certain type of effect by non-causally ensuring the presence of a suitable lower-order state that produces the effect. All actual causal production happens at the lower level, but the same 'program' can be realized by different micro configurations. The only thing all these micro configurations need to share is that they realize the same program. So the realizing state can be any one of these possible micro configurations.[12]

According to Pettit a higher-order property programs for certain result *E* when:

1. Any instantiation of the higher-order property non-causally involves the instantiation of certain properties – maybe these, maybe those – at a lower level.
2. The lower-order properties associated with instantiations of the higher-order properties, or at least most of them, are such as generally to produce an *E*-type event in the given circumstances.
3. The lower-order properties associated with the actual instantiations of higher-order properties do in fact produce *E* (Pettit 1993a: 37).

Jackson and Pettit call explanations in terms of higher-order properties *program explanations* and explanations in terms of lower-order properties *process explanations*. According to them, the properties mentioned in process explanations are causally relevant because they are *causally effective* properties. Properties at the higher level can also be causally relevant, although they are not causally efficacious. They are relevant because they are *causally programming* properties. Both kinds of properties can explain, but they offer different kinds of information. The process explanation refers to causally efficacious properties and it offers *contrastive information*. It tells us what kind of causal processes work in the actual world. This information helps us to distinguish actual processes from processes in alternative possible worlds. The program explanation offers *comparative information*. It tells how the actual world and some other possible worlds are similar apart from their differences at the lower level. It tells us that under certain conditions the same effect would have taken place no matter how the world in question differs from the actual world. The presence of a programming property ensures that there actually is a causal chain that produces the causal result. (Jackson and Pettit 1992a: 119, 1992b: 15: Pettit 1993a: 232.)

The program explanation tells us what remains the same when other factors change. It identifies a stable macro-level pattern. Recall the example of a cracking flask. There is a stable pattern or regularity between a rise in temperature and cracking. This pattern is invisible at the micro level. The micro-level account tells us everything that happens as the flask breaks, but the program explanation adds to this information about other conditions in which the same effect would have been produced. In this sense it is an improvement over the process explanation. The program explanation seems to provide some information that is not available at the process level.

For Jackson and Pettit, program explanations are valuable because they offer information about the causal history of the *explanandum* that

we would not otherwise have. This leads them to explanatory ecumenism: explanations at various levels can be irreducibly informative, and these different explanations complement each other (Jackson and Pettit 1992b: 16). One consequence of ecumenism is that the respectability of special science explanations seems to be restored. Independently of the future successes of 'microphysics', the program explanations in the special sciences will preserve their value.

Fitness seems to fit this model nicely. As noted, it is in the final analysis a causally impotent property, but it is still used in causal explanations. It also seems to provide the comparative information that Jackson and Pettit take to be characteristic of program explanations. It tells us what is common between different organisms and it presupposes that there really are causally powerful properties that will do the actual causal work. By abstracting from micro-level details, fitness tells us what various types of organisms have in common. In addition, by leaving out idiosyncratic details, it can make explanation more understandable and theoretically interesting. Jackson and Pettit's account also seems to do justice to the intuition that explanations making references to fitness are irreducibly valuable. Thus, it seems that evolutionary explanations that make use of the concept of fitness are program explanations. This sounds like a nice confirmation of Jackson and Pettit's account: it saves our intuitions about fitness explanations.

Should we be satisfied with the account provided by Jackson and Pettit? I think we should not. Although the model seems to do justice to some of our intuitions about explanation and it makes the right properties explanatory, it does not tell us why this happens. Why is the information that the program (or process) explanation offers about the causal history of the *explanandum* event explanatory? Neither does the account help us account for the *explananda* of these explanation. For example, what do evolutionary explanations explain, or what are they able to explain? The only answer Jackson and Pettit are able to provide is: there are some explanatory questions for which the provided information is explanatory. This is hardly a satisfactory answer.

The basic problem is that the model simply claims that macro properties are explanatory and then accounts for this by using metaphors like 'programming' or 'non-causally ensures'. Instead of providing a model of causal explanation, Jackson and Pettit give us a couple of metaphors that do not get us very far. The reason for this failure is quite clear. If one wants to say something substantial about explanation, one should have a substantial account of explanation, but this is something Jackson and Pettit do not have. They want to make some general points about explanation without committing themselves to any particular account of explanation or causation. This strategy may seem tempting when one considers the wide array of philosophical theories of explanation available, but the temptation is misleading. One cannot have

philosophical results without philosophical assumptions. At most, one can provide examples of plausible explanations, as Jackson and Pettit have successfully done.

Let us take one more example of process and program explanations. I smell a cigar and I raise the question: Why do I smell a cigar? The program explanation is quite simple:

> Someone lighted a cigar in the same room.

The respective process explanation is much more complicated:

> A person x lighted a cigar y at moment t_1, and the molecules m_1, m_2, …, m_n, traveled routes r_1, r_2, …, r_n with speeds s_1, s_2, …, s_n, reached my nose at t_n, and interacted with my senses in the following way ….

Both explanations seem to be acceptable causal explanations. But by using the heuristics developed in Chapter 2, we can see that these explanations have different *explananda*. One plausible *explanandum* for the program explanation is:

> I smelled a burning cigar
>
> [I did not smell a burning cigar].

The exact *explanandum* for the process explanation is very hard to characterize, because it requires a special vocabulary and concepts that are not accessible in the vernacular. However, it is something like the following:

> I smelled a burning cigar exactly the way I smelled it
>
> [I smelled a burning cigar in some other way].

For the program explanation it does not matter how the smell reached my nose. We can change a great number of facts about molecules, their velocities, and their routes and still give the same program explanation. Changes in these facts do not make any difference from the point of view of the intended *explanandum*. This is why the process account is not the right explanation for this *explanandum*. With the *explanandum* of the process explanation, things are different. The resolution of its *explanandum* is now much higher. Only a small change in facts that were irrelevant from the point of view of program explanation, and the explanation would not remain true. The outcome would be different. For this *explanandum* the program story would be too general to be acceptable.[13]

These results seem quite intuitive, and they point to an underlying account of explanation. The example suggests that both explanations work by providing counterfactual information about the causal history of the event to be explained. They seem to be explanations of the same kind, but their *explananda* are different. This conflicts with the conclusion drawn by Jackson and Pettit. Their discussion gives an impression

that these are different kinds of explanations of the same *explanandum*. When we explicate the *explanandum* with the contrastive-counterfactual model, we find out that they are explanations of slightly different *explananda*. Micro and macro explanations have different foils and thus they provide different explanatory information. The impression that micro and macro explanations are in competition is a result of the *explanandum* not being explicit enough. The example also illustrates nicely how the resolution required from the explanation depends on the resolution of the *explanandum*. If the difference to be explained is sophisticated, then also the explanation has to be sophisticated.

This analysis of the situation has some conceptual advantages. We do not have to explain differences between program and process explanations, since there are none. In contrast, Jackson and Pettit owe us further explication of these explanations and their differences. Furthermore, we can drop the strange talk of 'causally effective properties'. The approach also avoids the conceptual problem that Jackson and Pettit face. Recall that they accept the multi-layered ontology. Let us suppose that the social level is the macro level and the psychological level is the micro level. In this case, social structural explanations would be program explanations, whereas explanations at the psychological level would be process explanations. Now consider the situation where we add the neurophysiological level to our considerations. Obviously, explanations at this level would be process explanations in comparison with explanations at the other two levels. But what happens to process explanations at the psychological level? It seems that they would have to be both program and process explanations at the same time. This might be otherwise acceptable, but the assumption was that process and program explanations are different kinds of explanations. How one thing can be both similar and different with regards to the same characteristic at the same time?

Let us now get back to fitness, but without the metaphors employed by Jackson and Pettit. Is there something that we miss without them, and/or is there something we do not get if we stick to them? Let us ask some basic questions. What is the explanatory role of fitness attributions? What is the *explanandum* in selection explanation? Evolutionary explanations make use of population genetics. Models in population genetics provide biologists with a general scheme for answering 'what if things had been different?' questions. The differences in contrasted end-states are accounted for by differences in the variables and parameters of the model. The point of these explanations is to explain differences in success between competing variants. The *explananda* of natural selection explanations are explicitly *contrastive*. They explain why a given trait (or organism) was selected *rather than* its alternatives.[14]

This comparative context is reflected in the fact that a selection coefficient is usually presented as *relative fitness* and not as absolute

reproductive capacity. The genotype with the highest level of fitness is usually given value 1, and then less fit genotypes get multipliers (< 1) that reflect their reproductive capacities. The alternative genotypes considered in this comparison are the variants present to the process of selection. Evolutionary explanation starts from the situation where we have more than one variant present in the population and then goes on to account for the fact that the population reaches equilibrium. *Usually* the equilibrium reached is one in which there is only one variant present in the population (in significant numbers). The differences in the fitnesses of traits explain the differences in their evolutionary success. The evolutionary explanation focusses on *population*. It explains changes, or the absence of changes, in the population. It does not explain the success or the properties of individual organisms. Although one can explain the adapted traits of an individual by referring to natural selection, this explanation will always include essential references to the population dynamics within the prehistory of that individual.

These observations fit the model of causal explanation developed above rather nicely. The *explanandum* is clearly contrastive. We want to explain why variant *a* rather than variant *b* became dominant in the population. We also have an explanatory mechanism that has an essential role in explanation: the process of natural selection. With this account of explanation Jackson and Pettit's metaphors are not needed. Furthermore, there are important things that my account helps to illuminate, but which remain in the dark with the program model. The program model does not emphasize the contrastive nature of evolutionary explanations. It also misses the central role of the explanatory mechanism. Without these ideas, making sense of evolutionary explanation is quite difficult.

The contrastive-counterfactual model seems to provide a very natural way to account for intuitions behind the program model. With it we can state all the interesting things found by Jackson and Pettit. Furthermore, it allows us to explain why the program and process explanations explain. We do not have to take their explanatory status as given. My account shows that both process and program explanations are examples of the same pattern of explanation.

There is one more advantage for my account of explanation. Jackson and Pettit (1992b; Pettit 1993a: 248-265) defend so-called explanatory ecumenism. They argue for this position by showing that what they call a *fine grain preference* is misguided. The fine grain preference can be described as a desire to have the greatest possible causal detail in explanation. According to this view, getting a more detailed explanation is an end in itself. The finer the grain of an explanation, the better the explanation is. The fine grain preference can work in two dimensions, and we can distinguish between close grain and small grain preferences. *The small grain preference* shows in the preference for micro-

level explanations. For example, in our example of the cracking flask, the molecular account would be better according to the supporter of the small grain preference. *The close grain preference* shows itself in historical explanations. According to it, it is always preferable to have a closer distance between causes and effects. An explanation that refers to a temporally close causal antecedent is always preferable to one with more distant antecedent.

The basic argument against fine grain preference can be seen by considering Richard Miller's (1978: 410-411) explanation for the change from carbon steel to stainless steel as the main material for knives. According to Miller, the change can be explained by the greater capacity of stainless steel to keep its edge and by reduction in the relative cost of stainless steel in the 1920's. To explain this change we do not need to give detailed historical description of the episodes in which individual cutlery executives made actual decisions to switch production to stainless steel. We can abstract from these details, since if the change had not happened this way, it would have happened some other way. If these executives had not made the decisions, some others would have. Why can we make this abstraction? The answer lies in the competive nature of the market for knives. Since the users preferred knives made of stainless steel, there was selection for firms that produced them. In the long run, only those firms that switched their production to stainless steel were able to stay in the business of knife manufacturing. If the social scientist is interested in explaining the ultimate victory of stainless steel rather than the exact timing of its adoption, she can abstract from the details of individual executive decisions. This abstraction does not worsen the explanation, it makes it more robust.

Jackson and Pettit argue – convincingly to my mind – that neither version of the fine grain preference is supported by good arguments, and that they should be discarded as general methodological rules. I agree with this conclusion. The problem is with the role of the model of program explanation in their argumentation. It is used to argue against the small grain preference, but since it does not apply to the temporal case, the case against close grain preference is made by analogy. Basically, Jackson and Pettit raise exactly the same argument against both preferences: if we follow these preferences, we lose explanatorily valuable comparative information (Jackson and Pettit 1992b: 15-17).

The trouble is that the program explanation model can explain why the small grain preference is false, but it cannot do it for the close grain preference. The model speaks about the relation between higher-order properties and lower-order realizations. It does not apply to the relation between earlier and later parts of an causal chain. However, Pettit and Jackson employ similar considerations to rule out both preferences. Why is this possible? The natural conclusion is that the program model does not make deep enough claims about explanation. There has to be

some more general account that could explain why both preferences are misguided. If we can use the same account to explain why program explanations work and to justify why the preference for small grain is misguided, we have a clear advantage over the program model.

The contrastive-counterfactual account of causal explanation is not only the account to do this. It can clearly do the job; it can also be used to explain what is plausible in fine grain preference. It fails as a methodological rule because it requires explanatorily irrelevant information to be included in the explanation. The valid core for the fine grain preference comes from the demand for explanations with mechanisms. The search for mechanisms leads us to the small and close grain details. But the point is that the explanation does not improve just by any detail, as the fine grain preference would have it. We want details about the mechanism. This leaves most of the fine details uninteresting and irrelevant from the point of view of explanation.

4. Intentional explanation as causal explanation

After having discussed the problem of macro explanation at a general level, we can return to the case of intentional explanation. The application of my account of explanation to the intentional explanation is quite straightforward. I will now briefly characterize my view on intentional explanation since the discussion in Part II of this work will build on it.

Common sense regards beliefs and desires as causes of behavior. We explain actions by reference to the aims and desires that motivate it and by beliefs that underlie it. For example, we might say that "She quit smoking *because* she believed it was affecting her health". The 'because' here seems to work the same way than in non-intentional causal explanations. If she have had different beliefs, or desires, she would have behaved differently. Beliefs is assumed to make a causal contribution to the process. This can be seen, for example, by manipulating the beliefs of somebody. Our intervention, if it is successful, brings about changes in the agent's behavior. The same applies for aims and desires.

Further evidence is provided by the everyday distinction between rationalization and explanation. This distinction only makes sense against the background assumption that there is a difference between the motives that causally affect the behavior and the motives that are falsely afterwards said to have affected the behavior. For common sense, there is a clear difference between the motives that explain the behavior and the motives that could have explained the behavior.

The contrastive counterfactual model of explanation is in accordance with these intuitions. In it, the contents of beliefs, desires and other intentional attitudes are explanatory when they make a relevant difference in agent's behavior. My account of explanation can be seen as an elaboration of accounts first developed by Schiffer (1991), Ruben

(1994) and Baker (1995), whose counterfactual theories of causal explanation are developed in the context of intentional explanation.

Let us see how the contrastive counterfactual deals with intentional explanation. The *explanandum* of intentional explanation is an agent's action or choice. This suggests that natural foils are alternative actions or choices that are incompatible with the *explanandum*. When we want to know why Jill went to Florida for a holiday in November 2000, natural alternatives are, depending on the context, other holiday destinations and staying at home. The explanation for going to Florida rather than to Uganda might be that the latter never came to Jill's mind as a holiday resort. In contrast, the explanation for going to Florida rather than Jamaica might be that Jill found the unique opportunity to observe the legal and political wrangling at the presidential elections on location more appealing than Jamaican attractions that can be seen any time.

What about my mechanism requirement, is it satisfied in the case of intentional explanation? The example suggests that the central explanatory mechanism in intentional explanation is practical reasoning by the agent. An intentional explanation shows how the agent's beliefs, desires and intentions led her to the choice she made via a process of deliberation. The idea is that the differences in the inputs for the process of practical reasoning make the difference in the outputs, the agent's actions. Had the agent had different beliefs or desires, she would have behaved differently. The explanation gives practical reasoning a causal role: its presence and its contents made a difference in the process that resulted in the agent's behavior. Beliefs, desires and goals get their explanatory relevance by their role in the agent's practical reasoning.

In Chapter 2, I identified the explanatory mechanism functionally: it is something that ensures that *f* occurs instead of *c* because of *a*. An agent's practical reasoning fits this bill. It tells us *how* the beliefs and desires led the agent to act in a certain way rather than some other way.

Practical syllogism is a traditional way of representing practical reasoning. I do not regard practical syllogism to be the fundamental explanation model for the human sciences that von Wright (1971: 29) takes it to be. However, I find it useful in bringing out some important characteristics of practical reasoning. Consider the following extended version of practical syllogism:

1) *A* intends to bring about *p*

2) *A* considers that she cannot bring about *p* unless she does *w* at time *t*

   i) *A* believes that she can do *w*

   ii) *A* does not have other goals or desires that overrun or compete with intending *p* or doing *w*

iii) *A* believes that she can do other things *x*, *y*, *z* etc. that bringing about *p* requires

iv) Either a) A does not believe that there are alternatives to *w* that *A* could do to bring about *p*, or b) she considers *w* to be preferable to these alternatives.

v) *A* is capable of completing the inference from her premises to the conclusion

vi) Nothing interrupts *A*'s inference process

vii) *A* does not change her mind during the process of deliberation or before the time *t*

viii) *A* does not forget her intention during the time between deliberation and time *t*

C) *A* sets himself to do *w* at time *t*.[15]

1) and 2) are the major premises of practical reasoning, and C is the conclusion. This syllogism is not a logically valid deduction, and I do not have any interest in transforming it to one. The point I want to make concerns the auxiliary assumptions i) - viii). They emphasize the comprehensive character of practical deliberation, which includes much more than the intentions and beliefs mentioned in premises 1) and 2).

A great number of other beliefs and desires have relevance to the practical reasoning, as assumptions i) - iv) bring out. Practical reasoning is holistic by its nature: the agent's other beliefs and desires constitute the context in which practical reasoning takes place. If this context is different, the conclusions of the inference can also be different. This creates some problems for the construction of intentional explanations. Since the process of attribution beliefs and desires to an agent is also holistic, the reconstruction of an agent's intentional states is forced into a hermeneutical circle. This has some methodological consequences for the human sciences, but these consequences are not my concern here.

In contrast, assumptions v) - viii) bring out points that are directly relevant to my argument to the effect that practical reasoning should be understood as a causal process. Let us start with assumption v). It brings out an important competence assumption: only behavior by agents with relevant inferential capacities can be explained intentionally. Why do we make this competence assumption? The only way to make sense out of it is to think that practical reasoning is a causal process. Only causal processes have causal preconditions. If it were not a causal process, our practice of attributing practical reasoning only to beings that satisfy the competence assumption would be completely arbitrary.

Assumptions vi) - viii) provide additional support for my thesis.

They show that practical reasoning is assumed to be a process that takes time and is open to a causal interference. We can give an intentional explanation only when the agent's reasoning is not subject to such interference. It is an essential feature of causal processes that they take time. Secondly, only processes that are themselves causal can be subject to causal interference. I conclude that practical reasoning is a causal process.

It is not my aim to argue that practical reasoning is an ideal example of an explanatory mechanism. I only want to claim that it is an explanatory mechanism. Clearly, there are several points where our account of this mechanism could be improved. First, the competence assumptions presupposed by the attribution of episodes of practical reasoning should be clarified. It is probable that in our everyday explanatory practice we attribute practical reasoning too liberally. Second, we have quite a limited understanding of the patterns of inference used in practical reasoning. Finally, it would be illuminating to understand how intentional states and our cognitive life are realized at the neurophysiological level. This understanding would lead to a better understanding of the limits and the possibilities of intentional psychology, and it would also connect intentional psychology to the explanatory mechanisms of natural sciences. The process of acquiring such an understanding will probably bring about some changes in intentional psychology and in our explanatory practice relying it. However, the essential elements of it will remain.[16] Despite these shortcomings of intentional psychology, practical reasoning is a causal mechanism.

Note that if we understand causal explanation as I have suggested, the usual anti-causalist arguments against interpreting intentional explanations as causal explanations are disarmed.[17] First, my account does not require that there are psychological laws covering intentional explanations. The existence of such laws can be left completely open since singular causal explanation does not include essential reference to laws.

Secondly, as the earlier discussion in this chapter shows, my account provides a solution to the problem of explanatory epiphenomenalism. There is no reason to fear that physical causes will deprive intentional states of their explanatory relevance. The causal powers of mental states can be inherited from the microphysical level, as the physicalist claims, but the explanatory relevance of the former is retained, since there are some contrastive questions for which they offer more adequate explanations.

Thirdly, my account does not create any special problems for answering the so-called logical connection argument (cf. von Wright 1971: 91-96). The usual arguments against it are still in force. The fact that there is sometimes a conceptual link between intentions and actions does not show that there cannot be a causal process that relates the agent's beliefs and desires to her behavior. And it is the latter that inter-

ests social scientist. When we categorize a certain piece of behavior as an action, we assume that it was brought about in a certain manner, that is intentionally. This categorization creates the conceptual link. In such cases, the relation between an intention and the respective action is too close to be explanatory. However, this does not show that intentional explanation is impossible. To the contrary, it seems that the existence of the conceptual connection *presupposes* the possibility of intentional explanation. After all, it is inbuilt in the concept of action that it is a piece of behavior that is brought about intentionally.

Finally, what about those who say that the aim of human sciences is to understand human action instead of explaining it causally? The reply is that they are not two different things: to explain intentionally is to understand. In order to causally explain human action, we have to interpret it. The construction of interpretation requires looking at things from other person's perspective and it also requires some empathy. These things, properly understood, are not obstacles to causal explanation. Rather, they are preconditions for finding one. The perspective of causal explanation has an answer to the question why it is important to study cultural meanings and how people interpret them. The answer is that these things are important because they make a difference in the way people behave.

Notes to Chapter 4

1 One can encounter similar problems, for example, when using a D-N account of explanation, see Kim 1993: Chapter 13.

2 This argument does not make any special assumptions about the relationships between vocabularies of different levels. There may, or may not, be bridge principles that connect the predicates between the levels. We are not translating descriptions of one level to descriptions of another level. Rather, our inference is based on assumption 2), which says that the lower level facts fix the higher level facts. This leads us to the conclusion that the ultimate truth has to be at the bottom level.

Furthermore, the argument does not say that there must be an explanation at level L_p, because the *explanans* at that level explains the *explanans* at level L_1. That would presuppose transitivity of explanation, which is a controversial assumption. Rather, the argument says that assumption 2) implies that there is a description of the causal history at level L_p. Of course, the argument does not deny that level L_p can provide instantiation explanations of the properties and states at level L_1.

3 The following draws from Ylikoski (1997), which includes a more detailed discussion.

4 In this account, population genetics is not to be equated with any specific models or laws (like Hardy-Weinberg theorems). In different evolutionary contexts these principles are quite different, and furthermore these principles themselves are also molded by evolution.

5 Fitness can be attributed to various kinds of entities: genes, genotypes, phenotypes etc. There is also a further question concerning whether fitness can be attributed to tokens of these types. For simplicity I will only talk about fitness of organism types.

6 The notion of adaptedness must be distinguished from two other notions of adaptation. The first refers to the process of (directional) natural selection and the second to the products of this process. Adaptedness refers to capacities of a given biological unit, not to its origin or to the process that created it. Thus, adaptedness is an ahistorical notion in distinction from these two notions of adaptation. One can further distinguish various special notions of adaptedness. For these notions, see Burian 1984.

7 Sometimes a distinction is made between Darwinian fitness and Fisherian fitness (Ettinger *et al*. 1990: 501-504). The former refers to what I have called adaptedness and the latter to fitness proper. It does not help in this conceptual confusion that Fisherian fitness is sometimes called Darwinian fitness, or that the actual number of offspring is also called Darwinian fitness (Burian 1984: 299). As Darwin lacked a way to measure fitness, his notion of fitness remained purely qualitative. He used the notion of fitness for the first time in the fourth edition of *Origin of Species*. It was not until the 1930's that R. A. Fisher introduced the modern concept of fitness. Concerns about the tautological nature of fitness were raised only after this change in the use of the notion (Settle 1993: 64).

8 For further discussion of the statistical nature of evolutionary theory see, Beatty 1984, Beatty and Finsen 1989, Richardson and Burian 1992, Horan 1994, Rosenberg 1994, and Brandon and Carson 1996.

9 For later, and far more sophisticated accounts of supervenience, see Kim 1993 and Horgan 1993. I have discussed supervenience in Ylikoski 1997.

10 This is not to deny that the explanation would be better if the details about realizing states or properties were also included. But in this case we would have two explanations: a natural selection explanation and an explanation of a given level of fitness. It is better to have both, as long as one remembers that they are separate things.

11 Jackson and Pettit do not want to commit themselves to any particular theory of causation (Pettit 1993a: 32-33). Similarly, they do not commit themselves to any specific theory of causal explanation, except to the idea that to explain an event is to provide information about its causal history (Jackson and Pettit 1992a: 119).

12 Consider a real computer analogy. There are application programs that are similar in both Windows and Macintosh operating systems. From the point of view of the user, these programs are exactly the same and they work similarly. However, if one considers the operating system, processor, and the physical configuration of the computer, quite different processes realize the same operations.

13 Elliot Sober (1999: 551) argues that the choice between micro and macro explanations is a matter of taste. According to him, there is no objective reason to prefer one to another. From the point of view of my analysis, Sober gives up too easily. If we look at the *explanandum* more carefully, we can observe some differences between macro and micro explanations.

14 Although the notion of fitness can be used to compare a) different organisms in the same environment and b) similar organisms in different envi-

ronments, only the former use is explanatory. In the latter, one is comparing two different fitnesses. As these fitnesses are measured according to the same standard (reproductive success) one can find out to which environment the organism is better adapted. This comparison does not make explanatory use of the notion of fitness.

15 This reconstruction is inspired mainly by discussions by von Wright (1971: 83-131) and Tuomela (1977: 170-205).

16 There are those who think that the future development of neuroscience will force us to give up the intentional psychology completely (cf. Churchland 1991). These eliminative materialists fail in their argument by misunderstanding intentional psychology, or folk psychology as they call it. Intentional psychology is not a theory about the internal workings of the brain (cf. Graham and Horgan 1988; Jackson and Pettit 1990c; Blackburn 1991; Horgan and Woodward 1991; Baker 1995; Egan 1995). However, even if the eliminative materialist were right, my thesis that practical reasoning should be understood as a causal mechanism would still hold. Eliminative materialists agree that it is an explanatory mechanism, they just hold that is a mechanism that should be given up because it is not compatible with the rest of science.

17 The anti-causalist camp has suffered a number of significant desertions since the 1960's. For example, Peter Winch has turned to causalism (Winch 1990: xi-xii).

Part II

Explicating interest explanations

Chapter 5
Interests and science studies

In the following three chapters I will take a look at one of the central explanatory resources of the social studies of science. Historians, philosophers and sociologists of science use interest explanations, although sociologists use them most visibly. Interests belong also to the explanatory repertoire of scientists themselves. Their widespread use does not mean that there is a consensus concerning the proper use of interests in explanation. The users of interest explanations differ from their critics in their respective understandings of both the nature and the scope of these explanations, and opinions are not unified even within these two camps.

When an explanatory concept is as widely used as 'interest', one might expect that some methodological discussion about its nature underlies its use. Such is not the case, for methodological discussion of the nature of interest explanations is nearly nonexistent. Of course, if the concept were clear and if everybody understood what it entails, there would be no need for a methodological discussion. But even a brief look at the published discussion shows that self-evidence or clarity of interest explanations cannot be the reason for the absence of methodological reflection. The advocates of the interest perspective have not been very informative about one of their central theoretical concepts, and as consequence their critics seem to have great problems hitting the target.[1]

The aim of the following chapters is to improve this situation. My goal is not to correct particular misunderstandings, but to clarify the *explanatory models* behind interest explanations. I approach the task both by explicating the explanatory apparatus and articulating the *explananda* of these explanations. This chapter is divided into three parts. The first part will make a conceptual analysis of the notion of interest and the second will take a look at discussion about interests in the social studies of science. The intended *explananda* of interest ex-

planations are discussed in the third part. The detailed analyses of interest explanations is left to Chapter 6, where I will discuss examples of interest explanations and sketch a taxonomy of these explanatory models.

This chapter does not intend to present a historical account of the influences or the details of various applications of interest concepts in the social studies of science. Such historical work is undoubtedly important, but my interest lies in conceptual analysis. Nor will I deal with the uses of interest concepts outside the sociology of scientific knowledge. There are many interesting things to say about interests in everyday talk, political theory and jurisprudence, but these issues remain outside the scope of this work.[2]

Before starting a more systematic discussion, it is good to reflect on the relevance of interests as a topic of philosophical analysis. This is especially important, since a look at recent literature on the sociology of scientific knowledge seems to suggest that interest explanation is no longer a very central issue. I would claim that this impression is wrong. It is true that many practitioners of sociology of scientific knowledge claim that they are not interest theorists and that their explanatory resources would rather include things like forms of life, practices, norms, paradigms, tacit knowledge, technological trajectory, social (and material) construction etc. However, this is only part of the story. This can be seen by considering how these notions are used in explanations. They are all descriptive concepts that leave open the question *why* things are the way they are rather than some other way. In trying to account for the changes (or the stability) of all these fancy things, one has to refer to things like goals and interests. The discussion of interest explanations might be out of fashion but not out of relevance.

1. Interest as an extension to folk psychology

The concept of interest is part of the everyday vocabulary that we use when trying to make sense of each other's behavior. It is part of mundane reasoning (Pollner 1974) or folk psychology, as philosophers call it. In the following I will explicate the everyday concept of interest in order to bring out the specific character of the sociological concept of interest.

Applying the concept of interest to a particular agent requires making certain assumptions. The first assumption is that the agent in question has at least some aims, goals or desires. I call this the *intentionality* assumption. There are not very strict limits on what the agent can want, but he or she must want something. The second assumption is about *rationality*. We assume the agent is rational in the sense that she does things that she believes to be conducive to her goals or aims. We also presume that if there is more than one way to advance agent's goals,

the agent chooses the one that she believes to be the most conducive to her goals. The third assumption is about *knowledge*. We assume that the agent is able to collect reliable information about her action environment either by her own observations or by using other social agents as the sources of information. In addition to this assumption about capability to acquire knowledge, we assume that the agent has also used these abilities to the effect that she has acquired beliefs about her action environment. We do not have to assume that the agent can take note of everything or that she knows everything about her action environment. These assumptions about knowledge give some space for mistakes and omissions, but do not allow that the agent can be completely wrong about her action environment. Naturally, it is possible that the agent does not fulfill our requirements of intentionality, rationality, and knowledge, but in such cases we will not try to use interest concepts to make sense of the behavior of that agent. In such situations one has to use some other scheme of analysis.

With the help of these assumptions we can make many useful inferences about each other. When we know the goals of an agent, we can predict what she will do in a given situation. If we know what the agent does, we can make informed guesses about her motivation and goals. We can also sometimes infer interesting information about the action situation from the knowledge of the agent's goals and her choices of action in that situation. With the help of these assumptions we can make the agent's behavior understandable and we can explain why she acts in a certain way rather than other. To understand this reasoning better, let's have a closer look at the concepts of goal and interest.

Interests and goals

In ordinary language the notion of interest is used in various ways. A look at a large dictionary confirms this observation. It can, for example, be used to refer to a share in a business venture, to a legal concern or right in property, to money paid for the use of lent money, and to a hobby. These uses are peripheral from our point of view. Interest can also refer to things or topics that draw one's curiosity, concern, or attention. The attitude can be expressed by phrases like 'being interested in' and 'being interesting'. One can also make somebody interested in some issue or item, which means that one succeeds in attracting that person's curiosity or succeeds in enlisting a person in a given cause. Some writers in science studies literature have drawn attention to this kind of interest talk among scientists (Callon & Law 1982; Latour 1987). Studying these processes is an important topic of research, but it is not among the concerns of this chapter.

For our current concerns the most important use of the notion 'interest' is related to things that are conducive to one's goals or aims.

Things are in one's interest, or one's concerns, because of their relation to the achievement of one's goals. These things are not 'interesting' by themselves, they are the *means* to achieve one's goals or to fulfill one's wishes. This connection ties interest closely to the cluster of goal concepts at the heart of intentional psychology. Among these concepts (in English) are such notions as: aim, desire, end, goal, objective, purpose, and want. There are important differences between these notions, but they behave quite similarly in relation to the notion of interest. As my goal is only to explicate the concept of interest, I will not complicate my discussion by separating these other concepts from each other.

Things that are in one's interest depend on one's goals. Interests tie goals to the action environment by picking up those things that are necessary for the realization of these goals or are otherwise conducive to them. If one does not and will not have any desires or aims, then one does not have interests either. Interests are *goal-dependent*: there are no interests without goals or aims. There is one sense in which one could have interests without *currently* having any goals and desires. In this case the interests refer to things that are conducive to the goals that the agent will or can have in the future. For example, we can speak about the interests of a child without attributing any current goals to that child. These future interests are not very helpful in accounting for agent's current behavior, so I will leave this very special sense of interest outside my discussion.

The goal dependence of interests makes them goals themselves. If the achievement of *B* is the (only) means to achieve my goal *A*, and I know this, then the achievement of *B* becomes my goal also. I will call a goal formed by this kind of practical reasoning a *subgoal*. Subgoals derive from having some other goal. They are pursued only because of their connection to some more fundamental goal. Being a person or a human is not a prerequisite for having interests. All kinds of agents can have interests, as long they are agents with goals.

The next important conceptual difference between goals and interests concerns knowledge. Although it might be sometimes difficult to say what one exactly wants or what one is aiming at, it is clear that there is a sense in which one has to know what one wants or aims for. It is not *my* goal if I do not know that I have it. One can for a short period of time forget what one was trying to achieve, but if this forgetfulness endures too long, one simply does not have that goal anymore. The same goes for beliefs about one's goals. One cannot have completely wrong beliefs about one's aims and desires. What I believe to be my goals, are my goals.

Interests are different. It is possible that the agent is completely ignorant of the things that are in her interest, but those things are nevertheless in her interest. She can also have false beliefs about her interests. She can believe that certain things are in her interest, when in fact

they are not. These states of affairs do not produce any conceptual problems. How can we account for this conceptual difference between goals and interests? Agent's goals and desires are subjective in the sense that it is wholly up to the agent what they are. They are facts about the agent's attitudes. There might be an important difference between goals and desires to the effect that one cannot change one's desires at will in the sense that one can change one's goals, but both concepts are still subjective in this sense. On the other hand, *once there is a set of goals, it is an objective matter which things are in agent's interest*. What things are conducive to one's goals depends on causal facts about one's action environment, not on one's wishes or desires. I cannot decide that some things are in my interest, except by changing my goals. This difference makes it possible that one can have false beliefs about one's interests or one can even be completely ignorant about them.[3]

It is important to distinguish clearly between the interests of an agent, her beliefs about her interests and her public claims about her goals and interests. In sociology of science some critiques of interest explanations have rested on confusion on these issues. (For example Woolgar 1981, for a reply see Barnes 1981.) Although interest talk is an interesting phenomenon as a topic of study, it is not the same thing as an agent's interests or her beliefs about them. Agents' discourse about interests (and other motives) is an important part of agents' *interest cognition* and goal formation, but only one part. Interest talk may change constantly, but it does not mean that interests or beliefs about interests are changing at the same pace. Interests are not created in discourse, although discourse can prompt agents to change their goals. From the methodological point of view it is important to remember that neither an agent's own reports about her goals and interests nor another agent's accounts of her motives and interests are fool-proof sources of information about the agent's goals and beliefs about her interests.

It is possible to have *conflicting* goals. The pursuit of one aim can hinder or completely prevent the achievement of another. For example, one's short-term desires can be in conflict with one's long-term projects. The same situation can also be described as a conflict between an agent's interests. It is always up to the agent's own decision to solve the conflict. It can be solved either by weighting the various interests or by some other decision-making procedure. In this context I want principally to point out the methodological import of this phenomenon: how an agent's certain interest contributes to her behavior depends on her overall interests and goals. Just by knowing that a thing is in an agent's interest, one cannot infer the prediction that the agent will pursue the course of action that serves that particular interest. Interest reasoning is clearly holistic.

In a case where some interests are in conflict, it is easy, after the

fact, to see which interest is strongest just by looking at the agent's choices. But in the cases where the same action is conducive to two different goals, it can be difficult to decide which interest was the decisive factor. This is analogous to the cases of causal overdetermination discussed in Chapter 2. Of course, in such cases the agent does not need to decide about the relative importance of her goals. The problem is only for the analyst, not for the agent herself. The analyst's problem is that she cannot justify her claims about the explanatory relevance of the particular interest simply by noting its compatibility with her choices and actions. The agent may well have the goal attributed to her and have the relevant knowledge about the things that are conducive to the achievement of that goal, but these factors are not sufficient to explain her behavior. Her action might have been due to some other interest. To pick the right explanatory interest, the analyst has to know what other interests the agent has and also have evidence concerning the cases where these interests are in conflict to the effect that the interest she has picked is the overriding one.

Do one's interests change every time one changes one's aims? Not necessarily. It is possible that the things that are in one's interests remain the same. The same things can be conducive to both one's earlier and one's later aims. Some things, like nutrition, are preconditions for the achievement of almost any goal. One does not even need to know the specific goals of an agent to know her interest in adequate nutrition. In a modern society money is a generalized means of achieving things. No matter what one aims to achieve, having money at one's disposal is helpful. Wealth does not guarantee the achievement of one's goals, but it certainly helps in many cases. As a consequence, most people have at least some interest in having money at their disposal, and this interest remains stable while agents' specific goals constantly change. *Stability* of interests is one of the main reasons for the usefulness of interest concepts.

Changing one's goals is not the only way to change one's interests. A change in one's action environment can alter one's interests, although not all changes in the environment are relevant. Most changes in my action environment happen independently of me or as unintended consequences of my action; but there are also situations in which I might try to change my action environment deliberately. In such cases we can speak about agents trying to change their interests without changing their goals. It might be in my interest to change my interests. Naturally, it is more common that my interests change because of the changes in the action environment that are not initiated nor affected by my own action.

It is often thought that interests are always somehow selfish. Is this true? I would say that in a trivial manner, yes. One's interests are always related to one's own goals. My interests cannot be related to your goals

unless I have adopted the advancement of your goals as my own goal. I call this trivial, because the concept of interest does not set any limits for the kinds of goals an agent can have. The goals can be as selfish or altruistic as the agent wishes. If my goal is to save humankind, then it might be in my interest to solve the world's hunger problem. It is only essential that my interests derive from *my own* goals. When in ordinary parlance interests are sometimes set against the common good or altruistic behavior, the reference is actually to interests that derive from selfish goals. Naturally people often have selfish goals, but the use of the concept of interest does not commit one to the view that *all* human action is selfishly motivated. That position would require a separate argument.

Interests are often attributed to collective agents such as groups, institutions, professions, classes, firms, and states. The attribution of interests to collective agents is often taken as quite unproblematic: if we can attribute goals to a certain collective, then we can also attribute interests to it. However, there are two problems. First, it is not always legitimate to attribute goals to a collective. Second, the attribution of interest to a collective is often ambiguous. I will not discuss the first problem here since there is an extensive literature on social action that deals with this issue. (See for example Tuomela 1995.) I only want to make an observation concerning the second issue. When one is speaking, for example, about the interests of a profession, one can refer to two different things. First, the reference can be to the interests of the members of that profession. In this case, all (or most) members of the profession have similar individual interests. I call this *a shared interest*. In such cases all members can pursue their (shared) individual interests independently of each other. This is not so when one attributes *a collective interest*. In this case the members of the profession have an interest as a collective. Although even in this case the ultimate benefit from satisfaction of this interest goes to the individual members, the means of attainment of this satisfaction are different. The pursuit of a collective interest presupposes collective action, the members of the group cannot pursue it separately. It is not obvious that the agents can in fact act collectively. For example, there might be a conflict between the interests of the collective and the individuals who are its members which makes the incidence of collective action improbable. Only if the members can solve the problem of collective action can they act as a collective and pursue their collective interest. Usually this requires some internal organization. Firms and states often have such an organization, but social classes usually do not. Consequently, the attribution of interests to states is less problematic than the attribution of interests to social classes.

If we take interests to be a part of folk psychology, there are some interesting philosophical implications. Recall the discussion of the sci-

entific status of folk psychology initiated by Paul Churchland. His central claim is that folk psychology is a non-scientific proto-theory that should be excluded from the scientific worldview. His central argument is that beliefs as they are characterized in folk psychology are not to be found in the human brain, and consequently, they do not exist. (Churchland 1991.) It is controversial whether we should interpret folk psychology as making any claims about the neurological realization of beliefs as Churchland does (cf. Graham and Horgan 1988; Jackson and Pettit 1990c; Blackburn 1991; Horgan and Woodward 1991; Baker 1995; Egan 1995). I will not go into the details of this thorny discussion, since the point I want to make can be made without taking sides in the dispute. The point is that if interest concepts are a part of folk psychology, as sounds plausible, some parts of folk psychology would remain intact even if Churchland were correct. Interest theory does not make any *specific* psychological claims, which makes it immune against the eliminativist attack. No matter how radically future neuroscience changes our idea of workings of the human mind, it will not deny that human agents have representations of their environment and that they are goal-directed organisms. And this is all interest theory needs to be applicable. More realistically, I think, we should consider the existence of interest theory as a support for the critics of Churchland, who take a more shallow view of folk psychology. In this account folk psychology is not interpreted as making any specific claims about the realization of beliefs in human brain.

I hope that this discussion clarifies the concept of interest. Interests are goal-dependent. Interests pick up things from the agent's action environment that are conducive to the realization of her goals. Interests differ from goals in that an agent can be wrong about them. One cannot be unaware of one's goals, but one can be ignorant about one's interests. Interests derive from one's goals, but they are not by necessity selfish.

What about the use of interests in explanation? As mentioned before, there are requirements for the attribution of interests to agents. An agent must have goals, elementary rationality, and some knowledge about her action environment. These requirements apply to and restrict the explanatory use of interests. Interest explanation is based on the model of intentional explanation. In the model, an agent's action is explained by her beliefs and desires. The concept of interest allows us to make some interesting extensions and modifications to this pattern. For example, there are forms of interest explanation, filtering explanations, whose connection to the basic model of intentional explanation is very distant. I will return to these extensions and modifications in Chapters 6 and 7, but now I will briefly comment on interest in simple intentional explanations.

From the point of view of explanation, it is very important to no-

tice that there is a distinction to be made between action simply serving a certain interest and the pursuit of that interest. In the first case, the results of an action are conducive to the interest, but the connection between the action and its serving a certain interest is coincidental. The action was not taken because of that interest. As a consequence, the interest cannot explain the action in such a case. Only the pursuit of the interest can explain the action.

As intentional explanation is explanation by an agent's beliefs and desires, it is clear that one cannot explain the agent's action by interests that she does not know she has. The beliefs of the agent always mediate her interests in an action. This makes interests redundant in the sense that when we know what an agent's beliefs and goals are, her interests do not add anything causally relevant to the explanation. However this does not make the whole concept redundant. First, as noted, interests are important for the reconstruction of the agent's beliefs and goals. Secondly, a description in terms of interests allows the explanation to abstract away from the many causally irrelevant details of the agent's mental life. Not all of her beliefs and goals or changes in them are relevant from the point of view of explaining her particular behavior. Description in terms of interests allows picking (relatively) stable parts of the agent's practical reasoning. This can be seen most clearly in the case of the filtering explanation, to be discussed in the next chapter.

Interest talk among scientists

As interests are part of the repertoire of our mundane reasoning about social action, it is not a surprise that scientists also use the concept of interest to make sense of each other's behavior. This gives us an extra motive to study interest explanations: we need to understand them not only to make sense of what people in science studies are saying, but also to make sense of how scientists themselves make sense of the social world around them.

Geologist Edward Bullard (quoted in LeGrand 1988: 10) offers an example of scientists' use of interests (although not mentioning the term) to account for their colleagues' actions:

> There is always a strong inclination for a body of professionals to oppose an unorthodox view. Such a group has a considerable investment in orthodoxy; they have learned to interpret a large body of data in terms of the old view, and they have prepared lectures and perhaps written books with the old background. To think the whole subject through again when one is no longer young is not easy and involves admitting a partially misspent youth. Further, if one endeavours to change one's views in midcareer, one may be wrong …. Clearly it is more prudent to keep quiet, to be a moderate defender of ortho-

> doxy, or to maintain that all is doubtful, sit on the fence, and wait in statesmanlike ambiguity for more data ….

This quotation displays a use of an interest perspective that is comparable to that of many sociologists in terms of its sophistication. The fact that informants are versed in the vocabulary of interests gives an advantage to a historian or sociologist studying scientific episodes: informants are making part of the analysis. However, their accounts are not to be taken without a grain of salt. There are some features in the interest talk of scientists that are very important to keep in mind, especially when compared with the sociological use of interest concepts.

In scientific controversies it is often found that the disputing parties try to explain their disagreements by referring to social interests (and also other 'non-cognitive' influences). In Gilbert & Mulkay's terminology this use of interests belongs to *the contingent repertoire* of the scientists (Gilbert & Mulkay 1984). This repertoire is used to explain away discrepancies between what is believed by the agent herself to be true or rational and what the other people claim to be the truth. This explanatory setting is familiar from everyday life: an agent is trying to explain the difference between her own view and the position of someone else. The agent typically assumes that her own position is well justified. Why otherwise would she hold such a view? The existence of disagreement creates an explanatory problem for the agent: why others do not hold the position that she judges to be the right one. In such a situation, the position held by others is regarded as exceptional or abnormal, and it is accounted for by referring to some factors that are absent in a normal (or ideal) epistemic situation.

Many sociologists of scientific knowledge have noticed the *asymmetrical* nature of these explanations. The beliefs that are believed to be true or rational are thought to be natural and not in need of any special explanation, whereas 'false' beliefs are accounted for in terms of 'non-cognitive' factors like psychological idiosyncrasies, incompetence, social or career interests. (Gilbert & Mulkay 1984.) However, it seems to be an overstatement that scientists *always* offer asymmetrical explanations when accounting for divergence in opinions. It is also possible that they offer symmetrical analysis, at least for some episodes (Tammi 1999). In such a case, both opinions are seen as reasonable, and the divergence is attributed to insufficiency of data and inconclusiveness of arguments.

Apart from the cognitive aim of explaining discrepancies in opinion, there is also another motivation for this asymmetrical setting: the wish to discredit the competitor's positions. The idea is that by revealing the 'true' motives or causes of opponents' views, they can be disregarded as non-serious contributors to the debate. (Shapin 1979b; Brante 1984: chapter 6.) We might call this *the exposing stance*. The true motives are contrasted with the claimed motives in a manner that ques-

tions the reliability and credibility of the opponent. The implicit claim is that opposition supports its view because it serves its interests rather than on account of its epistemic credentials.

It is important to note that in the everyday talk of scientists (and philosophers), the term interest is only used for goals that are non-epistemic. So when they find a sociologist of scientific knowledge discussing social interests in science, they naturally assume that the sociologist is speaking about the influence of non-epistemic factors in science. Usually it is also wrongly assumed that sociologists of science are only interested in the non-epistemic aspects of science. However, sociologists of scientific knowledge have not limited the use of 'interest' in this way. They do not equate social with non-epistemic, but treat it as a general category which includes both epistemic and non-epistemic factors. Donald MacKenzie (1981: 219) points out that to say that the research choices or evaluations are goal-directed does not mean that they are inadequate, unscientific, or biased. To the contrary, it seems that in most case studies in the sociology of scientific knowledge the authors explicitly deny this kind of connection.

This denial has two aspects. First, sociologists of scientific knowledge claim that the impact of social interests is not always negative. Even completely non-epistemic interests can contribute to production of knowledge by leading scientists to carefully scrutinize some aspects of nature which might otherwise have remained unexamined, or they can make scientific research possible by making resources available for research. Thus, social interests can also be accelerators and motivators. (Shapin 1979b: 171.) Second, sociologists aim to understand the historical developments of scientific episodes, not to evaluate the validity of claims made by participants. Critics do not often recognize how consistently sociologists of scientific knowledge pursue the program of value-free social science. In the case of the study of science this means that sociologists fully abstain from evaluating participants' positions. Evaluative concepts like truth, rationality, successful, or progressive, are treated consistently as actor's categories. (Collins 1981.) Of course, sociologists of scientific knowledge have to make some epistemic, political and ethical value choices or commitments in their own research, but this only means that a sociologist cannot study herself (in real time) from an outsider's perspective. It is someone else's business to study the sociologist of scientific knowledge. To emphasize these two points, the sociologists of scientific knowledge have often aimed to study science at its best.

As it is not very common to associate goal-directedness with bias, why are sociologists of scientific knowledge often taken as exposers of biases in scientific development? The problem is not with the idea of the goal-directedness of science, but with the use of the concept of interest. The word interest often has a negative connotation that the con-

cept of goal does not have. As I noted, in much everyday talk, interests are associated with factors that can bias or corrupt an agent's judgment or make her action deviate from proper conduct.

An example of this usage can be found in the discussion of conflicting interests in science. In the context of this discussion, interests are understood as biasing factors that harmfully interfere with ethically sound scientific conduct, most commonly in the form of financial or personal interests. Financial interests refer to economic stakes scientists (or their relatives) might have in the results or interpretations of a scientific study. The much fuzzier class of personal interests covers such personal relationships as being a close relative, having a personal vendetta towards somebody, having an ongoing scientific priority conflict, having a teacher-student relationship, *etc.* (Resnik 1998). This conception of interest is extremely limited in contrast to the concept used by sociologists of scientific knowledge. The prevalence of this negative concept of interest has caused many commentators outside social studies of science to take all sociological claims about interests as claims about biases in scientific research.

It is not correct to say that sociologists of scientific knowledge do not deal at all with biases in the sciences they study. Scientists' own discourse is full of discussion and analysis of biases, lines of demarcation and of credibility in general. Sociological study of science would be invalid if all these aspects were left out. Instead of using normatively loaded concepts like bias, epistemic or convincing when describing of scientific episodes, the sociologist makes them a topic of her analysis. While scientists, and philosophers, regard themselves as participants in a discussion, the sociologist of scientific knowledge takes the position of an outside observer who does not have any stake in the issue. Instead of giving her own meaning to the concepts of science and epistemic credibility, she tries to follow how objects of her study themselves understand and demarcate these things, and to determine the consequences of this boundary work. The point is not to maintain, move, or change the borders between good and bad science or between science and non-science, but to observe how these boundaries are created, interpreted and sustained. And indeed, this is one of the most interesting aspects of the sociology of scientific knowledge. An outsider perspective on scientific discourse can give quite interesting results. (Gilbert & Mulkay 1984; Gieryn 1999.)

Purely descriptive approach is not without its problems. The problem is that the participants in the dispute under study, as well as outside commentators, often tend to take sociological descriptions as normatively loaded. When a sociologist of scientific knowledge shows that the production of a given piece of scientific knowledge is a process of social construction involving various social interests, outsiders might take these claims as some sort of criticism. Social construction can be

understood as some sort of fabrication. Especially those supporting the marginal position tend to hijack the sociologist to their side (Lynch 1993: 77-80). This is clearly a misunderstanding of the sociologist's position, which follows from the fact that sociologist uses the same vocabulary as the people they study.

And indeed, the descriptions the sociologist of scientific knowledge provides *can* be normatively relevant. The motives of the agents, research practices and standards are what the evaluation of the quality and relevance of research is about. They matter to the users of the knowledge claims and to people who are affected by them. Epistemic evaluation is much like ethical evaluation: it is difficult or impossible to completely step outside of it. Although the sociologist of scientific knowledge wants to remain an outside observer, she quite often has also some other hats to wear. As a user of knowledge, a tax-payer, a social activist, or as a citizen affected by expert decisions, she cannot keep the stance of a mere observer. This is the source of the inherent ambiguity in her position. Descriptive claims made from the observer's perspective can be read as normative comments from the user's perspective. I do not claim that this observation makes the sociology of scientific knowledge version of value freedom impossible; rather it calls for more care when publicly presenting sociological studies. What is needed is a clear distinction between the perspectives of an observer of knowledge production and those of a user of the knowledge produced.

2. Interests in the sociology of scientific knowledge

This discussion rests on the assumption that the concept of interest and interest explanations belong to the repertoire of everyday folk psychology. This assumption contradicts a widely shared view according to which interest explanations in the social studies of science are most often related to *the interest theory* developed by the members of the Strong Program, especially by Barry Barnes. (Barnes 1977; Barnes & MacKenzie 1979.) I claim that this is an 'optical' illusion: almost everybody uses interests in their accounts of science, but only the members of the Strong Program have elevated this concept to the status of a central theoretical term. Nevertheless, the Strong Program uses the same folk psychological concept of interest as everybody else. The only difference concerns the scope of that concept. To substantiate this claim I will first take a brief look at the early work of Barry Barnes.

The central epistemological idea and the justification for the possibility of a sociology of scientific knowledge in Barnes' early writings is the thesis of *instrumentalism*. It claims that knowledge-producing activities are goal-directed processes in which agents use their earlier knowledge as a resource in order to create new knowledge and to achieve their other goals. They are seen as innovative and selective users of their

cultural inheritance. Barnes contrasts this conception of knowledge production with the *contemplative* account of knowledge. According to this account, disinterested individuals produce knowledge either by passive perception or by contemplation. Knowledge is taken to reflect or correspond to reality in a very straightforward manner. In this conception only the mistakes and invalid descriptions of reality are taken to be due to the influence of social or personal interests disturbing the normal process of perception. (Barnes 1977: 1-2.) By arguing for instrumentalism Barnes tries to establish the following three theses:

1) both the production and the maintenance of knowledge are essentially social processes;

2) knowledge production is an active and goal-directed process;

3) the relation between knowledge production and certain goals and interests is constitutive rather than external.

Given that these theses are accepted, the possibility of sociology of scientific knowledge is justified.[4]

According to Barnes, the connection between knowledge and interests is neither internal nor logical. Rather, he claims that interests inspire the construction of knowledge out of available cultural resources in ways that are specific to particular social and cultural contexts (Barnes 1977: 58). He also notes that knowledge does not appear and disappear as various kinds of interests wax and wane. It often continues in an intellectual tradition, as a resource to be deployed in the furtherance of whatever interests are institutionally predominant. In this way, yesterday's 'ideology' can sometimes be transformed imperceptibly into today's 'science'. The validity or later usability of a certain piece of knowledge is not tied to the interests that originally created it. (Barnes 1977: 41-42, 1982: 112-113.) This is an important point. The fact that the choice or the development of a belief, idea or theory has been influenced by some interest, or that it has served some interest, does not have any automatic consequences concerning its validity or its later applicability for some other purposes. The sociological study of scientific knowledge does not rest on the genetic fallacy.

Instrumentalism is a general epistemological position, not a substantial sociological theory about science. Its main sociological implication is that scientific activity can be studied just like any other social activity. Barnes' discussion of instrumentalism takes place at same level of generality as Jürgen Habermas' theory of knowledge constitutive interests, which served as an inspiration for Barnes (Habermas 1972; Barnes 1977: 12-18). Neither of these theories is intended to be used directly in the explanation of specific scientific episodes. They both share the general thesis that scientists have an interest in 'prediction and control'. This thesis does not say anything about particular scientists or their goals. It is a claim about what is *common* to all knowledge production.

Theories on this level of generality are not able to account for the details of the concrete episodes in the history of science. To be able to do that, instrumentalism needs to be supplemented with a theory of specific goals and interests that drive scientists and other relevant social actors (Yearley 1982: 359-361).

The thesis of instrumentalism is the core of Barnes' position. I do not want to make any claims about whether the contemplative account of knowledge was accepted at time of the publication of Barnes' book, but it seems to me that by the 1990's it had been rejected by almost everybody in the science studies community. Almost everybody across disciplinary boundaries would accept Barnes' conclusions 1) - 3) above, provided that not very strict limits are put on the scientists' goals. Thus, I can skip further discussion about general epistemological issues and go directly to the substantive ideas about the goals and interests that drive science.

The next interesting question is whether Barnes or other advocates of the Strong Program have offered any substantial theory of interests to supplement their general epistemological thesis of instrumentalism. The answer is that they have discussed various suggestions toward a theory of goals and interests, but no single theoretical model has achieved dominant status. For example, analyses of historical case studies in terms of class interests were made in the latter part of the 1970's, two most notable examples being Donald MacKenzie's work on British statistics (MacKenzie 1978, 1981; MacKenzie & Barnes 1979) and Steven Shapin's work on Edinburgh phrenology (Shapin 1979a, b). These two works together with Barnes' (1977) critical review of various Marxist theories of class interests clearly show that the theories of class were seriously considered as a theoretical resource for understanding science and its relation to the surrounding society. However, class interests were never claimed to be the only or even the most relevant interests in explaining scientific episodes. Nor were they claimed to be an essential part of the Strong Program. Subsequent developments have shown that class interests as determinants of scientific development were only a temporary hypothesis. This hypothesis was dropped because the case studies conducted showed that class interests are a relatively unfruitful explanatory resource when the aim is to explain scientific episodes at a detailed level. The references to class interests vanished completely from the sociology of scientific knowledge literature after the beginning of the 1980's. But this does not mean that the concept of interest was also dropped. For example, Donald MacKenzie and Steven Shapin still use the concept of interest in their later works, but without references to social classes (Shapin & Schaffer 1985; MacKenzie 1990).[5]

Contrary to a common stereotype, the early Strong Program did not have any special externalist preference for their explanatory fac-

tors. For example, Barnes wrote in 1977 about his and MacKenzie's case study of the controversy between the biometricians and the Mendelians:

> In the case of scientific controversy, technical factors and esoteric professional interests must always be looked to first as a source of explanation. But in the present instance no sufficient basis for the dispute can thereby be found. The opposed forces did not have access to different kinds of evidence with conflicting implications, nor were differences in their training, and in the skills and competences they possessed and valued, of such magnitude and significance as to account for their different theoretical perspectives. Nothing in the esoteric scientific context satisfactorily accounts for the controversy. Nor there was any technical reason why the disputants should not have agreed to await further evidence, or accepted that both their accounts might have had merit and applied to different kinds of evolutionary change (Barnes 1977: 59).

This passage clearly shows that the intention was not to explain everything in terms of class interests or some other external factors. To the contrary, the passage suggests that Barnes and MacKenzie had a clear preference for an 'internal' or cognitive explanation for the controversy. Only when explanations in these terms were found wanting did the authors turn to the wider social influences. For sociology of scientific knowledge, the extent to which history of science displays external influences is purely a matter-of-fact issue.[6]

After the demise of theories of class interest no explicit theories of interests have been advanced. The 'goals and interests' (Barnes 1982; Barnes, Bloor & Henry 1996) have remained an abstract formula that has been filled in variously in different case studies depending on the details of the case. Nor has there been any theoretical discussion about the *kinds* of interests that influence science. The model of scientists' professional interests to be discussed in the next chapter has been the most popular scheme of analysis, but it has not received much theoretical attention either.[7]

The supporters of the Strong Program could have been clearer and more explicit in their discussions of interests. But in their defense, I have to point out that there is also something essential about their position in their openness about the kinds of interests that have or could have influenced the development of science. According to their naturalistic position, it is a matter of historical and sociological investigation to find out what kinds of interests are at work in a certain historical episode and how. The question is empirical and should not be prejudged by philosophical or theoretical arguments. This makes it important to keep the issue open. Secondly, which interests are relevant depends crucially on the intended *explanandum* the sociologist has. It would be foolish

to limit the list of possible explanatory factors before one has chosen which aspect of the scientific episode one wishes to explain. These two considerations show that the programmatic nature of the theoretical discussion is not a simple failure. There is a general justification for the abstractness of the formula 'goals and interests'.

This discussion might lead one to the conclusion that instrumentalism and the interest perspective are both rather trivial claims. Even the cognitive goals discussed in the traditional philosophy of science can fit the bill, because there are no limitations to the kinds of interests that are allowed to figure in interest explanations. Furthermore, nobody has denied the goal-directed nature of scientific activities. I take this to be a valid point, at least to a certain degree. It seems that most of the critical discussion of interest explanations in science is based on misunderstandings. Once the misunderstandings are resolved, there is not much to disagree with. However, instrumentalism includes two important ingredients that are missed by the triviality accusation.

The first ingredient is a heuristic for empirical analysis. Instead of unproblematically assuming that all prominent scientists try to achieve the cognitive goals that an analyst believes to be the proper goals of science, the interests perspective suggest that the analyst should ask who the principal agents in scientific developments are and what their characteristic goals and interests are? Although all scientists might have an abstract 'interest in prediction and control' (or an equally abstract goal of truth), instrumentalism advises us to look for differences in the agents' interpretation of this general goal. It might be that although at the verbal level everybody subscribes to the same 'official' goals, under the surface one finds a great deal of variance in accounts of how these abstract goals are translated into concrete actions. It probably also turns out that the scientists have a number of non-official interests that also influence the direction of the research and the interpretation of its results.

The second important ingredient is the idea that it is both legitimate and important to ask *why*-questions about the cognitive goals of scientists. The sociology of scientific knowledge does not differ from traditional history and philosophy of science by replacing cognitive goals by social interests as the central explanatory factors. Its novelty is in its insistence that it is also important to explain cognitive goals. Traditionally students of science have not been eager to ask why scientists have the scientific goals they have. The sociology of scientific knowledge emphasizes the importance of this *explanandum*.

The critics of the sociology of scientific knowledge have not always observed this extension of the list of *explananda*. They have instead interpreted it as a replacement. Surprisingly many philosophical critics of sociology of scientific knowledge have taken the supporters of interest explanations as opposing explanation of action in terms of rea-

sons (Laudan 1981: 188-193; Brown 1989: 24-30; Bohman 1991: 40-48; Niiniluoto 1991: 139; Roth 1996: 48-54). In the light of my previous discussion, this conclusion seems absurd. Interest explanations *presuppose* that agents' practical reasoning causally influences their behavior. I have not found a single explanation in sociology of scientific knowledge in which the causal relevance of agents' practical reasoning is denied.[8]

It is true that the Strong Program denies the causal and the explanatory relevance of *our* evaluations of the historical agents' rationality, but there is no reason to conclude that agents' reasons cannot affect their own behavior. There is a clear difference between what an agent believes to be 'good reasons', and what an outside observer takes to be 'good reasons'. And only the former are *causally* relevant to the particular historical episode. This is the true content of the sociological denial of the explanatory relevance of philosophical theories of scientific rationality, and I think that every self-respecting historian of science would agree. Sociologist of scientific knowledge also want to underline the fact that we can also try to explain *why* agents accept precisely those reasons as *good* reasons. One should sometimes try to transform one's *explanans* into an *explanandum*. In fact, sociologists of scientific knowledge think that this is a mark of a serious scientific attitude. After all, the immediate explanation might not be the most interesting.

In this context it is useful to bring up another claim that is often raised against interest explanations. Many critics have claimed that interest explanations reify social action and relations in undesirable ways. The users of interest explanations have been accused of taking society for granted (Latour 1987, 1988a; Pickering 1995: 151-152), treating interests as unmoved movers (Woolgar 1981: 372; Pickering 1995: 63-65) and taking social agents as 'interest dopes' (Woolgar 1981: 375; Jardine 1991: 178). I cannot here evaluate these claims as claims about individual uses of interest concepts. However, as general claims about interest explanations, these accusations have to be rejected as wrongheaded. There is nothing inherent in the idea of interest explanation that forces one to commit oneself to these fallacies. In fact, the interest perspective draws from exactly the same view of scientific work as the goal-directed, situated action that critics are offering as an alternative (Shapin 1988: 544).

Similarly, it is difficult to see why the explanation in terms of interests should in any way make it impossible to allow that interests can change with time or that new interests can arise. To the contrary, because interests are related to an agent's action situation and goals, it is difficult to see how this perspective could even deny the possibility of changes in interests. It does not follow from the fact that one cannot explain everything at the same time that the certain factors are given the status of unexplained explainers. Interest explanation takes the

agent's action situation and her goals as given in a particular explanation, but this does not preclude the possibility that these might change or that they might in turn be objects of explanation. Every explanation has to take something as a given, and not even a social constructivist can escape this basic fact about explanation.

What about the claim that the agents are being treated as 'interest-dopes'? I think that this characterization gives a very misleading picture. Interest explanation draws from the fact that an agent's action situation narrows the avenues of action that are available to her and that the agent knows this. It is not committed to the idea that an agent cannot change her goals or to the denial of the important fact that the formation of an agent's goals is often a social process. Both of these issues can be taken into account and analyzed within the interest perspective.[9]

Interests are widely used as explanatory resources in the social studies of science. It does not make sense to catalog all the users of the interest vocabulary in the field of science studies or to pinpoint slight differences in their applications of it. Interests belong to the repertoire of many historians of science (for example, Rudwick 1985; Sapp 1987; LeGrand 1988; Biagioli 1993; Kim 1994; Geison 1995; Proctor 1995; Lenoir 1997) philosophers of science (for example, Giere 1988; Hull 1988; Chalmers 1990) and sociologists of science (for example, Bourdieu 1977; Whitley 1984; Latour 1987; Stewart 1990; Fuchs 1992; Gieryn 1999; Segerstråle 2000) who do not explicitly support the Strong Program.[10]

The belief that there is something like 'the interest theory' is a myth. Like everybody else, the advocates of the Strong Program draw from common sense when attributing interest. What separates the Strong Program from other schools in science studies is the fact that they have raised the concept of interest to the status of an emblem. The difference is not in their use of the concept or in the way they use it. This observation makes understandable the seemingly confused nature of critiques of the interest explanations presented in the 1980's by people outside the Strong Program (Woolgar 1981; Yearley 1982; Latour 1987). They were not so much directed against the concept of interest itself, but against the vision of the goals and methods of sociology of scientific knowledge it was taken to symbolize.

With this discussion in mind, it is important to note that the interest perspective can work both as an *explanation-seeking heuristic* and an *explanatory pattern*. When we are using the interest perspective as an explanation-seeking heuristic, we are not assuming that explanatory assumptions of the interest explanation are satisfied. Rather, we are using the idea that the agents use the available resources to advance their goals to make sense of scientific activities. An explanation-seeking heuristic helps in the search for candidates for explanatory factors and even in making sense of empirical material before any explanatory

questions have been raised. The value of this kind of heuristic does not depend on whether the search will end with an actual interest explanation. The use of the heuristic can also facilitate research by showing that certain interests are not a relevant explanatory factor in the case under study. In my opinion it is beyond doubt that the interest perspective has proved itself useful as a heuristic in the sociology of scientific knowledge. What remains to be seen, however, is whether it has also been acceptable as an explanatory pattern. This is the issue on which the rest of this work will concentrate.

3. What the sociology of scientific knowledge aims to explain

I will discuss the explanatory patterns used in interest explanations in the next chapter. Before that, I will take a brief look at various *explananda* the sociology of scientific knowledge has at its disposal. The *explananda* are too varied to be discussed comprehensively. For example, I leave out *explananda* related to the use of scientific ideas and authority in culture and at society at large. I will focus on explanations directed to the inner workings of science. I will start with relatively simple *explananda*, and then proceed to more complicated, and interesting, examples. I will also point to the ways in which contrastive analysis can be used to clarify the intended *explananda*.

Sometimes one gets the impression that the following research heuristic is used: *all scientific disagreements can be traced to conflicting goals and interests*. This heuristic leads to the false diagnosis of many scientific episodes. Clearly it is possible that scientists have been exposed to different kinds of data, that their evaluations of methods and instruments are different, or that their belief in the strength of a support for a claim can be different without these differences being traceable to some differences in goals and interests. Not everything is explainable in terms of interests.

One factor that can create an impression of explanatory overambition is an imprecise *explanandum*. Quite often sociologists (and other students of science) do not state precisely what they aim to explain. There can be various reasons for this. For example, the explainer does not want to limit the scope of his explanation too prematurely; it might be that the same explanatory factors can also explain some other *explananda* apart from those under consideration. In some other cases the explainer might be insecure about the precise *explanandum* she is after. Third, often the emphasis in investigation is on descriptive validity, and explanation is only a secondary aim. For example, the presence of some interests might be an interesting feature of the historical episode, although their role in explanations might just be one of the background factors.

The idea that *explananda* of the sociology of scientific knowledge are contrastive is quite natural.[11] The sociology of scientific knowledge aims to explain the variation between cognitive practices and attitudes between individuals and groups. The contrastive setting has a fundamental role in its question setting. For example, one of the central background assumptions of the Strong Program is that everybody shares the same reality. If we combine with this assumption the interest in cognitive variation, the result is the typical explanatory setting in the sociology of scientific knowledge. In this setting the sociologist of scientific knowledge cannot appeal to reality or 'Nature' as an explanatory factor. Since everybody shares the same reality, it cannot be the explanation for the differences. Social causes and explanations are a more promising direction for research. (Barnes & Bloor 1982: 34; Fuller 1993.) Bruno Latour cannot be more wrong when he claims that the supporters of the Strong Program aim to explain everything in terms of 'Society' (Latour 1988a; Bloor 1999).

In the case of an individual, an agent's interests can be used to explain her beliefs and behavior. Let us start with beliefs. The simplest case of belief is a personal belief. I take personal belief to be an involuntary conviction about some state of affairs. Involuntariness means that personal beliefs cannot be (directly) changed at will. One cannot simply decide that one believes something. This means that interest-based practical reasoning cannot be a direct explanation of one's beliefs. But interests can still have explanatory value in indirect ways. First, an agent's interests can unintentionally affect the kinds of observations and evidence the agent will face by directing her information gathering, and by making a difference in her beliefs indirectly. For example, the agent's interests might cause her to pursue a particular avenue of research that happens to support certain beliefs and to disregard other avenues of research that happen to offer conflicting evidence. Secondly, an agent can affect his beliefs in indirect ways, for example by character planning. Here interests influence the goals of planning. Thirdly, a psychological mechanism of wishful thinking offers a way in which an agent's hopes and desires can causally affect her beliefs. To my knowledge, only the first mechanism has been used in the sociology of scientific knowledge. (Miller 1978 includes examples of explanations that combine interests and the mechanism of wishful thinking.) However, personal beliefs are rarely or never the principal object of explanations in the sociology of scientific knowledge.

There are methodological problems in accessing the private beliefs of historical or contemporary agents, but the principal reason for them not being objects of sociological investigation is that they are not very interesting. Sociology of scientific knowledge understands science as public and communal activity, not as the private mental life of an individual. The object of interest is primarily collectively held knowledge, not individual beliefs. The idea is to study how a claim gets the

public status of credible knowledge, not how individuals end up with their personal beliefs. The latter is left to the cognitive science and psychology.

More important than personal beliefs is public acceptance, which refers to a public commitment to a certain proposition by an agent. Although we normally suppose that personal beliefs are not in contradiction with publicly stated beliefs, it is possible that an agent does not believe personally something that he publicly presents as her belief and vice versa. Similarly, there can be variation in the strength of conviction between publicly stated and privately held beliefs. In contrast to personal beliefs, public acceptance is an action that is voluntarily initiated. It is also notable that groups, institutions and organizations can publicly accept propositions, but they cannot have personal beliefs.[12]

A public commitment to certain ideas, or their public rejection, is not an absolute necessity for scientific research. A scientist can proceed with her work infinitely with a kind of 'wait and see' attitude. There will always be some further evidence, and possibly alternative theories, that could affect her judgment in the future. This makes the incidence of public opinion about a theory somewhat arbitrary. As sustaining judgment is always an option, it forms a contrast against which the scientist's action can be seen. This contrast is something that the sociologist of scientific knowledge can set herself to explain. One can try to explain the timing, the forums, and the strength of public acts of acceptance or rejection. Why did the scientist decide to take a stance when she did? What goals did she think she could achieve by it? And, could these goals also be used to explain the forum and the manner of acceptance? Similar questions can be asked about changes of opinion. Why did the scientist think it prudent to change her position and to contradict her earlier views? All these questions make sociological sense, and the answers to them can reveal interesting things about the social structure of science.

There are obviously more modalities of belief than simple belief and disbelief (both in private and public contexts). First, it is possible to have no beliefs at all about some issue. Let us call this attitude nonbelief, to distinguish it from disbelief, which is a belief in the negation of the statement. Second, there are degrees of belief. The degrees of belief can be modeled in terms of probability calculus, but I do not take this approach to be very fruitful in this context. Scientists do not generally, at least not yet, use probability calculus to characterize the strength of their conviction. A qualitative approach is more fruitful. We can start with Larry Laudan's distinction between acceptance and pursuit of a scientific theory. When an agent accepts a theory, she treats it as true, whereas when she only pursues it, she only works with it without any belief as to its truth. (Laudan 1977: 108-114.) Similar distinctions can be developed for attitudes towards hypotheses, empirical findings and research methods. We could also follow LeGrand in adding the modali-

ties of rejection, entertainment, and employment to our taxonomy (LeGrand 1988: 92). It is important to note that scientists themselves recognize this variety of attitudes towards their theories. This should be taken into account in descriptions of scientific activities, because different attitudes have important behavioral consequences. The analyst's account should be at least as sophisticated as the participants' own accounts. Naturally, these subtleties of attitude offer interesting contrasts for explanation. Why does a scientist adopt the attitude of entertainment rather than acceptance towards a theory, or why does she reject an empirical claim instead of remaining noncommittal toward it?

Public acceptance is not the only kind of act that can be an object of explanation in sociology of scientific knowledge. Even more important *explananda* are the choices scientists make. In constructivist studies the decision-ladenness of scientific work is one of the central theses (Knorr-Cetina 1981). Scientists choose their graduate school, institutional affiliation, co-workers, research staff, instruments, theoretical interpretations, research topics, publications forums, participation in public controversy etc. They also choose to trust (or distrust) their colleagues, experimental results, reliability of their instruments, etc. All these choices can become objects of sociological curiosity, and the contrasts are easy to find. It is obvious that agent's goals and interests show up in these explanations in some role. After all, the supposition is that scientists' behavior is intentional. The sociologist of scientific knowledge can ask why scientists chose to act this way rather than some other way? Or she can compare how two groups of scientists act and try to explain the difference. Similarly she can contrast scientists' behavior with their earlier (or later) behavior and try to account for the difference.[13]

However, it is important to keep in mind that the sociology of scientific knowledge is not mainly in the business of explaining the actions of individual scientists. The more important *explananda* are at the collective level. The sociologists of scientific knowledge are more interested in explaining why a certain claim has the status of scientific knowledge (or pseudo-scientific rubbish) than in explaining why a particular person believes it. In fact, individual-level *explananda* are interesting because of their connection to more general issues. *Explananda* at the collective level are more difficult to categorize than those at the individual level, but it is not impossible to give an idea of this variety by using some representative examples.

The simplest phenomena at the social level are the aggregate effects of individual behavior. By aggregating individual level attitudes and choices one can try explain the fortunes of scientific ideas, theories, research methods, etc. If one assumes that individual scientists or research groups act more or less independently, their success or failure

is basically an aggregate effect of individual choices. For example, a theoretical interpretation will not enjoy much success if it is not useful to scientists in their pursuit of their goals. On the other hand, if it useful and attainable and it is clearly better than alternative interpretations, one can expect great success for it. The changes in the fortunes of a theory might be explained by changes in the interests of research groups, which in turn are explained by changes in their action situation. The availability of new equipment and methods, the appearance of new competing research groups, or changes in funding sources are examples of such changes of situation.[14]

Scientific controversies are especially interesting objects for explanation. Controversies provide sociologists, historians and philosophers of science a window to the factors and assumptions behind scientific work that would otherwise stay invisible. In public controversies, scientists are forced to articulate positions and arguments that would otherwise remain unarticulated. This has made controversies one of the central objects of study in the science studies.

There are various things about controversies that one might wish to explain. The first is the very existence of a public controversy. Not all scientific disagreements lead to a public controversy. Sometimes even very large differences in scientific opinion remain behind the scenes, whereas sometimes seemingly quite insignificant issues give rise to a huge scientific controversy. This suggests that the causes for the existence (or non-existence) of a scientific controversy are not to be found entirely in the contents of science, but in the goals and interests of the scientists. When there is a controversy to be explained, a number of important *explananda* arise.

First, one could wish to explain the timing of the controversy. Why did it start when it did? Why did it last as long as it did? Why did the controversy reach closure when it did? What were the mechanisms of the closure? Why did it close the way it did? The contrasts used in these explanations are obvious. The second important class of *explananda* is the participants. Why did some scientists participate in the controversy, while other remained outside? Why did the participants choose to take the positions they did? Why did some participants lost their interest in the controversy, and why did some others joined in so late? Again, the contrasts arise naturally from the historical context. One can also try to explain the character and the forum of the controversy. A controversy can take place in the public press, in specialized journals, in scientific meetings or in a court of law, so one natural *explanandum* is the choice of the forum for the controversy. One can also compare the tone, the emotional engagement, the intensity, and the scope of the controversy with other controversies and ask for an explanation of these differences.

For sociologists of scientific knowledge, references to the incommensurability of competing positions or to the inability of the partici-

pants to understand each other's positions are not an adequate explanation of the controversy. To the contrary, they are things to be explained. The problems related to understanding technical details and aims of the opposed theory can normally be overcome, given that participants have some incentive to do it. A controversy does not derive from philosophical problems of translation or understanding, since scientists can in principle acquire competence in the competing traditions. The real problem is not that of understanding, since the disputants can understand each other's positions and goals perfectly, and still disagree. The explanation is to be found in the different goals and interests that motivate the participants. Note that incommensurability, and the breakdown in communication that goes with it, can sometimes even be an achievement that requires much effort from the participating scientists. (Barnes & MacKenzie 1979: 51-59; see also Biagioli 1993: chapter 4.)

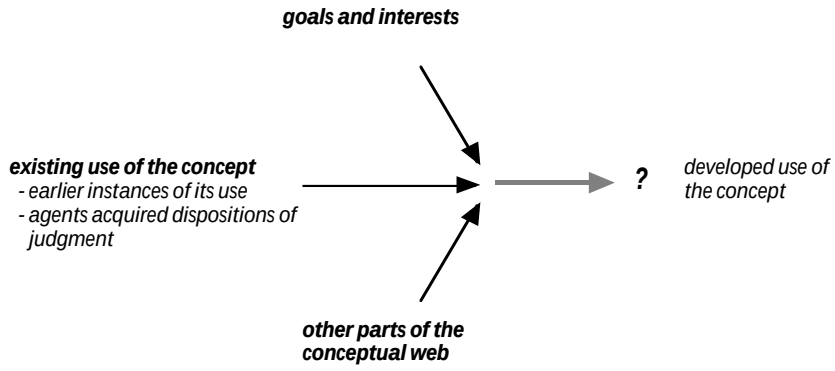
Another important class of explananda consists of the institutions[15] created and maintained by scientists. Among these institutions are definitions of concepts, conventions and standards. Institutional *explananda* are closely related to the *finitist* account of meaning and concept application. This account is inspired by Wittgenstein's discussions of rule following (Wittgenstein 1953; Barnes 1982; Kripke 1982; Bloor 1997, see also Pettit 1993a), and it is based on the following five basic theses. According to the first thesis, *the future application of a concept is open-ended*. The idea is that a previous application of the concept is not able to fully determine its future use. There are always other (formally) possible ways of continuing the application of a concept apart from the one chosen. The second thesis says that *no act of concept application is indefeasibly correct*. The idea of this thesis is that collective judgments have a central role in defining what will count as a correct way of using the given concept. There is no naturally correct way to apply it. The third thesis, which is a natural corollary to the second, is that *all acts of concept application are revisable*. This thesis emphasizes the fact that apart from future use being open-ended, the earlier use of the concept is in principle also revisable. Individual earlier uses can always be later judged to be mistaken. The fourth thesis says that *the successive applications of the concept are not independent*. This thesis emphasizes the fact that the earlier concept use has a role in determining the future use of the concept. Finally, the fifth thesis generalizes this holism by claiming that *the applications of different concepts are not independent of each other*. The claim here is that the use of a concept is related to a web of other concepts. This web of concepts works as a whole and it is subject to the requirement of coherence. (Barnes, Bloor & Henry 1996: 54-59.)

These five claims are formal claims about concept application (or more generally about meaning). The finitist account of meaning is interesting in this context because of its role in the creation of sociologi-

cal *explananda*. The finitist account of meaning can be regarded as a generator of sociological *explananda* and appropriate contrasts. Its main consequence from our point of view is that meaning has the character of a social institution. Because of the finitist nature of meaning, it is always possible to use the concept differently, and this opens up the contrast space for the creation of interesting sociological *explananda*. Why was the extension of the concept continued this way rather than some other salient way? The question can be expanded to the more general and interesting question about the cognitive order: which persistent and systematic causes explain the development of the use of the concept over time? (Barnes, Bloor and & Henry 1996: 119-120.)[16]

In this account of meaning, the earlier use or definitions in terms of other concepts are not *sufficient* to determine the correct use of a concept. The concept user's dispositions and habits account for the routine way the concept is usually used. In normal situations the concept usage is automatic and 'natural' or as Wittgenstein puts it, blind. However, this routine cannot be used to account for the changes in the use of a concept, or its application for the novel situations. Similarly, the dispositions acquired by the users cannot be used to make sense of situations where the competent users of the concept disagree about the proper use. One needs one more element to this picture of the social constitution of meaning. The required extra ingredient, according to the supporters of this view, is the concept users' goals and interests. The habits of concept usage are seen as immediate causes of routine action, which action would be modified if the goals and interests require it. The routines are sustained, developed, modified and abandoned as the users see fit. (Barnes 1982: 101-114; Barnes, Bloor & Henry 1996: 124-127.) These elements can be summarized in the following figure.

Figure 1. The finitist account of concept application



Goals and interests inform agents' judgments in novel situations and in situations where there is a disagreement about the proper usage of a concept. They can also be used to explain the stability of the application of a concept over time.

Examples of interesting institutional *explananda* are the cultural maps of science and its surroundings. The lines of demarcation and attributions of epistemic authority are objects of much sociological curiosity. Cultural boundaries between social activities and other actors' categories are here understood as institutions. They are collectively created and maintained cultural maps that tell the members where they may and may not go and how to behave in the social landscape. The agents in the pursuit of their various goals use them as cultural resources. One can raise interesting questions about these boundaries. Why are the boundaries drawn the way they are? Why was one of the suggested cultural mappings accepted instead of others? Why is the boundary work done by certain groups more successful than that done by other groups?[17]

Characteristic qualities of research groups or scientific fields can also be regarded as institutional features and related to the interests of the participating scientists. For example, Robert Kohler (1994) accounted sociologically for what he calls the moral economy[18] of *Drosophilia* geneticists, and Harry Collins (1998) set himself to explain the differences in the evidential cultures[19] of collaborating Italian and American research groups studying gravitational waves.

A sociologist of scientific knowledge is principally interested in explaining why institutions are the way they are, rather than in some other conceivable manner. The controversies in and around science show that the goals and interests of scientists and other relevant parties are important parts of the process of creating and maintaining these institutions. Consequently, there is a role for interests in the explanation of the institutions in science.

In this chapter I characterized the concept of interests and its application in the social studies of science. I have also drawn attention to the kinds of explanatory questions that interest explanations try to answer. However, I have not yet said much about the interest explanations themselves. This will be the topic of the next chapter.

Notes to Chapter 5

1 Naturally, in the beginning of a research program some of the central concepts are left open for further explication (Barnes 1974, 1977), but when after 20 years things seem almost as unclear as in the beginning (Barnes, Bloor & Henry 1996), there is cause for concern. On the other hand, the critics have not done any better. For example, the standard reference for a critique of interest explanations, Woolgar 1981, apart from its being based on a misunderstanding of the basic tenets of interest explanation, is actually

directed against of whole idea of explanation than against interest explanation. Yearley 1982 includes a somewhat more informed critical discussion. Examples of philosophers of social science that have contributed to the confusion are Roth 1987 and Bohman 1991.

2 For a history of the notion of interest in political theory, see Hirschman 1977 and Myers 1983. For current discussions in political theory, see Reeve & Ware 1984 and Hindness 1986. Peillon 1990 reviews how interests are understood in the sociological tradition. For a discussion from the point of view of moral philosophy, see Feinberg 1984.

3 In political philosophy some theories use the notion of real interest in a sense that seem to make interests wholly independent of the agent's goals. This impression is wrong. It is true that in these theories it is possible that the agent's real interests and his actual goals and desires are almost totally separate due to a false consciousness. However, these theories presuppose a normative theory about goals and desires that are used to determine the real interests. For example, Steven Lukes takes real interests to be things that would be conducive to the goals and desires that an agent would have in a situation without the influence of coercion and ideology (Lukes 1974: 40-42).

4 The Strong Program advocated by Barnes and Bloor does not reduce to the thesis of instrumentalism. For a fresh introduction to the basic themes, see Barnes, Bloor and Henry 1996.

5 For Barnes' later thoughts about class interests, see Barnes 1995.

6 It is not possible to explore the internalism/externalism issue in depth in this context. For an excellent discussion of this topic, see Shapin 1992. As Shapin notes, sociologists of scientific knowledge have always been more interested in the empirical study of 'boundary work', rather than doing it by themselves. See, for example, Gieryn 1999.

7 Bloor (1983) suggests that we should interpret Wittgenstein's references to needs as references to social interests in the sociological sense. Wittgenstein uses the concept of need when he tries to account for changes in language games. As Bloor notes, these suggestions are never developed either by Wittgenstein himself or by his followers. According to Bloor, 'needs' should not be construed as individual appetites, but as collective phenomena. Changes in language games are to be explained by references to shifts in the goals and purposes of the participants. (Bloor 1983: 47-49.) This is certainly an interesting suggestion, but it is difficult to flesh it out in the absence of a more developed account of interests.

8 However, there are plenty of opposite instances. For example, as early as 1974 Barry Barnes wrote: "Despite a large body of opinion to the contrary, there is no necessary incompatibility between causes and reasons as explanations of action, indeed reasons can be listed among the causes of action" (Barnes 1974: 70).

9 In fact, a similar criticism can be raised, for example, against many of Bruno Latour's analyses. Some of his studies show actors as slightly different kind of 'interests-dopes' who, for some unexplicated reason, seem to accept whatever goals and interests Latour's central actors rhetorically impose on them. Of course, this is an artifact produced by Latour's way of story-telling and his focus on successful network builders. However, to avoid this accusation Latour has to allow his (human) actors to think for themselves about their

goals and interests, which brings him in line with other interest theorists. Latour would probably object to this claim, but apart from his claims that he is using the concept of the interest in a different sense than the Edinburgh School (Latour 1988a: 260), it is very difficult to find in practice any substantial conflicts between the use of the concept between the actor-network theory and the Edinburgh School. (For example, see Latour 1988a: chapter 3; see also Shapin 1988; Bloor 1999: 99-100.) The real differences, and there are many, between the approaches are to be found elsewhere. (See for example Bloor 1999.)

10 For extensive reviews of the work done in social studies of science, see Shapin 1986, 1995b and Golinski 1998.

11 This thesis can be supported by observing that for example Barry Barnes regards explanation in contrastive terms (Barnes 1974: 71-78; Barnes, Bloor & Hendry 1996: 119-120). Unfortunately he has not developed the contrastive idea further.

12 In the recent philosophical literature there are various distinctions between belief and acceptance (Engel 1998). It is notable that all the concepts in the philosophical discussion are confined to the perspective of agency and in that way to the 'private' sphere. My distinction between private belief and public acceptance should not be confused with these distinctions. See Geison 1995 for an interesting account of the difference between private and public beliefs in the case of Louis Pasteur.

13 For a clear example of a question setting of this kind, see Collins 1999.

14 An example is Andrew Pickering's case study of high-energy physics that will be discussed in the next chapter.

15 The concept of institution is here used in the philosophical sense (Searle 1995 and Bloor 1997), not in the more traditional sociological sense, which refers to social organizations.

16 Notice one interesting consequence for social explanation. The finitist account makes explanation of action by reference to rules and norms problematic. The problem is that rules cannot tell agents how they are to be applied. The full explanation always presupposes a richer account of the practices underlining the norm-governed behavior. Consequently, the norm-based explanations are not alternatives for other forms of explanation, but presuppose them. (Barnes 1982: 101-104; 1988: 26-32.)

17 For examples of studies of 'boundary work', see Gieryn (1999) and Barnes, Bloor, and Henry (1996: Chapter 6).

18 According to Kohler, moral economy regulates both authority relations within the group and the access to the means of production and the rewards for achievement (Kohler 1994: 6).

19 Collins distinguishes three dimensions of the evidential culture. The first dimension concerns the question whether it is the job of an individual scientist (or laboratory) to take responsibility for the validity and interpretation of scientific results, or whether this responsibility rests on the wider scientific community. The second dimension concerns the interpretative boldness that characterizes the interpretations scientists make of their data. Finally the third dimension concerns the evidential thresholds scientists set for the significance of their results (Collins 1998: 302-307).

Chapter 6
Interest explanation in action

In this chapter I will discuss some representative examples of interest explanations. The aim is to illustrate explanatory mechanisms presupposed by these explanations. The idea is to make the nature and the scope of these explanations more visible by explicating their underlying structures. All examples are from the field of social studies of science.[1] With the help of these examples I wish to show that there is no unique pattern of interest explanation. There are different kinds of interest explanation and their structures are quite different. Situated goal-directed human action has a central role in all of my examples, but there is some variation in the details of this role.

I will start by describing redundant uses of interest vocabulary and then proceed to more interesting examples of interest explanation. I will distinguish three forms of interest explanation that are based on different explanatory mechanisms. The first is based on an agent's practical reasoning. The idea is that her choices are rational responses to the challenges her action environment sets for the advancement of her objectives. My examples include both social and professional interests as *explanantia* of choices by scientists. I also draw attention to the role of unintended consequences of action in more complex interest explanations.

The two other forms of interest explanation I will discuss are intentional and non-intentional filtering explanations. In an intentional filtering explanation, certain agents are able to control or filter the work of other agents in a way that is conducive to their goals. As a consequence, the actions of the controlled agents can be explained by interests of the controllers even if the two groups do not share the same interests. In non-intentional filtering explanations the explanatory work is done by a cultural or social selection process analogous to natural selection. In such explanations interests can have an important role in the selection environment that determines the fate of beliefs and prac-

tices to be explained. These three patterns of explanation cover most of the informative uses of interest explanation in social studies of science.

In the final section I will discuss the model of cycles of credibility originally developed by Bruno Latour and Steve Woolgar. I argue that this model provides a theoretical elaboration of the sociological notion of professional interest. The model can be used both to explain some general characteristics of modern academic science and to understand the relationship between social and epistemic goals in scientific work. Especially, it can be used to show that sociological explanations of science in terms of professional interests are not explanations that refer purely to non-epistemic goals. This observation shows that most criticisms of sociology of scientific knowledge by philosophers of science are misguided.

1. The redundant use of interest vocabulary

Sometimes the concept of interest is used very loosely, as in Everett Mendelsohn's paper "The political anatomy of controversy in the sciences" (1987). The paper makes references to "cognitive, metaphysical and social interests" (Mendelsohn 1987: 105), "philosophical, political and religious interests" (110), "professional interests" (115), "upper middle-class interests" (116), and the "self-interest of the scientists" (118). Clearly, all these interests are not interests in the same sense, and apparently he uses the term as a kind of umbrella-concept to cover various ways in which science is goal-directed activity. Generality of this kind is not helpful. It is difficult to see how cognitive, metaphysical or philosophical interests could be understood in a more specific sense of interest. It seems that Mendelsohn is mostly speaking about goals, commitments or just simply preferences.

From the point of view of the analysis of interests presented in the previous chapter, most of Mendelsohn's uses of the concept are redundant. The term 'interest' can either be dropped from the sentence without any loss of information, or it can be easily replaced by terms like goal, desire, or commitment. Often this simple substitution makes the sentence more comprehensible. These uses of the concept are redundant because they do not offer any information about the agent's action situation. As a consequence, the special explanatory import of the concept of interest is lacking. (See also Yearley 1982: 371.)

A couple of examples suffice to make this point. Mendelsohn's liberal use of the interest vocabulary is clearly one. It would be more appropriate to replace cognitive interests with cognitive goals, philosophical interests with philosophical commitments or goals, metaphysical interests with metaphysical preferences etc. The concept of interest is not doing any theoretical work in these cases and even its stylistic effects are questionable. Similarly, Donald MacKenzie's early discussion

of Pearson's "interest in maximizing the analogy between nominal and interval variables" (MacKenzie 1978: 49-50) can be better understood as a simple (cognitive) goal. I am not claiming that these uses of the concept of interest are totally uninformative or non-explanatory. They might well refer to important explanatory factors, but the concept of interest is not required for this purpose. Furthermore, if we take the project of sociology of knowledge seriously, the things Mendelsohn and MacKenzie mention are more naturally treated as *explananda* rather than *explanantia* for the sociological study of cognitive activities.

2. Interests in practical reasoning

The core component of most interest explanations is an agent's practical reasoning toward achieving her goal(s) within a given action situation. The agent is assumed to be rational in the sense that she is able to choose adequate means to advance her objectives. The idea is that an agent's choices are rational responses to the challenges her action environment sets for the advancement of her objectives. As an agent's ultimate goals are often quite distant and require long and complex series of actions, it is rational to divide goals into series of subgoals. These subgoals can then work as proxies for the original goal. The advantage is that their more concrete nature helps the agent to reason more easily about appropriate actions in her everyday engagements. It is important to keep in mind the means-ends structure of the subgoals: the achievement of subgoals is a means to achieve the more distant goal. Interest explanations of this kind are instances of the standard pattern of intentional explanation. An agent's beliefs and intentions together with her practical reasoning bear the explanatory burden.[2]

Mackenzie on Pearson and eugenical interests

We can find an example of this sort of subgoal formation in MacKenzie's work (1981) on Karl Pearson. According to MacKenzie, Pearson had political and social goals related to the eugenics movement that motivated his scientific work. More specifically, his explanatory claim is that Pearson's commitment to the eugenics movement was an important influence on the formation of the cognitive goals of his statistical work. MacKenzie's explanatory claim rests on the following counterfactual:

> (EC 1) If Pearson had not been influenced by eugenics and committed to it, he would not have chosen the same topics of scientific research as he did and, as a consequence, he wouldn't have developed such statistical methods as he did.

The nature of MacKenzie's explanatory claim is often misunderstood. He is not claiming that social interests had a direct influence on Pearson's

scientific choices and decisions. Rather, the eugenical motivation influenced his evaluations of the importance of certain statistical research problems, and his evaluation criteria for the results of such research. In more general terms, the basic idea of this explanatory pattern is the following. Political and social goals explain scientist's cognitive or scientific goals, which in turn explain many of the details of her scientific work.

MacKenzie clearly assumes that Pearson's choice of the topics for his scientific work was based on more or less rational considerations about the best means to advance his eugenical goals. There is no evidence that he would deny that Pearson has reached his conclusions using ordinary practical reasoning. MacKenzie's explanatory practice does not differ from the standard 'internalist' practice in the history of science by replacing epistemic factors with social factors, but by extending the list of things requiring explanation. For example, he is not denying that reasons can be causes; he is rather suggesting that there are causes for scientists' reasons. The sketchiness of Mackenzie's explanatory apparatus in his writings has led some of his critics to misunderstand his position, but the fact is that his position is not very radical. Clearly, if there are problems with MacKenzie's explanation, they are not due to his reference to interests.

The real challenge for explanatory claims of this kind is to provide empirical evidence for the explanatory relevance of the counterfactual. To provide such evidence MacKenzie examines two scientific controversies that Pearson participated in. The first is the public controversy between biometricians and early Mendelians in the beginning of this century (MacKenzie & Barnes 1979; MacKenzie 1981: Chapter 6). The second is the more scholarly debate between Karl Pearson and George Yule over the statistical analysis of nominal variables (MacKenzie 1978; 1981: Chapter 7). The basic explanatory pattern in both of these cases is similar. The controversy and its continuation is explained by the different cognitive goals of the participants, and these differences are in turn explained by their different social and political goals. As the basic explanatory pattern is similar in both cases, I will only discuss the latter example.

In MacKenzie's analysis of the Pearson-Yule debate the *explanandum* is the difference between Pearson's and Yule's mathematical statistics and the *explanans* the difference in the goals of their statistical activities. According to MacKenzie, Pearson showed both great effort and scientific integrity in his pursuit of the research program of eugenics. This program created a specific data-processing demand for Pearson. He needed measures of the associations of nominal data that were numerically comparable to the interval-level correlation coefficient. To meet this demand, Pearson devised a series of measures of association. MacKenzie notes that Pearson's interest in measures of as-

sociation diminished when his practical statistical concerns shifted. This suggests that there is an important connection between these two issues in Pearson's scientific work. In contrast, Yule did not have any involvement with any eugenics research program. His practical commitments did not give rise to a similar dominant *desideratum* and, as a consequence, he was prompted to develop a looser and more pluralistic approach to the measurement of association. In MacKenzie's account, this rather esoteric controversy was a result of different cognitive goals, and it was sustained not because the participants did not understand each other's positions, but because their cognitive goals were proxies for their different practical goals. Interestingly, MacKenzie did not use incommensurability as an *explanans*, but rather as the *explanandum*. (MacKenzie 1978; Barnes & MacKenzie 1979; MacKenzie 1981: Chapter 7.)

I take MacKenzie's explanatory pattern to be sensible. Of course, there might be some other plausible explanation candidates for this particular difference, but the evaluation of the explanation against the historical data is not my concern here. This particular explanatory factor might not have much wider application, but it is strong enough for MacKenzie's purpose, which is to show that social interests had a significant role in the genesis of scientific results that are not, in the light of our current knowledge, false or fabricated. This aim might now sound very modest, but in the 1970's it was ambitious enough to motivate MacKenzie to write a book-length study.[3]

The trouble with MacKenzie's approach is not with the explanation pattern he uses, but in the scope of the explanatory factors he has chosen. They are sufficient for the existential claim MacKenzie wants to make about the influence of wider social interests on British statistics of the time, but they are not very useful for a study of the details of statistical research. When one considers the amount of historical work that would be required in order to substantiate MacKenzie's relatively modest claim, it is understandable that later sociologists and historians of science have turned to other explanatory factors. In subsequent sociology of scientific knowledge, the typical explanatory factors used in association with this explanatory pattern are the professional interests of the scientists that promise to be relevant for a wider array of *explananda*.

Professional interests

The starting point of the model of professional interests is the observation that within any given scientific specialty there exists a distribution of different skills and scientific competencies. Some individuals are more skillful in mathematical or theoretical analysis, whereas some others might show excellence in using some specific experimental tech-

nique or apparatus. What these competencies are depends on the scientific field under consideration. From the point of view of scientists these competencies and abilities represent a considerable *investment*. They have spent a lot of time and other limited resources in acquiring and developing their skills. These investments serve as a basis for their professional interests as scientists. If their goal is to continue their professional existence, and especially if they wish to advance in their career, they had better make good use of their investments.

Within the reward system of modern academic science, scientists have every reason to wish to show the value and the scope of what they can do. This can be done in various ways. The value of one's abilities can be shown positively by demonstrating them in the applications that are considered important or interesting by one's colleagues. By showing excellence in research and by producing results that other scientists can, and want to, use in their own work, a scientist shows the relevance and the value of her investment. Scientists do not just passively take the evaluations of problems and suggested solutions as given, they can be very active in their efforts to convince their audience of the quality and importance of their work. This makes the study of scientific rhetoric especially interesting. The general perception of the importance of a piece of research, or of a whole scientific field, is a discursive achievement. Furthermore, scientists are not limited to discuss just their own work. Since the value of a scientific contribution can be determined by a comparison with other contributions, a scientist can advance her interests also by criticizing both the quality and the scope of the scientific competencies and results of the rival colleagues. By finding faults in the contributions of others, scientists can raise the relative value of their own work.

Pickering on professional interests

Andrew Pickering's study of high-energy physics is a well-known example of the use of professional interests in the explanation of the dynamics of scientific work (Pickering 1980, 1984). In this case study Pickering aims to explain why the 'charm' model of the newly found elementary particles defeated the competing 'color' model in high energy physics (HEP) between 1974 and 1976. Pickering argues, drawing on the Duhem-Quine thesis, that the 'color' model was never conclusively falsified. This claim motivates his search for an explanation: if scientists were never logically forced to accept 'charm' and to reject 'color', why did the wide consensus for 'charm' and against 'color' establish itself by mid-1976? Pickering argues that one influential factor was the new experimental data collected during that time. However, for Pickering, this cannot be the whole explanation. Following his interpretation of the Duhem-Quine thesis, he claims that the scientists

who supported 'color' could have interpreted the experimental results in a way that would not have challenged their conviction. (Pickering 1980: 111-114.) Interestingly, the *explanandum* is explicitly contrastive: the aim is to explain why the majority of scientists working in HEP accepted 'charm' rather than 'color'.[4]

How does Pickering go about explaining this? He points out that the interaction between emerging experimental data and two different theoretical responses must be supplemented with an account of the preexisting matrix of 'interests' in the HEP community. In his Kuhn-inspired model, the central difference between the competing models was the difference in possibilities for future work for the majority of researchers in HEP community. 'Charm' allowed scientists to perceive the newly produced experimental data as 'puzzles' that had connections to the existing research practice. Furthermore, the proponents of the 'charm' model offered a set of tools to attack these puzzles. These tools were derived from the recently developed gauge theory in quantum field theory, which also offered theoretical support for the 'charm' interpretation. (Pickering 1980: 117-120.) On the other hand, the 'color' model could only offer solutions that were, according to Pickering, 'sociologically *ad hoc*'. The interpretations of the new experimental results that the supporters of 'color' model were able to provide had no connection to the existing research practice and consequently no connection with the interests of the HEP scientists. According to Pickering, this gave an undisputable edge to the 'charm' model and explains why the consensus emerged around 'charm' and not around 'color'. (Pickering 1980: 120-121.)

Pickering's explanatory model – encapsulated with the title 'opportunism in context' – is the following. Each scientist has at her disposal a distinctive set of resources for doing her research. The resources are results of the 'investments' the scientist has made during her scientific career. The resources may be either material or cognitive. An example of the first kind of resource is access to particular pieces of experimental apparatus, and an example of the second is expertise in a certain experimental technique or theory. The distinctive constellation of resources characteristic of a particular scientist is formed in the course of her professional career. The investment in a particular expertise creates an interest in the deployment of that expertise in the scientific work. Pickering characterizes this interest as 'a particular constructive cognitive orientation towards the field of discourse'. (Pickering 1980: 109; 1984: 10-12.)

This model of the dynamics of scientific research is based on the observation that a scientist's resources may be well or badly suited to particular contexts and issues. The research strategies of individual scientists are structured in terms of the relative opportunities presented by different contexts of research for the constructive exploitation of their

resources. According to Pickering, this explains the attitudes scientists have towards experimental results. If a theorist can find resources for a constructive analysis of data, it can expected that he will not spend a long time searching for ways to challenge the experiment. Similarly, if an experimenter finds the experiment, through the medium of theory, suggestive of new problems for a experimental inquiry, he will not be very prone to look for ways to challenge the results. The fortune of a scientific idea depends on the possibilities it creates for scientific puzzle solving. (Pickering 1984: 13.)[5]

In 1990 Pickering characterizes this opportunism-in-context model as follows:

> Doing science is real work; real work requires resources; different scientists have different degrees of access to such resources; and resources to hand are opportunistically assembled as contexts for constructive work are perceived. My claim, exemplified many times over in *Constructing Quarks*, was that if one understands scientists as working this way then one can understand, in some detail, why individuals and groups acted as they did in the history of particle physics (Pickering 1990: 692).

This succinctly summarizes the basic idea of Pickering's explanatory model. The scheme of analysis is not very sophisticated, but it certainly is in accordance with common sense. Pickering is primarily explaining why individual scientists and research groups choose to pursue the avenues of research they do. Scientists chose to work with the 'charm' model because it gave them challenging, but solvable, research problems. It also gave them an opportunity to make use of their existing competencies and to continue their earlier research agendas. The 'color' model did not have similar advantages.

Pickering is not claiming that the scientists accepted the 'charm' model as true because it fitted their interests. Rather, they chose to work with it because it promised them an opportunity to make use of their research competencies. Since the model provided positive ways to continue existing scientific work, the scientists did not have much incentive to find faults in it, or to develop alternatives to it. On the other hand, the 'color' model lacked these virtues. Furthermore, the supporters of the 'charm' model had an interest in criticizing it, because by finding problems in an alternative approach they could legitimate their own approach.

In Pickering's explanatory scheme, the almost general acceptance of the 'charm' model is an unintended consequence of the advancement of the professional interests of individual scientists and research groups. Since it is usually assumed that scientists work with the theory that they find most plausible, the constellation of interests described by Pickering leads to a situation where most physicists are regarded as

supporters of the 'charm' model. The same constellation of interests also ensures that the 'charm' model will be seen as a progressive research program, since most scientists will be working for the advancement of this program. There would be a constant flux of results supporting the model. On the other hand, the lack of interested scientists guarantees that the 'color' model will be seen as a stagnating approach since it does not produce any new positive results. Moreover, the constellation of interests creates incentives to produce arguments and data against it, which strengthens the general impression of its demise.

Pickering's explanation shows that not all interest explanations are explanations in terms of the *intended* effects of the interested action. Sometimes it is more appropriate to look for the *unintended* consequences of such action. Unintended consequences make explanations more interesting. Usually information about the goals and practical reasoning of the scientists provide news only for outsiders, but patterns of unintended consequences of action can be interesting for scientists themselves.

Intended and unintended consequences

The distinction between intended and unintended consequences requires some clarification. I suggest that we should distinguish between four different questions:

1) What did the agent intend to bring about?

2) Which consequences did the agent believe her actions would have?

3) Which of the anticipated consequences did the agent take into consideration in her practical reasoning?

4) Which actual consequences of her actions did the agent recognize afterwards?

The intended consequences refer to the expected consequences of action that are an agent's reason for her actions. She acts in order to bring about these effects. If the agent intended to bring about x, it is obvious that she believed that her actions would have that consequence, and that this consequence was taken into account in her practical reasoning. This is true on purely conceptual grounds. On the other hand, it is contingent on whether the agent was successful in bringing about x and whether she was in a position to recognize that her action in fact had this consequence.

Things become more complicated when we consider consequences that are not intended by the agent. First, there is an issue concerning the agent's beliefs about the consequences of her future actions. What did the agent expect to be the consequences of her actions? One can

also ask whether these beliefs were correct (or warranted)? The anticipated consequences can be divided into two classes. First, there are the consequences that the agent took into consideration in her practical reasoning. Second, there are those that she did not take into account. Questions 1)-3) are relevant for explaining the agent's action. The fourth question is irrelevant from this point of view, because a later cognition is not a causally relevant factor in explaining the agent's action.

On this basis, we can conclude that there are three kinds of unintended consequences of action:

> UIC 1) consequences that were not the agent's reasons for her actions, but were taken into account in her practical reasoning
>
> UIC 2) consequences the agent anticipated, but which were disregarded in her practical reasoning
>
> UIC 3) consequences the agent did not anticipate.

In practice, it is often difficult to draw the line between intended consequences and unintended consequences of the first kind. Similarly, differentiating between categories UIC 1) and UIC 2) is tricky. The folk psychological concepts underlying these categories are by their nature so fuzzy that there is bound to be many borderline cases. Usually this fuzziness does not matter. The important point is that the category of unintended consequences is more heterogeneous than the first impression suggests. When using the concept in this chapter, I will always use it in a broad sense that covers all three cases.

Gilbert and the policy of differentiation

Nigel Gilbert's (1976; 1977) study of competition among research groups illuminates the model of professional interests in a way that complements Pickering's discussion. It also provides further examples of interest explanations that build on unintended consequences of action. According to Gilbert, researchers (and research groups) use "the policy of differentiation" to avoid openly competitive situations. The differentiation is achieved by selecting problems of research that are believed not to be under current investigation by other scientists. This raises the question: why do scientists use this policy? Gilbert's answer in terms of research group interests suggests that scientist's success in terms of professional career and peer recognition depends on her abilities and opportunities to find and follow up research problems that differ substantially from those studied by others.[6]

In Gilbert's model, research groups avoid competition by concentrating, whenever possible, on techniques and problems that are not in the main focus of other research groups. The motivation for this 'niche

formation' is clear. Once one group has achieved a clear lead in a particular field, the other groups tend to avoid unrewarded duplication of the effort and open, and often disadvantageous, competition. The reasons for this are quite obvious. The rewards for the replication of earlier research results are small in comparison with new and original research. Secondly, competition with an already established research group (within its own speciality) is often unrewarding, since it is not enough to achieve their level excellence. To justify the duplication, the results should be significantly better, and this is difficult to achieve without a significant investment of resources. The established group has often a clear advantage in know-how due to its previous research on the topic. This makes improving on their results within a reasonable time even more difficult. Usually a challenge is launched only if there are some novel tools or a new approach to the problem. (Gilbert 1977: 109-110.)

According to Gilbert, the level of competition within a scientific speciality depends on a number of factors. First, differentiation of research is not possible if the number of topics to study is limited. For example, during the initial stages of the development of a research area, there may be only a very limited number of topics that have been recognized as suitable research problems. In such cases competition will be a natural result. Secondly, this differentiation is difficult when there is general agreement among the members of the speciality on the overriding importance of finding solutions to some clearly defined problems. The problem might be a 'bottleneck' problem for future research in the field, or its solution might be highly rewarded by actors outside the scientific speciality. (Gilbert 1977: 109-110.)

The third factor is the cost and availability of research apparatus and materials. If a widespread consensus on the significance of particular problems is accompanied by the fact that the research apparatus needed for research is cheap, unsophisticated and easily available, one can expect competition to increase. On the other hand, if one or a small number of groups are able to obtain monopolistic control over a scarce but necessary piece of equipment, other groups will avoid competition. The fourth significant factor is scientists' awareness of each other's research. In the initial phases of the formation of scientific field and its associated networks, the participants might not know enough about each other's research to avoid competition. Once the institutionalized means of communication are in place, the level of direct competition tends to go down. (Gilbert 1977: 110-111.)

Gilbert's account explains why open competition between research groups is an exceptional occurrence in science. Competition can be expected to occur only when the four factors mentioned above have enough influence. Naturally in the background there is another kind of competition: the competition for resources to do research. This is quite important. It is because of the scarcity of resources scientists *have to*

care about the novelty value of their research.

The policy of differentiation has also consequences from the point of view of a scientist's career. Gilbert considers the difficulties a scientist faces when she wishes to begin to work on a previously unexplored topic. The first task is to find an interesting problem that has some potential to be developed into a research program. Among the radar meteor researchers Gilbert studied, the most interesting research problems were found either as by-products of research on other topics or they were based on theoretical predictions. According to Gilbert, opportunities for new topics of study open up frequently for scientists, but most of these topics appear less promising when compared with their current topic. There is a natural explanation for this. As a result of previous research experience, the scientist has gained particular skills and knowledge and usually she has also acquired specialized equipment suitable mainly for the current research topic. (Gilbert 1977: 113-114.)

Abandoning work on a current topic involves learning new skills, modifying or changing existing experimental apparatus, and maybe acquiring completely new devices and materials. It also presupposes familiarity with some new sections of the relevant scientific literature. In addition, a move to a new topic may imply breaking some of the scientist's informal contacts with colleagues and possibly the neglect of a scientific reputation that was carefully built up while working with the old topic. This explains why those who start working on new topics tend to be either new entrants to the field or those who think that their current topic has run out of interesting or solvable problems. As a consequence, new topics are often initiated by research students and postdoctoral fellows directed to these topics by their supervisors or senior members of a research network. (Gilbert 1977: 113-114.)

In Gilbert's account, researchers are motivated both by the hope that they can make worthwhile discoveries and by a desire to gain recognition (and rewards associated with it, such as promotions). The research topic to be chosen must fulfill a number of criteria in order to promote a scientist's career and other scientific ambitions. First, to gain recognition for his work the topic must be such that others are not already concerned with it. As already noted, replication of someone else's results is not highly rewarded. Secondly, the results expected from the research must be such that they are of interest to the audience of fellow scientists. Here the 'interesting' problems are those whose solution has significant consequences for the current and future work of other scientists. Thirdly, at least a part of the research community must also judge the results to be more than moderately successful. These requirements limit severely the number of good research topics. Within these constraints, the scientist can still choose between the risk of opening up an entirely new area in the hope that it will deliver significant findings, or

a more timid strategy of selecting a topic that is unlikely to yield especially novel or unexpected results. Usually research students are advised to take the latter course, and more experienced researchers prefer the riskier strategy. (Gilbert 1976: 199-203; 1977: 114.)

Eventually, if everything goes well, the researcher choosing the riskier strategy comes to be regarded as the expert in her topic will therefore be favorably placed to attract funds and other forms of assistance. As a consequence, her laboratory comes to be seen as a 'center of excellence' for her research area. These processes are largely self-reinforcing, and they culminate in the professional recognition of the scientist's academic status by acceptance into the scientific elite. Successful scientists tend to become the foci of research networks, and as a consequence, they are in a particularly good position to become aware of new avenues of further research. Hence, they are also in a position to suggest topics to research students and others looking for research possibilities, and to change the course of their own research to more promising topics. Their position allows them to act as 'sponsors' for the work of junior researchers, not only providing ideas but also securing funds for their projects and legitimacy for their work. On the other hand, scientists choosing the safer strategy will not probably receive more than a moderate degree of recognition from their peers. Those scientists whose riskier strategy does not produce important results are confined to even less recognition. They face the choice of changing their topic once more or moving to some related but separate specialty. The continuation of their research on their original topic requires interested sponsors who can legitimate and support the continuation of research. (Gilbert 1977: 115.)

One important way to do successful research is to resolve longstanding problems by using a technique that has not been previously applied but with which one is familiar. This was a strategy that some radar scientists in Gilbert's case study used when they started to work in meteor astronomy. The application of methods from another field generated a continuing stream of further research problems, leading eventually to the growth of a new discipline. Of course, such 'interesting' problems are not easy to find. Most frequently they arise by chance, as consequences of unexpected observations made with an apparatus that was intended for other purposes. Sometimes they arise because it is realized that techniques currently in use in one area of research are applicable to unsolved problems troubling another area. A new speciality emerges only when the successful solution of such problem generates a stream of interesting research problems. This occurs quite rarely. Gilbert emphasizes that the possibility of achieving success in this way depends on the interaction of a number of social and intellectual factors. Among these is the prior competence in the appropriate methods and techniques, and the often accidental discovery that

these research tools might be fruitfully applicable in some new area. One also needs some contact with existing practitioners in the new field and an available pool of possible recruits to the new field. And finally, for a major success, one needs some luck or judgment to find a research problem whose solution has ramifications outside its immediate context. (Gilbert 1976: 199-203.)

As we have seen, Gilbert uses the model of the professional interests of scientists to answer a number of *explananda*. First, it is used at the individual level to explain choices by individual scientists and the differences between the choices of scientific novices and more accomplished researchers. However, his emphasis is on the explanation of more general patterns. Among these are the prevalence of the policy of differentiation, the emergence of new research areas, and the typical characteristics of the researchers who start the work in these areas. The explanatory power of his model is considerable when compared with MacKenzie's model. A combination of some simple assumptions about the motivations of scientists with some observations about the social structure of the science allows Gilbert to sketch a model that has a number of interesting applications. The application of this model to particular research areas allows for the construction of interesting explanatory hypotheses.

3. Explanation by other people's interests

An example of intentional filtering explanation can be found from Donald MacKenzie's work. He claims that eugenical interests had an institutional connection with the development of statistics at the turn of the century in Britain (MacKenzie 1981: Chapter 5). According to him, much of the statistical work done by people who did not have any personal commitment to eugenics can also be explained by the reference to eugenical interests (MacKenzie 1981: 107-109).

MacKenzie's empirical claim is that the work done by the biometrical school on statistical association can only be understood in terms of the goal to maximize the analogy between association and interval-level correlation. This goal was the result of the connection between biometric statistics and the needs of the eugenic research program. However, he claims that there is no way of telling whether this goal was personal motivation for all biometricians. Their motives could have been such things as the enjoyment of mathematical puzzles, advancement of a professional career, or whatever. (MacKenzie 1981: 107-110, 220.) Clearly, this sort of interest explanation cannot rest purely on agents' practical reasoning as in the previous examples. We need something else.

To understand this explanatory pattern, we need some background facts. First, the biometrical school (the Department of Applied Statis-

tics, which included both the Biometric and the Eugenic Laboratory) led by Francis Galton and after him Karl Pearson, was the only unit working on statistical theory in Britain during that time (MacKenzie 1981: 101-104). The second relevant background fact was that both of these leaders were committed supporters of eugenical ideas. The work on statistics was for them a subgoal proxy for the advancement of the eugenics program. The third relevant background factor is the tight control they supposedly had over the work done at the department and their use of their position to advance the goals of eugenics (MacKenzie 1981: 105-106). With the help of these facts we can see the structure of the explanatory pattern MacKenzie has in mind.

I call this type of interest explanation *intentional filtering explanation*. In it, certain agents are able to control or filter the work of others in a way that is conducive to their goals. This dependency of the controlled agent's action is the basis for the explanatory pattern. The controlled agents need not share the goals of their controllers. It is only required that they have goals which allow the controllers to steer their work by manipulating their action environment. For example in MacKenzie's case study, if we make a sensible assumption that members of the institution acted to further their professional interests, we would find a pattern where they also advanced the interests of eugenics as an unintended (but probably recognized) consequence of their action.

This explanatory pattern allows MacKenzie to strengthen his central claim about the influence of eugenical concerns in statistics. With this mechanism he can show that Pearson's eugenical motivations had a much wider influence than just his own statistical work without going too deep into the beliefs and goals of the statisticians working at the biometrical school. It also helps to back up MacKenzie's claim about the influence of the professional middle class' interests in the statistics. Given that both Galton and Pearson identified themselves with the professional middle class and saw eugenics as means to advance its interests, the above facts allow MacKenzie to make his claim without supposing that all statisticians identified themselves with professional middle class or with the eugenical movement. Again, the principal problem with this explanatory claim is not with the assumptions it makes, but in its explanatory scope. There simply are not that many details of the statistical research that MacKenzie's explanatory factors could account for.[7]

Hacking on military interests and form of scientific knowledge

A variant of the intentional filtering explanation is a case where the explanation refers to the unintended consequences of interested action. A clear example of this is Ian Hacking's thesis about the influence of

military funding on modern physical sciences. Hacking's *explanandum* is what he calls "the form of scientific knowledge". By this concept Hacking refers to the idea that existing knowledge somehow determines which issues are *possible* candidates for topics of scientific research. His claim in relation to the military involvement in physical science research is that it might have influenced the form of physical science knowledge in a manner that precludes some questions about physical reality outside of what is considered to be meaningful and feasible topics of research. Of course, this influence might have not been only negative: if there had not been huge military investment, certain aspects of physical reality might have stayed outside of what are currently considered as meaningful research problems in physics. Hacking's thesis is not very specific or detailed, but the general point he wants to make is clear: military funding made a difference in the form of current physical knowledge. (Hacking 1999: Chapter 6.)

Now let us take a closer look at what is going on in this explanation. The aim of the military and the defense industry is to promote research that is useful for the development of weapon systems and other military applications. This goal explains their funding choices. On the other hand, the scientists did not necessarily share this objective, for their idea may have been to use the resources provided to advance their own professional and cognitive goals. However, because of their dependency on funding and other resources, their research activities served the purposes of the military and can be partly explained by these interests. To this extent, the pattern is similar to MacKenzie's example considered above. The crucial difference is that in this case neither party had the goal of shaping the form of scientific knowledge to what it is. Actually, given that Ian Hacking originated this concept in the 1980's, neither party had the faintest idea of the issue. The *explanandum* is an unintended consequence of action.

Hacking's explanatory claim rests on the following counterfactual:

> (EC 2) If there had not been massive military interest in the physical sciences, the form of knowledge in physics today would have been different.

Now, in principle, this explanation sounds sensible. The fact that the *explanandum* is unintended by the agents in no way invalidates the explanatory pattern. The only problem concerns the ambiguity of the contrast. Hacking's sketchy discussion does not give a very concrete idea as to what could have been different. This is partly due his concept of form of knowledge, and partly due to the scale of issues he discusses. I guess we might try to imagine what physics would have looked like if it had not become Big Science, but thinking about alternative forms of knowledge might prove too difficult. This may raise the concern that there might not be any clear contrast for Hacking's *explanandum*. But if there is such a thing, then *intentional filtering explanation by unin-*

tended consequences of action seems to be a legitimate explanatory pattern.

The basic components of intentional filtering explanation are not different from the components of ordinary interest explanation. The crucial addition is the idea that some agents are able to manipulate or control the action environments of the others. This allows for the possibility that the actions of the latter can be explained by interests that do not derive from their goals. This is an important observation that underlines the fact that interest explanations cannot be reduced to a single pattern of intentional explanation.

4. Non-intentional filtering explanation

Donald MacKenzie's work on the history of British statistics also provides an example of what I call *non-intentional filtering explanation*. MacKenzie (1981: Chapter 2) aims to show a connection between the eugenics movement and the class structure of the nineteenth century Britain. According to him, eugenics fitted both the interests and the experiences of the professional middle class in Britain better than those of any other social class. He also points out that the development of eugenical ideas required knowledge and skills that only the educated professionals had. This explains, according to him, why most supporters and developers of eugenics were from this class. (MacKenzie 1981: 22-25.) I will not discuss the details of eugenical ideas or the structure of British society, but try to explicate this explanation.

MacKenzie characterizes this explanation as *structural*. He does not want to claim that the supporters of eugenics supported and accepted it as a consequence of a conscious calculation of the utility and of a comparison of the advantages of various social ideologies. He does not explicitly rule out that some individual members of the movement might have adopted their positions as a consequence of this kind of process. However, he does not want to rest his explanation on this presupposition. He explicitly allows that at least some of the supporters felt that their actions in the name of the movement were motivated by unselfish and disinterested moral reform (MacKenzie 1981: 24-25). His analysis also shows no evidence that his explanation presupposes unconscious processes of interest calculation. Clearly, this explanation is not intended to be an individualist one.

Neither does the explanation rest on a collective intentional action by the members of class or on an intentional action where the 'class' is the acting subject. The professional middle class of that time was not an organized group that could have acted as a collective agent. Its members shared an understanding of their social status as a 'professional middle class', but this was all. More significantly, not all people who shared the same class position were members or supporters of the eu-

genics movement. Some even opposed it explicitly. (MacKenzie 1981: 26.) Consequently, this is not a case of intentional explanation by group action.

Can this be some sort of a functional explanation? MacKenzie does not refer to 'consequence laws' (Cohen 1978) or any other functionalist ideas. As the background of his work is the causal program of the Edinburgh School, we can safely suppose that there has to be some sort of a non-intentional causal mechanism to back up his explanatory claim. What is this mechanism? My claim is that the only acceptable option is that it is some kind of non-intentional evolutionary filtering process.

But first we should arrive at some understanding of the *explanandum*. Is MacKenzie explaining why the supporters of eugenics were for the most part from the professional middle class, or is he explaining why there was a eugenics movement at all. MacKenzie is not very clear here. What he wants to achieve is some sort of connection between these two entities to make the relationship between the eugenics movement and the structure of British society understandable. Clearly the explanatory setting is quite ambiguous. For example, what criteria are necessary so that something might count as a eugenics movement or as a movement by the professional middle class? These are very difficult questions to answer. But let us suppose that we have some kind of answer to these questions to the effect that we can formulate some broad counterfactuals that can serve as explanatory hypotheses.

Let us concentrate on the first alternative: why the supporters of eugenics were for the most part from the professional middle class. We can distinguish earlier and later phases of the development of the eugenics movement in terms of filtering mechanisms that are operative. First, given that there is some kind of eugenics movement 'in the air', we can presume that a significant number of its supporters in the nineteenth century were from the professional middle class. This presumption is based on the MacKenzie's claim that the development of eugenical ideas required such knowledge and skills that only the educated professionals had. The idea is that these skills were necessary requirements for the development of this social program. If nobody had them, it would not be possible to develop the eugenical ideas to any significant extent. The professional middle classes also had the experience of being constantly evaluated and classified on the basis of merit and tests during their lives in educational institutions. This circumstance helps us to see that the eugenics movement might have been more congenial to the life experiences of the professional middle class than, for example, the experiences of less educated social groups. These sketchy ideas seem to suggest that, given that there was a eugenics movement, a signifi-

cant part of its supporters were from the professional middle class. However, they cannot show that there had to be a eugenics movement in the first place. The members of the professional middle class might have turned to some other social ideology. As a consequence, we do not have an explanation for our second alternative. Interests are not the right kind of things to explain historical contingencies like this. From the point of view of interests it was a pure historical coincidence that eugenical ideas were there to be developed.

Let us next turn to the mechanisms of the later phase. Now, assuming that some kind of connection between the professional middle class and eugenical ideas was developed, two auxiliary mechanisms come to our aid. These can be seen as positive feedback mechanisms that reinforce the original filtering mechanism. First, as eugenics was developed by members of the professional middle class, its later versions most certainly reflected more and more the experiences and the interests of the members of the professional middle class. This certainly enhanced its appeal among the members of this class. Secondly, as eugenics became the *de facto* ideology for the professional middle class (or some section of it), some kind of path dependence entered into the picture. As an already existing and supported ideology, it had a clear advantage in comparison to all other possible ideologies that the members of professional middle class might have adopted. The earlier use (and development) of the theory for particular purposes became an added reason for others to use it. As it already was developed and supported, it had a clear advantage when compared to the ideologies that were only 'possible'. Similarly, to shift to some other (more or less incompatible) ideology might have proved costly in terms of their credibility in the eyes of others.

As usually happens with evolutionary explanations, the setting of this explanation is retrospective: given that we observe the fact to be explained we try to figure out why it occurred, rather than something else. An evolutionary explanation always leaves much latitude for historical contingency. It is clear that we could not have predicted the career of eugenics, but seeing that it developed as it did, we can try to figure out why things happened this way.

MacKenzie need not claim that the eugenics was the best possible ideology for the professional middle class. He only needs to claim that it was better than its actual rivals. Some other social ideology might have been even more congenial to the interests and experiences of the professional middle class, but as a historically contingent fact it never developed to the point that it could have competed with eugenics. As eugenics was there first, it was locked in as the ideology of a certain section of the professional middle class.

Actual and imaginable alternatives

In this connection it is useful to comment on an argument that James Robert Brown has presented against the explanatory use of interests. Although it is presented in the context of theory choice, it can be modified to be applicable to MacKenzie's case study. According to Brown, this argument is 'the ultimate refutation' of the explanatory import of interests. The argument goes as follows. Let us assume that we have theories T_1, T_2, T_3, ..., all underdetermined by the empirical data. Now imagine that one of them, say T_2, is chosen because it serves some interest, say I_1. However, the question now arises, 'Why T_2?' In principle, there are infinitely many theories T_2, $T_{2'}$, $T_{2''}$, … that do equal justice to the data *and* to interest I_1, so why was one singled out over the others? Now, one answer might be that T_2 was chosen over $T_{2'}$,… etc., because it served interest I_2. But, again, there are infinitely many theories T_2, T_{2*}, T_{2**},... that equally serve interest I_2 (as well as serving I_1 and accounting for the data). According to Brown, this regress does not have happy ending for the sociologist of scientific knowledge. If eventually a final interest I_n is evoked to explain the choice of T_2, there are still infinitely many theories T_2, T_{2+}, T_{2++}, … which do equal justice to interest I_n (and to all of the other interests and to the data). As a consequence, the final choice is left completely unexplained. Neither data nor interest determines the preference for T_2 over T_{2+}, etc. On the other hand, if there are no "final interests", then the regress goes on infinitely and nothing is explained in this case either. (Brown 1989: 54-56.)

David Bloor has replied to this argument by taking Shapin's case study about Edinburgh phrenology (Shapin 1979a, 1979b) as an example. The same points apply directly to MacKenzie's study, since these two case studies are very similar. Bloor allows that other theories might also have served the interests of the Edinburgh middle-class as well as, or even better than, phrenology. It was a historical contingency that phrenology was available as a theory that could be used to challenge the *status quo* and to legitimate social reform. (Bloor 1991: 171-172.) We can generalize Bloor's point by noting that Brown seems to have misconstrued the contrast class of an interest explanation. Interests explain the choice between existing alternatives, not choices between T_1 and all the imaginable alternatives. Interest I_1 might have made a difference among the available theories T_1, T_2, and T_3, although it is not able to make a difference among all the possible theories. Brown's argument rests on a improper reconstruction of the *explanandum* of an interest explanation. Consequently, it is much less than the 'ultimate refutation'.

Bloor's reply takes the wind out of Brown's sails. However, even after this answer Bloor still has a more limited problem. What if the adoption of two or more existing alternatives would be equally conducive to

an agent's (or group's) interests? It is clearly possible that there are cases where an agent's interests *underdetermine* her choices. In cases like this it seems that interest cannot make the difference that explains the choice. The explanation has to invoke some other facts. This version of the problem of underdetermination is not fatal for the interest perspective, but it shows its limits. Interests are not the magic formula for explaining every aspect of science.

What about non-intentional filtering explanations in general? Are they viable? And what is the role of interests in these explanations? Let us start with some observations about the Mackenzie's example. First, its explanatory power seems very weak. This derives from the ambiguity of the explanatory setting and from the generality of the explanatory factors used. The intended *explanandum* is vague, and even if we construe the explanations as referring to a cultural selection mechanism analogous to natural selection, the contrast of the explanatory question is open. Natural selection always chooses between two or more concrete alternatives. It tells us why one rather than other was selected. Now, in Mackenzie's example both the existence and the identity of alternatives are left open. Consequently, it is very difficult to say why the eugenics movement prevailed. It might have been due to its 'cultural fitness' or due to some random effect analogous to the genetic drift. And as in earlier examples, the interests MacKenzie refers to in his explanation are such that they do not have very wide application. There are not that many *explananda* his explanatory factors can satisfactorily answer. Both these considerations make MacKenzie's explanation a less than ideal example of a non-intentional filtering explanation.

Many sociologists and historians of science who have experimented with evolutionary analogies, but their explanatory patterns have remained weak or sketchy analogies (cf. Gilbert 1977; Knorr-Cetina 1987). The most significant contribution in this category has been made by philosopher David Hull (1988). Hull's theory is based on explicit evolutionary considerations, and the professional interests of scientists play a central explanatory role in it. Scientists' professional interests work as a selection environment for theories, methods and research topics. Were the interests different, the outcome of the selection process would be different. However, Hull's actual explanations are sketchy, so I am not able to provide an example that fully shows the explanatory potential of the non-intentional filtering explanation in social studies of science. However, I will return to the nature of non-intentional filtering explanations in the next chapter.

5. Professional interests in a reputational organization

This discussion has shown that interest explanations vary from simple accounts of agent's practical reasoning to evolutionary explanations

invoking non-intentional filtering mechanisms. These explanations do not have a common essence, albeit they have common elements. It is notable that they are not based on any specific sociological theory. This is quite characteristic of social studies of science in general: they do not contain much explicit sociological theory in the traditional sense of a theory.

However, there is one theoretical idea without which no discussion of interest concepts in social studies of science is complete. This exception is *the model of the cycles of credibility* first sketched by Bruno Latour and Steve Woolgar (1986: Chapter 5). Although the authors themselves are not using this model any more, it is an important theoretical background for many sociological studies of science. For example, Steven Shapin (1988: 544) writes that

> … this approach has enormous potential. It may, indeed, be the most profitable way forward for sociological explanations of scientific action and of the closure of scientific controversies.

In the following I will try to explicate this model. I will not follow closely the presentation by Latour and Woolgar, since I will add elements by other authors who have developed and discussed similar ideas. (Williams & Law 1980; Knorr-Cetina 1982; Barnes 1985; Whitley 1984; Hull 1988; Mäki 1993; Callon 1995; Turner 1996.) I will also add some material of my own.

Latour and Woolgar summarize the basic idea behind their model of cycle of credibility as follows:

> We used credibility to define the various investments made by scientists and the conversions between different aspects of the laboratory. Credibility facilitates the synthesis of economic notions (such as money, budget, and payoff) and epistemological notions (such as certitude, doubt, and proof). Moreover, it emphasises that information is costly. The cost-benefit analysis applies to the type of inscription devices to be employed, the career of scientists concerned, the decisions taken by funding agencies, as well as to the nature of data, the form of paper, the type of journal, and to readers' possible objections. The cost itself varies according to the previous investment in terms of money, time, and energy already made. The notion of credibility permits the linking of a string of concepts, such as accreditation, credentials and credit to beliefs ("credo", "credible") and to accounts ("being accountable", "counts", and "credit accounts"). This provides the observer with an homogenous view of fact construction and blurs arbitrary divisions between economic, epistemological, and psychological factors. (Latour & Woolgar 1986: 238-239.)

For Latour and Woolgar, the concept of the cycle of credibility is a way to interrelate money, scientific prestige, credentials, research problems, arguments, empirical results, publications etc. It allows them to see how issues studied separately by biographers, economists, sociologists, epistemologists, and so on, are related to each other in the everyday activities of scientists. This is achieved by the multi-facetedness of notions of credit and credibility. For example, credibility can be an attribute of an experimental result, of a scientific argument, of a scientist's CV, or of a publication forum.

An analytically oriented philosopher might ask what the point is of having such a diverse array of things under one concept. Clearly one can distinguish between different senses of the notion of credibility according to their applications. This is true, but this reply would miss the point of Latour and Woolgar's exercise. Their main objective is not a conceptual analysis, but an empirical analysis of scientists' activities in a scientific laboratory. Their justification for treating various uses of credibility as a unified phenomenon is the interrelatedness of these various uses of the concept. As they argue, it is not possible to in everyday scientific work to separate the credibility of scientific results and the credibility of the scientist producing them (Latour & Woolgar 1986: 202).

Scientists as entrepreneurs

The basis of Latour and Woolgar's model is scientists' own talk and reasoning about credibility. Latour and Woolgar observed that the scientists they studied were constantly concerned with the allocation of credit, the evaluation of credibility of empirical results, the assessment of the trustworthiness of their sources, and the scientific profitability of their career choices or avenues of new research. They also noted that scientists made sense of these issues with concepts borrowed from economic life. There was constant talk about costs, profitability, investments, risks, opportunities, and returns. Scientists even talked about marketing or selling their research products in order to earn credit. (Latour and Woolgar 1986: Chapter 5; see also Knorr-Cetina 1982: 110-112; Latour 1993: 100-129.)

According to Karin Knorr-Cetina, two distinct forms of economic reasoning can be found among laboratory scientists. The first, which reflects how research decisions are made, occurs when scientists talk about their research strategies. They describe their choices as rational decision-making, and they invoke economics concepts in the process. The different avenues of research can be evaluated in terms of the costs, risks, and possible returns they involve. The second, and more implicit, form of economic reasoning occurs in scientists' valuations. They value research problems in terms of their challenge, the value of instruments

in terms of their cost and rarity, and the worth of publication forums in terms of their reputation. In all these instances the value is determined by the contribution these items make to the value of the scientist herself. (Knorr-Cetina 1982: 111-112.)

Latour and Woolgar's idea to regard scientists as entrepreneurs is a natural extension of these observations. They combine these two ideas by taking credibility as a kind of capital or currency, and then generalizing the concept. Latour and Woolgar are not the first ones to use economic analogies in this context. For example, Pierre Bourdieu (1975) characterized science as a field of competition where scientists are described as capitalists trying to monopolize the field by accumulating scientific authority. Once in monopoly position, they attempt to define the significance research problems, the appropriateness of methods, and the importance of theories in a manner that supports their position in the field. In Bourdieu's model scientists are entrepreneurs who are trying to maximize their symbolic capital. (Bourdieu 1975; for an application of these ideas see Sapp 1987.)[8]

There is one very important difference between the model proposed by Latour and Woolgar and Bourdieu's earlier model. As Latour and Woolgar note, Bourdieu only considers one part of the market: the entrepreneurs. Completely lacking in his picture is the demand for the products these entrepreneurs produce. Bourdieu fails to consider how the value of the intellectual products is determined, and this is Latour and Woolgar's most important contribution to the discussion. They regard science as a market that is characterized by the mutual dependence of the agents in the market. The success of agents in the market depends on other agents' products and the quality of these products. This brings the epistemic issues to the core of the model.

Mutual epistemic dependence

Scientists face the following fundamental difficulty. In their scientific work they have to rely on results produced by other scientists. The simple reason for this is the fact that the information they need in their work is typically costly (both in time and resources) to produce and check. Similarly, the competence required for the production and checking of this information is typically very costly to obtain. As a consequence, there is a very extensive division of labor among scientists. This leads to a situation where scientists are forced to rely on the authority of those colleagues who possess the relevant skills to produce the needed information. Use of false or inaccurate information is often very damaging to one's research and also to one's scientific reputation. Nobody wants to base research on faulty findings. This situation creates an interest in information that is screened with respect to its epistemic quality.

As Stephen Turner (1996) suggests, scientists can be thought to be

analogous to stockbrokers. Like scientists, stockbrokers possess knowledge, competencies and training that users of their services do not possess or are very costly to acquire in terms of time and effort. This implies that reputation is very important for both scientists and stockbrokers. The reputation is a kind of shorthand for both the possession and the trustworthiness of the requisite information, abilities, and training required for the production of the desired results. This makes reputation a paramount intermediate good in science, a kind of universal currency.

Why is reputation or credibility so important for scientists? One reason is that it is an indicator of quality and relevance for experts and non-experts alike. In the case of experts, reputation provides some clear advantages. Experts are in principle able to evaluate each other's work and competence directly, without using the indirect information provided by reputation. However, as this evaluation often consumes both time and other resources, it becomes sensible to use the shortcut provided by reputation. One's reputation can also be an important resource for convincing non-experts. Furthermore, anyone's expertise is highly specialized, and outside this speciality one has to rely on others. One is no longer an expert, and then one's reputation is not only a shortcut but a necessary source of information. Since non-experts are by definition incompetent, they have to trust some authority or other for information on esoteric issues of importance.[9]

One can distinguish three components in the evaluation of the credibility of a source of information. The first concerns the source's *epistemic competence*. Has she the appropriate training and experience to make the judgment in question? Does she have a sufficient amount of background knowledge and research resources to produce trustworthy results? Here one is basically evaluating the source in the same way that instruments are evaluated with respect to their reliability in producing valid results of the desired kind. The second component of evaluation concerns the *moral character* of the source. Is she trustworthy and honest? Can we trust that she has really done the tests she claims? Does she report her results correctly? What are her motives for making the claim and how does she benefit from it? These questions touch issues that do not have analogies in the evaluation of instruments: machines cannot fail morally. The third component is less fundamental, but still important. Here the issue is whether the source is *convincing in relation to third parties*. Could her testimony be used as a resource in trying to persuade other agents, or would her testimony decrease the credibility of the argument?[10] Here the issue is with the rhetorical or argumentative usefulness of the source. This component is clearly distinct from the first one. Due to differences in the background beliefs, the sources that convince me might not convince you and *vice versa*.

This discussion shows the epistemic importance of reputation, but

> in modern academic science a scientist's reputation is also closely tied to her access to the resources for doing research. Access to publication forums, conferences, research and teaching positions, research funding, and research facilities is at least partly controlled by evaluations made by scientific peers. Funding is, in principle, awarded on the basis of scientific merit and the promise of interesting results in the future, not on the basis of nepotism or the right political connections. This makes reputation the central currency in academic life. It is a key to the resources for doing research and for career advancement. Once achieved, one's credibility can be invested in future research, and more credibility one has, the more ambitious the projects are that one is allowed to launch. This is quite far from the earlier sociological account of science that regarded recognition as a psychological reward for scientific accomplishments. The current view does not deny that public recognition can also have this kind of motivating function, but it is not the most important aspect of the scientific reputation. (Latour & Woolgar 1986: 193.)

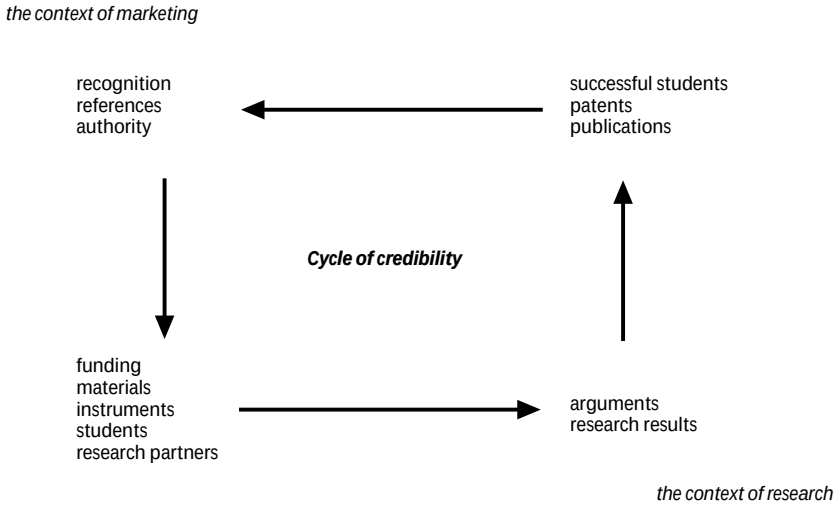
Credibility as currency and capital

A picture of a scientific capitalist begins to emerge. The scientist is an entrepreneur who produces products (for example theories, empirical results, and methods) for a market in which the principal buyers are her scientific colleagues. The key to success in this market is the relevance and the quality of products. Relevance is determined by the products' usability in the ongoing work of other scientists. The quality of the product depends on its credibility: its ability to enhance the credibility of the products the colleague intends to produce, and its success in helping the colleague to achieve the results desired. In other words, the market value of the product depends on its ability to confer value on products produced by others. If a scientist believes that her colleague's product is not credible, or she believes that it is not generally believed to be credible, she will not use it. The use of faulty results can cause huge damage to one's research, costing valuable time and other resources. In this setting, the success of a scientist depends on the use and evaluation of her products, and it is reflected in her scientific reputation, her central capital.[11]

The capital earned by earlier research must be invested in new research in order to augment it or at least to sustain its value. And here a scientist faces challenges analogous to those of a business entrepreneur. Which is the most profitable avenue of research, given the background of the scientist and the resources at her command? Should she stay with her current topic or change to a new one that promises to be more profitable in a long term? What are the risks involved in competing courses of action? How should one choose one's collaborators in a

research effort? Should one continue to work at the workbench or adopt more administrative duties as the research organization grows larger? How one should market one's results, and where and how should one publish or present results. The future success of a scientific entrepreneur depends on the choices she makes. A bad judgment or just bad luck can lead to a decrease in one's reputational capital. If one acquires a reputation for doing sloppy or defective work, or if one's products are no longer relevant to one's colleagues, one's scientific capital decreases and further options become increasingly narrow. A scientist's reputation starts to lose its value if it is not constantly renewed.

Figure 2. The cycle of credibility



This figure summarizes the most central aspects of the scientific cycle of credibility. Successful research produces results that can be converted to scientific publications and patents. Success in the context of marketing one's products leads to their use by colleagues, which in turn leads to public recognition and scientific authority conferred by the colleagues referring to one's work. This capital can in turn be converted to resources that are required for the continuation and advancement of one's research. Among such resources are materials, instruments, able students and useful research partners. This leads back to the context of research where the scientist can again try to make best of her (and her research group's) talents.[12]

A successful scientific career can be described in terms of a successful investment strategy. A young researcher starts with capital borrowed from her laboratory chief or professor. The patron provides the initial resources for doing research or lends his reputation in a form of

letters of recommendation for sources providing funding for research. The young scientist pays her patron back by working in his laboratory (for example, the results of her work are published as their joint work) or just by being successful and this way increasing her patron's scientific reputation. If successful, the young scientist can start to work on her own and if she is extremely successful, she can, in time, build up her own laboratory and research group. At this phase she has become a major scientific capitalist. The running of the research enterprise will take most of her time, and she would have to 'live' from the results produced by people working for her. The circle has closed.

Scientific reputation has a similar role as money has in economic life. It works as a generalized currency. People go to work to earn money, but rarely is having money their ultimate goal. Rather, they want money since it is a means to access whatever they happen to want. Money can be converted to (almost) everything. Similarly, one's reputation is a generalized means to achieve whatever one is after in academic science. It does not matter whether one is looking for the truth, power (in an academic community), a good salary, or the Nobel Prize: to be successful in the pursuit of this goal one has to be successful in the accumulation of scientific capital. One's goals cannot be achieved without it. (See Barnes 1985: 45-48.)

Some differences between philosophers and sociologists

What is the point of this model? It should be clear that it is not intended to be an empirical theory about psychological motives of scientists. More appropriately, it is a structural description of some central features of the social organization of modern academic science. It describes the structure of interests that characterize scientists' action environment. It is not about the motives of the scientists, but about the means by which they achieve their goals. The model cannot directly explain choices and actions by individual scientists, but it can be used explain the general patterns of behavior among scientists. It also works as a heuristic tool in the search for and construction of explanations for particular scientific episodes. The model creates a set of expectations against which more detailed explanatory hypotheses can be contrasted. In a sense, it is like 'an ideal of natural order' (Toulmin 1961) for a sociologist of science.

There is an interesting difference between philosophers of science and sociologists of science. Philosophers have a tendency to speak about credit whereas sociologists speak about credibility. This terminological difference signals a major theoretical difference. When philosophers describe scientists as looking for credit, they take it as a psychological motive. For example, David Hull speaks about 'desire for recognition', which he takes to be an unexplained starting point of his theory of sci-

ence (Hull 1988: 281). Sometimes this psychological motive is, moreover, contrasted with more 'epistemic' motives like truth or a good argument (Goldman 1999: 260-263).

In contrast, sociologists do not want to speak about psychological motives of scientists. They are more interested in the structure of situations where scientists have to act. Latour and Woolgar are careful to point out that they are not claiming that the pursuit of credibility is a psychological motivation for scientists. To the contrary, they leave personal motivations of scientists to biographers and psychologists (Latour & Woolgar 1986: 198, 207-208, 232.) Their statement that "there is no ultimate objective to scientific investment other than the continual redeployment of accumulated resources" (Latour & Woolgar 1986: 198) is often misunderstood. They are not claiming that the ultimate, and only, motive for participating in cycles of investment of scientific credibility is the desire to redeploy the accumulated resources. That is very rare motive, I guess. Scientists are not doing science just for the sake of getting public recognition from their peers. What Latour and Woolgar wish to say that there is no motive or goal that is common to all scientists. It is possible that the only thing they have in common is that they all are forced to participate in the cycle of credibility. As in business, a common currency allows people with different ultimate motives to participate in the same venture. No matter what the ultimate motive, one's success depends on one's success as an investor of capital.

Notice that if one takes 'the desire for recognition' as an ultimate psychological motive for scientists, as for example David Hull does, one has to explain why scientists want recognition, especially from their peers? Why cannot other groups provide this kind of emotional reward? This is a quite perplexing fact to be left unexplained. This explanatory challenge can be escaped by taking the sociological route and by giving up the psychological hypothesis. The mutual dependence explains why scientists want recognition from their peers independently of their ultimate personal motives.

The second difference between philosophers and sociologists is that the latter do not contrast the goal of credibility with 'proper' epistemic goals, as philosophers do. On the contrary, it is supposed to have an epistemic component. A sociologist claims that epistemic and social aspects of credibility cannot be distinguished empirically. Epistemic considerations are often built into the professional interests of scientists. As Bourdieu (1975: 21) puts it, scientists' choices are overdetermined by social and epistemic factors. (See also Biagioli 1993: 226-227.) And indeed, this sounds reasonable. After all, we are explaining choices by scientists, and as Giere (1988: 163) notes, people, including scientists, are very poor in distinguishing various factors that influence their judgments.

Two fallacies should be separated here. It is obvious that one can-

not epistemically justify a belief by referring to its positive consequences for anyone's professional interests. The good consequences do not make the belief true. The other fallacy is to infer the claim that scientists' choices are not responsive to proper epistemic concerns from the claim that their actions can be explained by referring to their professional interests. The validity of this inference is a contingent matter, and it depends on how the interests of scientists are structured. I have feeling that in the discussion of the validity of sociology of scientific knowledge these two issues are not always kept separate. However, I believe that both sociologists and philosophers can avoid both of these fallacies.

There is also a third difference between sociologists and philosophers who use the model of cycles of credibility. With the exception of Bourdieu (1975: 33), sociologists have not been interested in combining the model with the idea of an invisible hand. For philosophers (Hull 1988; Kitcher 1993) this idea has provided a way to hold on to the idea that science can produce objective knowledge despite its being inherently bound to social interests and dependent on material resources. The idea is that true or valid knowledge is an unintended consequence of the interested actions of scientists. Since I have discussed this idea earlier (Ylikoski 1995, 1998), I will not discuss it further in this context.

Possibilities for further development

There are interesting avenues for the further development of this model. I will mention here three directions. Richard Whitley (1984) takes the first by undertaking a comparative study of scientific disciplines. He shows how the way scientists depend on each other and on outsiders affects the form of knowledge their fields produce. The comparative perspective has not been widely used in social studies of science. This is a pity, since comparisons of different scientific fields are very interesting and promising sources of contrasts for explanatory questions. The comparative data would make it possible to formulate more specific explanatory questions, which in turn would advance the production of more detailed theoretical accounts.

Stephen Turner (1996) has made the interesting suggestion that the model could be used to explain both the existence of the patronage system and the existence of universities as mechanisms for organizing scientific work. These explanations can be understood as analogies to Ronald Coase's explanation for the existence of firms. Why is science so often done in universities? One explanation would be the fact that universities are useful means of enhancing the trustworthiness of scientists. The fact that the pronouncements of scientists from reputable institutions are taken more seriously by other scientists, by the general public and by the patrons of science can plausibly be accounted for in

these terms. During the relatively short life spans of their research, individual scientists face the difficulty of accumulating the trust that is necessary to compete with universities, which explains the predominance of universities in at least some relationships with patrons, the public and other scientists. Universities 'live' longer, have an interest in maintaining their reputations, and are also more able to nourish this interest effectively. Turner suggests similar explanations for the institutions of peer review. Peer-review serves scientists' interests in securing the confidence of the 'buyers' of their products by collectivizing the assurances, i.e. by making their individual claims into claims warranted by the scientific community. Similarly, the complexities of scientists' responses to the crises of trust and scientific fraud could perhaps be understood in terms of this self-regulation. Scientists in general have an interest in maintaining the image of their profession as trustworthy.

The third important direction of extension is given by the observation that scientific research is usually done with other people's money and in connection with other activities. These obvious facts are not currently reflected in the model. This gives a very internalistic picture of scientific work, and it does not apply in the same measure to all scientific research. It is mainly relevant to modern academic science, and more specifically to basic research, since it deals solely with relationships between scientists themselves, and not between scientists and other groups. Industrial and governmental research, which are nowadays the most significant employers of scientists, are beyond its scope.

The model also gives a very narrow and simplified picture of academic science itself. It is not true that academic scientists are totally insulated from the rest of the society. Secondly, research is usually connected with other academic activities. For example, when the research is done in universities, teaching and administration are activities that mix with the 'pure research' that the model describes. These aspects are not included in the model. Nonetheless, the model says something essential about the scientific enterprise as it is currently organized. I also think that there is no reason why the influence of these other factors cannot be added to the model.

In this chapter I have reviewed various forms of interest explanation. I have also characterized at some length the central explanatory resource used by sociologists of science: the model of professional interests. I expect that I have been successful in arguing that interest explanations are legitimate and interesting tools for a social scientist. Furthermore, I have been able to show the heterogeneity of interest explanations. Not all interest explanations are alike.

Notes to Chapter 6

1 The explanations discussed in this chapter have many factual presuppositions that I am not in a position to evaluate or discuss. I will concentrate on their merits as explanations, not on whether they are true explanations. I will unrealistically presume that the authors have all the relevant facts both included and right in their explanations. The reader is asked to consult original publications for the historical details of the cases discussed.

2 My view is intended to be compatible with plan-based theories of intentional action, like that advanced by Bratman 1987.

3 Later MacKenzie is quite explicit in stating that Pearson's eugenical commitments do not explain everything about his statistical work but that they explain this particular difference between his and Yule's work (MacKenzie 1999: 226-227). Alan Chalmers (1990: 96-104) criticizes Mackenzie for failing to show that social interests can influence the content of good science. According to Chalmers, MacKenzie only succeeds in showing that social interests were present in Pearson's (and others') intentions, but not in statistics itself. According to him, the fact that mathematical statistics was useful for eugenists and hence for professional middle-class interests, does not show that there is an intrinsic connection between the two. Mathematical statistics could also have been used to serve the interests of some other social group. I take this to be a misunderstanding of MacKenzie's aims. He was not trying to show that the social interests that informed the creators of these statistical techniques and methods would remain with these techniques and methods for good. Certainly he was not trying to provide any rigorous generalizations about the relationship between the techniques and certain social interests. He was simply trying to show that social interests had a significant role in the genesis of 'valid' scientific results. More generally, the position Chalmers attributes to MacKenzie has never been a part of the Strong Program, nor of the sociology of scientific knowledge movement in general. Apart from this misunderstanding, there is not much disagreement between Chalmers and MacKenzie concerning the interpretation and explanation of this historical episode.

4 Pickering later became more critical of interest explanations. In his 1995 book, The Mangle of Practice, he raises a couple of criticisms. First, he claims that interest theorists (like David Bloor) perceive social factors and interests as fixed, rather than constantly changing. According to him social interests cannot be taken as 'unmoved movers'. Second, he claims that interest theory is not able (or theoreticians supporting it are not willing) to account for the changes in interests and to account for the emergence of these interests in the first place (Pickering 1995: 63-65; 151-152). I discussed these claims in the previous chapter.

5 Pickering's model is certainly underdeveloped. It takes into account neither the credibility of the experimental results nor the reputation of the scientists producing them. After all, scientists do not work on any theory/experiment simply because it suits their resources. They choose only those that are credible in the eyes of their colleagues. Only in this way they can advance their professional aims and interests.

6 Schmaus 1992 develops a very similar account of scientific niche formation. For an application of Gilbert's account of competition in science, see Edge 1990.

7 MacKenzie does not claim that he can explain why Galton or Pearson became supporters of the eugenics movement. Although their social and class background makes this choice understandable, it does not make it necessary. There were other people with similar backgrounds who did not support eugenics.

8 Segerstråle 2000 presents an interesting extension of Bourdieu's ideas. She takes moral recognition as a form of symbolic capital, and then regards the main participants in the sociobiology controversy in the 1970's as entrepreneurs trying to maximize their moral capital (Segerstråle 2000: Chapter 15).

9 There might be interesting differences between various audiences (colleagues, funders, the general public) and the ways in which a reputation is created, maintained and evaluated.

10 For a historical account of the practices of evaluating scientific competence, see Rudwick 1985: 418-428 and Shapin 1994. Shapin 1995a is a review of the relevant sociological literature. For philosophical accounts of testimony, see Hardwig 1985, 1991 and Coady 1992.

11 Considering how much this model borrows from economics, it is surprising that economists have not developed similar models. For a review of traditional work on economics of science, see Diamond 1996 and Stephan 1996. For an exception see Earl 1983.

12 This model presupposes that two important norms are in force in a scientific community. The first norm demands that scientists cite the authors they use in their work, thus giving the credit, or the blame, to the ones who have done the work. The other norm requires that the credit be given to the person (or group) who presents the result for the first time, thus forcing scientists to produce new knowledge in order to gain credit.

Chapter 7
Interest explanation compared

This final chapter aims to further clarify the nature of interest explanation by comparing the interest approach with rational choice theory (hereafter RCT). I will also discuss Jon Elster's critique of functional explanation in the social sciences. This comparison will shed some light to the legitimacy and scope of interest explanations.

In the first section, I will briefly compare the interest approach with RCT and argue that these two approaches have some important differences in their methodologies and their assumptions about the cognitive capacities of an agent. The second section will take a more detailed look at two interpretations of RCT. Both interpretations suggest that RCT does not need to be interpreted as a description of agents' practical reasoning. These suggestions are worth discussing for two reasons. First, they bring RCT closer to the interest approach as I have reconstructed it. Second, these ideas are important in clarifying the ways in which interest explanations can be something else than simple intentional explanations.

The first suggestion, made by Philip Pettit, views interests as standby causes. I will argue that Pettit's 'rational interest theory' makes stronger motivational assumptions than the standard interest approach. Furthermore, I argue that Pettit's valuable idea of resilience as an *explanandum* can be separated from his thesis about the relative importance of self-regarding motives in human action. The second suggestion, by Debra Satz and John Ferejohn, proposes an externalist interpretation of RCT. In this interpretation, preferences are understood as descriptions of the structures of interaction rather than subjective mental states of individual agents. The externalist interpretation is interesting since it seems to fit quite nicely with standard explanatory practice in the RCT tradition. Secondly, it describes RCT as structural, non-individualistic, social theory.

In the final section I will discuss Jon Elster's critique of functional

explanation in the social sciences. This discussion is relevant since Elster's critique threatens the legitimacy of the pattern of non-intentional filtering explanations sketched in Chapter 6. I demonstrate that Elster's critique is flawed in a number of ways, and I therefore maintain that explanations referring to non-intentional filtering mechanisms are indeed legitimate causal explanations in the social sciences.

1. The interest approach and rational choice theory

Some commentators have noted similarities between RCT and the interest approach (Yearley 1984: 72). Both theories describe agents as rational pursuers of their goals, and consequently they both employ an intentional pattern of explanation. This sets them apart from some popular approaches in the social sciences. Furthermore, they do not intend to describe agent's reasoning in the extremely detailed and faithful manner characteristic of more hermeneutically oriented approaches. Instead, these approaches abstract away from some details and with a help of rationality assumptions try to give only the essential elements of agents' practical reasoning. Both are also more interested in typical agents, rather than in displaying idiosyncrasies of individual agents. In sum, compared with other approaches in social sciences, the interest approach and RCT seem quite similar. However, when these approaches are compared with each other, some very significant differences stand out. In this section I will concentrate on these differences.

The difficulty in discussing RCT is that it is a theory only in a very broad sense. In this sense it is very similar to 'the interest approach'. It can take various forms, depending on the specific application and field. Rational choice theories employed by economists, philosophers and historical sociologists differ greatly from each other. There is also much variation in opinions concerning the nature of rational choice approach.[1] For some, RCT equals the formal tools of neoclassical economics, and for others it does not seem to include much more than a commitment to the central explanatory role of intentional psychology (with some minimal rationality assumptions) in the social sciences. A comprehensive classification and discussion of these various definitions of RCT would require an extensive study that cannot be attempted here. Instead, I will first make some general observations, and pick up a couple of suggestions for interpreting rational choice explanations. Them I will see whether they can be helpful in making sense of interest explanations.

There are two very clear differences between the interest approach and RCT. The most obvious is the fact that the interest approach in the social studies of science does not take formal decision theory as its model of rationality. There are no references to the decision theoretical literature, or to concepts like preference, utility function and subjective

probability. Furthermore, there is no indication of a commitment to an assumption that agents' preferences are complete. Similarly, the reasoning abilities of agents are not assumed to reflect normative theory of decision. For example, scientists are not assumed to be Bayesian agents. Finally, although interest theorists often describe agents in strategic situations, they do not help themselves with the common knowledge assumptions that have a central role in game theory.

When compared with decision theory, the sociological conception of rationality is modest in its assumptions. Barry Barnes calls it natural rationality (Barnes 1974; 1976). This is a purely descriptive account of human reasoning processes. It takes agents' habits of belief formation and their methods of decision-making to be whatever empirical psychological research shows them to be. Normative theory of rationality does not have any role in this approach to human reasoning. This is the same standpoint as that typically taken by naturalistic philosophers of science, like Ronald Giere (1988: Chapter 6), and empirically minded psychologists of science (Faust 1984). Agent's reasoning is taken as 'natural cognitive activity' that is directed toward reaching specified goals (Giere 1988: 161).[2] This is in sharp contrast with RCT, which takes the normative decision theory as its starting point.

There is not much discussion about rationality in the sociology of scientific knowledge literature. This can be interpreted (charitably) as an open attitude towards the future results of research in cognitive psychology. Sociologists are ready, in principle, to incorporate any findings of cognitive psychology into their account. The problem is that cognitive psychology has not provided many undisputed results that challenge the existing cognitive assumptions of the interest approach. While cognitive psychology matures, sociologists seem to be relying on common sense psychology as the main source of their cognitive assumptions. For their applications, this common sense account has served sufficiently well.

It seems to me that the model of bounded rationality (Simon 1982; Conlisk 1996) is the theory of rationality sociologists would choose from the existing alternatives if they were forced to choose. Interpreting scientists as satisfiers would still make sense of most interest attributions. As some RCT theorists are also building on the premises of the theory of bounded rationality, there is room for more substantial overlapping between these two approaches.

The second major difference between RCT and the interest approach is methodological. Typically, rational choice theorists in economics and in political science proceed by building idealized and simplified models of their objects of study by making rationality assumptions. A good example is the problem of voter turnout in political science. The rational choice theorist starts from some premises that seem plausible in the light of her theory, and derives a prediction that the

voter turnout in normal political elections would be very low. As this prediction is falsified by empirical data, the rational choice theorist turns back to her assumptions and tries to revise them in manner that would resolve the huge discrepancy between theoretical predictions and empirical observations. A political scientist using RCT is not typically interested in any specific elections or in any specific political system. She is dealing with the general problem of voter turnout. Consequently, she does not predict the turnout in any specific elections, or explain differences in turnout between different elections. She deals with highly 'stylized' *explananda*.

This theory-driven methodology is absent in sociology of scientific knowledge. The preferred methodology for a sociologist of knowledge is to make empirical case studies based on either contemporary or historical material. In these studies detailed description of characteristics of one particular case are highly valued. Rather than trying to find something in common with various cases, the historical approach emphasizes the differences between the cases. The *explananda* are concrete and historically specific. The interest approach works more as an interpretive scheme than as a source of theoretical predictions. Consequently, sociologists of knowledge do not show much interest in explicit model building, they seem to prefer more informal characterizations of their explanations and theories.

Critiques of interest explanations

In the light of these two important differences, the very common criticism of interest explanations that they inherit the all problems of *homo economicus* does not seem to be entirely correct. First, there is nothing in interest accounts that commits their users to take scientists as maximizing subjective utilities in the sense of decision theory. The agents can as well be considered to be satisfying agents in Herbert Simon's sense. But not even this is necessary, since the interest perspective does not commit its supporter to any particular model of human decision making. The assumption of natural rationality as a purely descriptive account of human reasoning processes leaves a door open for various theories of human reasoning. Secondly, interest explanations do not make ambitious assumptions about an agent's knowledge of her action environment and other agents' preferences and beliefs. More specifically, these explanations assume neither complete nor common knowledge. Thirdly, there are some very clear methodological differences between the approaches. This has the consequence that, for example, nearly all criticisms presented by Green and Shapiro (1994) against RCT do not apply to the interest approach.

This conclusion should not be taken as a completely positive result. The interest approach escapes most problems of the standard RCT,

but it also misses some of its greatest merits. Theoretical explicitness, mathematical sophistication and other advantages of formal modeling are missing from it. These virtues would certainly help both in the development and in the evaluation of sociological accounts of science. Common sense folk psychology is safer in its assumptions, but also more vague in its results.

It is not my intention to claim that interest explanations are always faultless. Although the interest approach is much weaker than RCT in its cognitive assumptions, it still makes substantial assumptions about what I call *interest cognition*. The difference is that there are some theoretical problems in regard to cognitive assumptions in RCT, but in the case of the interest approach the problems concern only the applicability of the scheme to particular cases.

Interest explanation can make excessively strong assumptions about interest cognition in various ways. To begin with, there are failures in assumptions about agent's cognitive processes and capacities. First, there is *failure by a competence assumption*. An interest explanation can be too generous in attributing reasoning capabilities to an agent. It might be that some agents are less competent in processes of inference and memory than the explanation assumes. It might also be that an interest theoretician is too generous in her competence attributions in general. Clearly this is a failure, since a causal explanation cannot assume capacities that are not realized in the case in hand. The fact that interest explanations derive their assumptions about human cognitive abilities from an educated common sense is not a perfect safeguard against this failure. It might be that our common sense is too optimistic about our cognitive capacities in general, or that these common sense assumptions just do not work is some particular cases.

Secondly, there is *failure by an opportunity assumption*. A theoretician might be too optimistic in her assumptions about the performance of cognitive operations. It might be that the agent has not had an opportunity to perform the cognitive operations the theoretical account assumes, although she is capable of performing those operations in favorable conditions. For example, there might be some interfering factors that the theorist has not taken into account. Again, a true causal explanation cannot be built on processes that have not taken place.

An explanation can also fail by its assumptions about the goals of an agent. One way to failure is *attribution of excessively determinate goals to an agent*. It might be that the agent is not as clear about her goals as the theorist assumes. In some cases it might even be wrong to attribute any goals at all to the agent. Similarly, the theorist might disregard the complexity of agent's goals. This might be called *a failure by excessively drastic simplification of the goal structure*. The fact that an agent can have a number of conflicting goals can in some cases make the whole idea of providing an interest explanation implausible. The third source

of difficulty concerns the dynamics of goal formation, that is the emergence of an agent's goals and changes in them. Here we have *failure by assuming excessively stable (or elastic) goals*. The theorist must allow that the agent can change her goals and adopt new goals, but it is often difficult to find evidence for these changes. Suppose that there is evidence that the agent has a set of goals g_1 at time t_1, and that she has changed her set of goals to g_2 at time t_n. The difficulty concerns the attribution of goals between t_1 and t_n. Did she suddenly change her goals or was there a slow process of reconsideration? It is very difficult to say in the absence of evidence. An intentional explanation that refers to nonexistent goals or that characterizes the agent's goals wrongly cannot be the correct explanation. Since the attribution of goals is an interpretive process that is strongly influenced by the interpreter's background beliefs, these three failures are constant threats to the validity of interest explanation.

The third group of possible failures concerns assumptions of what the agent knows. An explanation can fail by assuming that the agent knows more about her environment than she actually does. It is normally reasonable to assume that agents are knowledgeable about their action environment, but how knowledgeable? Problems are also created by the possibility that the agent has false beliefs. What are the criteria for attributing false beliefs? An interest theorist proceeds by making assumptions about the agent's knowledge she considers reasonable, but common sense is not an infallible guide. People can be wrong about their action situations and about the goals and beliefs of other agents, and quite often they are.

These observations show that the construction of an interest explanation is not a foolproof procedure. The explanation can easily fail owing to its assumptions. However, I want to emphasize that these problems are empirical. They concern particular empirical applications, not the possibility, in principle, of providing true and sensible interest explanations. Furthermore, these problems are not unique for the interest approach. They are the same problems that every intentional explanation faces.

Despite these differences between the interest approach and RCT, the evaluation by Green and Shapiro of the applicability of RCT applies only with small modifications to the interest approach as well:

> Rational choice explanations should be expected, *prima facie*, to perform well to the extent that the following five conditions are met: (i) the stakes are high and the players are self-conscious optimizers; (ii) preferences are well ordered and relatively fixed (which in turn may require actors to be individuals or homogeneous corporate agents); (iii) actors are presented with a clear range of options and little opportunity for strategic innovation; (iv) the strategic complexity of the situation is

> not overwhelmingly great for the actors, nor are there significant differences in their strategic capacities, and (v) the actors have the capacity to learn from feedback in the environment and adapt (Green & Shapiro 1995: 267).

In this account, RCT gives up its universalistic ambitions. It applies only in some situations, not in all. I think the same can be said about the interest approach. There are some situations where it applies as an explanation scheme, and some other situations where it does not apply at all. The crucial question for a sociologist of science is how many aspects of science belong to the former category.

2. Two ways to expand the scope of explanation

The interest approach and RCT face similar challenges. One of these concerns the theory's explanatory power when it does not describe the agents' actual reasoning. If an explanation accurately describes the actual reasoning of an agent, then it fits to the intentional pattern of explanation nicely. Sometimes both approaches provide ordinary intentional explanations of this kind. However, although it is often denied that the explanation describes the agent's actual reasoning, this failure is not treated as a fatal blow to the explanatory power of that particular account. Indeed, the special advantage of these approaches is often claimed to be their ability to abstract away from the details of the mental lives of individual agents. The tricky question is to explain how this is possible. How can a causal explanation be explanatory when it does not seem to state the facts as they are?

Consider for example the model of professional interest discussed in Chapter 6. It does not claim that scientists are motivated mainly by their wish to enhance their credibility. Nor does it claim that scientists deliberate their actions in terms of credibility. Credibility is a theoretical concept developed by sociologists who study scientists. It does not belong to the native vocabulary of scientists; nevertheless the model is considered to be explanatory. How does this kind of explanation work? Since sociologists are not very informative on the issue, I will turn to the literature on RCT explanations for guidance.

Here, I want to consider two suggestions that have been made in order to deal with this problem in the case of RCT. The first suggestion, made by Philip Pettit, builds on the idea of standby causes. In this interpretation interests are 'virtually' present in the agent's reasoning as standby causes. The suggestion made by Debra Satz and John Ferejohn gives up these psychological assumptions. In their externalist interpretation, RCT describes the interests of agents, not their mental life. The central explanatory work in their model is done by a selection mechanism that guarantees that only agents who act *as if* they were RCT agents are present.

Lewis and Pettit on explaining resilience

Philip Pettit's (1993a; 1995a; 2000) account starts with the observation that the *explanandum* in RCT explanations is often the *resilience* of a phenomenon or a pattern of behavior. Pettit claims that in such cases rationality considerations can work as *standby causes*. By this term he refers to the idea that these causes might not have any role in the actual causal history of the *explanandum* event; however if there were some factors that would threaten the regularity, these causes would work to the effect that the regularity would remain stable. (Pettit 1995a: 323.)

How does this idea work? Pettit draws his idea from David Lewis' (1969) analysis of conventions. Lewis explains the existence of (some) conventions by the fact that they often serve to resolve problems of coordination. Lewis used a game-theoretical apparatus in the exposition of his ideas, and this raised the question: how is the reference to considerations of rationality in these game-theoretical models to be understood? According to Pettit, Lewis is not explaining the historical emergence of these conventions, nor is he providing us with an account of the reasoning that sustains these conventions through time. Lewis admits that agents might not be aware of the coordination problem their convention solves and that they may stick to their behavior for a variety of reasons. These reasons might include ideology and a sheer inertia of routine, among others. (Pettit 1995a: 326.) Lewis states his position in the following terms:

> An action may be rational, and may be explained by the agent's beliefs and desires, even though that action was done by habit, and the agent gave no thought to the beliefs or desires which were his reasons for action. If that habit ever ceased to serve the agent's desires according to his beliefs, it would at once be overridden and corrected by conscious reasoning. Action done by a habit of this sort is both habitual and rational (Lewis 1983: 181).

In Pettit's interpretation, Lewis indicates here that he is explaining the resilience of the conventions. The game-theoretical reconstruction does not explain by describing the actual practical reasoning by the agents, but by giving a structural description of their action situation that gives rise to their preferences.

Pettit builds his own account by expanding this idea. For Pettit, 'the economic mind' is only virtually present in agents' actions. Ordinarily, agents in non-market contexts routinely act based on a cultural framing of the situation. In such a frame of action the agents do not explicitly consider whether their actions are in accordance with their interests. In a sense, the agents proceed under a more or less automatic cultural auto-pilot. They proceed according to a culturally salient frame or

script without explicitly deliberating the conduciveness of that behavior to their ultimate goals. According to Pettit, the virtual presence of an economic mind is displayed in the fact that the behavior by the agents within the standard cultural frame is usually in accordance with their interests, at least approximately. How is this possible? Pettit suggests that there are in the background some 'alarm bells' that 'wake' the agent to consider her choices explicitly if her interests are in danger. These alarms remain silent in most situations where 'the cultural auto-pilot' is on. The idea is that these alarms start to work in situations where an agent's interests might be compromised, and direct the agent to switch to 'manual steering'. The agents start to consider their choices explicitly in terms of interests only in these alarming situations. (Pettit 1995a: 319-320.)

Pettit stresses that the alarms in his model have to be informational. They have to be signals that notify the agent that her advantage may be compromised if the action continues to proceed according to the salient cultural frame. According to Pettit, the following are examples of alarming situations:

1) the agent's decision situation is non-routine;
2) the agent has already had her fingers burned;
3) the peers of the agent (her reference group) are doing significantly better than the agent; and
4) some conventional or other assurances as to the responses of others are lacking.

States of affairs like these serve as signals that the agent's interests can be in danger, which leads the agent to explicitly consider her options. (Pettit 1995a: 321.)

Pettit illustrates his notion of resilience explanation by an analogy. He asks us to imagine a set-up in which a ball rolls along a straight line, but there are little posts on either side that are designed to protect the ball from the influence of various possible but non-actualized forces that might cause it to change its course. The posts are able to inhibit incoming forces and if nevertheless have an effect on the ball, the posts can restore the ball to its original path. Now, according to Pettit, the posts might be outside the complete causal story describing the history of the movements of a ball. The ball never touches the posts as it rolls down. In such case the posts do not have any role in the explanation of actual details of a ball's movement. However, they can have an explanatory role if the *explanandum* is the *resilience* of the ball's path. Thanks to the presence of these posts, the fact that a ball rolls in a straight line is robust under various contingencies, and it can be relied upon to persist. As a result, the posts do have a clear role in the explanation if the *explanandum* is this modal persistence of a straight line. (Pettit 1995a: 324-325.)

The analogy works as follows. As the reference to the 'virtually' efficacious posts explains the resilience of a ball's rolling on a straight line, so a reference to the 'virtually' present interest considerations may explain the resilience with which people maintain certain patterns of behavior. The actual history of a pattern of human behavior might be described in terms of the sheer inertia of people's routine way acting within a certain cultural framework. However, this pattern might have the modal property of being robust in various contingencies. An example of such a contingency might be that some people deviate and offer an alternative way of acting. Now, the factors that are needed to account for the actual history of a pattern might not be able to account for this modal robustness of the pattern. There might not be any reason why people could not display an alternative cultural framework. Pettit suggests that one possible explanation for this robustness is the virtual presence of interest considerations. If the contingencies were to produce a different pattern of behavior, the alarm bells of self-interest would ring, and the considerations prompted by interests would lead people back to their original behavioral patterns. The practical reasoning in terms of interests would have caused the behavior had it not been pre-empted by routines. (Pettit 1995a: 325.)

Pettit assumes that motives of *homo economicus* have to be self-regarding or selfish, at least virtually. More specifically, he claims that economists using RCT assume that self-regarding motives are generally stronger than other-regarding ones. (Pettit 1995a, 2000: 36-40.) Some of Pettit's critics have replied that a rational choice theorist need not make an assumption of this kind (for example, Tuomela 1994). The idea is not essential to the rational choice paradigm. It can be admitted that quite a few supporters of RCT make this kind of empirical assumption, but it is not essential for the rational choice approach. It is more like an optional extra assumption. In reply to such comments, Pettit has later specified his intentions. He now says that his discussion of explanation in terms of what he calls *rational interest theory*, which, by definition, is about rational pursuit of self-regarding concerns (oral communication, 12 July 2000, Rotterdam).

Evaluating Pettit's proposal

Should we assume with Pettit that *interests* are always self-regarding? I think we should not. Interests derive from persistent goals of an agent, whatever these goals may be. Why should interests be attached only to selfish goals? I cannot see any motivation for a conceptual limitation of this kind. As I argued in Chapter 5, the starting point of the interest approach is the agent's goals, but those goals do not have to be self-regarding. Similarly, we need not to make any *a priori* assumptions about the strength of the various motives of an agent. Why should we assume

that self-regarding motives are always stronger or overriding? Again, this seems to be a separate assumption that can be kept separate from the interest perspective proper.

I am happy to notice that I am not alone in opposing Pettit's constraints for interests. To an objection that truthfulness cannot be convention because truthfulness is a moral requirement that is independent of any interest, David Lewis has the following response:

> The objection plays on a narrow sense of "interest" in which only selfish interests count. We commonly adopt a wider sense. We count also altruistic interests and interests springing from one's recognition of obligations. It is this wider sense that should be understood in the definition of convention. In this wider sense, it is nonsense to think of an obligation as outweighing one's interests. Rather, the obligation provides one interest which may outweigh the other interests (Lewis 1983: 184).

I share this wider account of interests with Lewis. It is clear that scientists' professional interests are self-regarding in Pettit's sense. However, I do not see why one should consider the explanatory use of interests that are not self-regarding to be illegitimate. This is especially important since there are number of motives that are not easily characterized either as self-regarding or as other-regarding. For example, is the desire to make an important scientific discovery selfish or non-selfish? Fortunately, we do not have to answer questions like this. The possible ultimate selfishness of agents' interests and the form of interest explanations are clearly separate issues. Pettit's idea of interests as standby causes can be considered independently from his theory of 'rational interests'. Similarly, the idea of resilience as an *explanandum* is conceivable without it.

Does Pettit's theory provide any insight to the nature of interest explanations in social studies of science? I think he at least provides some interesting clues. Let us start with the idea of resilience. It is clear that sociologists and historians sometimes endeavor to explain the stability of some institutions and patterns of behavior. These *explananda* fit Pettit's description of resilience explanations. Similarly, the idea of interests as standby causes is useful in these cases. It is not plausible that agents constantly have in their minds the interests served by their actions. Clearly, interests figure in their practical reasoning only in times of challenge and novelty. It is only in such situations that they will consider the point of their routines and the possible advantages of alternative avenues of action.

Similarly, considering how Barnes and Bloor understand the role of goals and interests in rule-following, the idea of routine action being backed up with interests has some appeal. It is not reasonable to assume that agents constantly reflect their interests when applying con-

cepts. In fact this is not just unreasonable, but impossible. If applying concepts were always conscious decision making, it would lead to an infinite regress. For this reason, the application of concepts has to be ultimately based on non-intentional habits, not conscious reasoning (Pettit 1993a: 58-60). Goals and interests are explicitly considered only on special occasions. Examples of such occasions are situations where the agents have problems categorizing anomalous instances and settings where there are interpersonal inconsistencies in the use of a concept.

Pettit's idea seems to make sense about explanations in terms of interest in credibility. It is not plausible that scientists would consider all their actions in terms of their contribution to the accumulation of scientific credibility. It is more plausible to think that these considerations come up only in special circumstances. Recall Pettit's examples of alarming situations. Certainly scientists face non-routine situations where they are prone to consider their choices in terms of their contribution to their scientific career. Such situations include such crossroads as applying for a new appointment or planning a new research project. Similarly scientists can compare their success to the success of their colleagues and observe whether they are doing exceptionally badly. An analogy to burning one's fingers is also possible. Not getting funding, troubles getting one's research published, or complete ignorance by colleagues can work as such alarms. These alarms are certainly such that they bring considerations of credibility explicitly into a scientist's deliberations.

Pettit's model has considerable intuitive plausibility. There is only one aspect that troubles me. Pettit's idea is to come up with a plausible interpretation of *homo economicus* as a general thesis about human motivation. In this sense it is a thesis about human nature. It aims to say something about all human beings. Now, if we transfer the idea to the professional interest of scientists, we come up with a thesis of a common motivation for all scientists. It would say that all scientists are motivated by the interest to accumulate scientific credibility. This is a respectable empirical hypothesis, but I doubt that it is true. People end up doing scientific research for various reasons and their motives are as varying. Not everybody is concerned with their professional existence or career even in such alarming situations as those described above. Some people are ready to commit what may be called professional suicide by sticking with topics that they find interesting or important. There are also people who use their personal wealth to support their research and who are not interested in evaluations by their colleagues. Not all scientists are scientific capitalists, not even virtually. Nor it is plausible to think that this account would be about *the rational interests* of scientists. That would make rationality a categorical property of some motivations. I cannot see why a person who just wants to

do research without any academic monkey business could not be rational in choosing to disregard all considerations of professional interests. It is wiser to abstain from interpreting the model of cycles of credibility as a thesis concerning the ultimate motives of the scientists.

These observations suggest that Pettit's ideas can be useful in making sense of explanations in social studies of science. How well do Pettit's ideas fit with my theory of explanation? There are no special problems here either. For example, standby causes are really one example of preempted causes. As I argued in Chapter 2, the counterfactual theory of causal explanation can easily accommodate such cases. However, the idea of a standby cause is not relevant in many common explanatory uses of interests. Sociologists often try to explain choices by scientists in non-routine situations, or in situations where the routine pattern of action is changing. These situations are often considered especially interesting from the point of view of sociological analysis. In such cases, interests cannot be considered as standby causes. If they are to have any explanatory import, they should figure in an agents' reasoning. Otherwise the relationship between interests and choices would be purely accidental.

Pettit's ideas can be strengthened with the idea of subgoals developed in Chapter 5. The subgoals are devices to keep one's ultimate goals from complicating one's practical reasoning. They simplify one's practical reasoning and make it possible that all one's goals and interests are not constantly in one's mind. In this way the agent does not have to consider every choice she makes in terms of all of its consequences. She can concentrate on a limited number of issues at a time and leave more general goals in the background. In such cases interest considerations can explain both the origin and the continuity of particular routines. All these points help to make Pettit's reference to routines and non-reflective behavior more acceptable.

I would like to make one more critical observation about Pettit's discussion. In explicating his idea of *resilience* as an *explanandum*, Pettit contrasts it with two other *explananda*. These are *emergence* and *continuation* of a pattern of behavior or regularity. (Pettit 1995a: 324.) I do not find anything wrong with historical emergence as an *explanandum*. Sometimes we want to know how a given social pattern of behavior started. However, I have my doubts about whether continuity is a sensible *explanandum*. As Pettit describes them, these explanations amount to a description of the (causal) history of a behavioral pattern. There is nothing wrong with the idea of providing a detailed description of a historical process. But what kind of explanation would this description be, what is its *explanandum*? For sure, individual parts of that story are explanatory when the *explanandum* is some particular event in this history. But which explanation-seeking question would the whole history of the pattern answer? To my mind, a simple chronicle

like this is not an explanation, but a store of explanatory information. I conclude that there is no such *explanandum* as the simple continuity of a process. The *explanandum* social scientists have contrasted with the emergence has always been more like Pettit's resilience than pure continuity. Consequently, Pettit has not found a third kind of *explanandum* for social scientists, as he seems to suggest (Pettit 1993a: 276). However, he has helped to explicate an already existing one.

The externalist interpretation of rational choice theory

Debra Satz and John Ferejohn (1994) distinguish between two interpretations of RCT. In *the internalist interpretation*, RCT is seen as describing what actually goes on inside the minds of agents. In this interpretation, mental entities involved in a rational choice account, preferences and beliefs, describe real psychological states that are causally responsible for agents' choice behavior (Satz and Ferejohn 1994: 73). The trouble with this interpretation is obvious. As Satz and Ferejohn point out, RCT is too simplistic to be able to describe complex psychological processes realistically. Various empirical studies show that decision theory is not an accurate description our mental life. Based on this evidence they conclude that RCT fails as a psychological theory. (Satz and Ferejohn 1994: 73-74; Clark 1997; see also Kahneman, Slovic & Tversky 1982.)

In order to justify the use of RCT in the social sciences, especially in economics, Satz and Ferejohn propose an alternative interpretation. They claim that this *externalist interpretation*, as they call it, does justice to the most central uses of RCT in the social sciences. The basic idea is that a rational choice theorist usually aims to illuminate *structures* of social interaction in markets or in institutions. She is not interested in explaining the behavior of any particular agent, but general regularities governing the behavior of all, or most, agents. In such applications the burden of the explanation shifts from the psychologies of individual agents to the environmental constraints they face. (Satz and Ferejohn 1994: 74.)[3]

Satz and Ferejohn distinguish between two versions of the externalist interpretation. They call the first one *radical* externalism. This position denies the existence and the causal efficacy of mental entities. In this account, preferences are interpreted behavioristically as choices. For Satz and Ferejohn this position is too strong in its extreme instrumentalism. They support a *moderate* form of externalism. This account shares with radical externalism the deflationist interpretation of formal rationality. In this interpretation all that formal rationality entails is that an agent's action be explicable *as if* she were maximizing preferences. It simply consists of the claim that human action is 'consistent' with the hypothesis of goal-seeking behavior. (Satz and

Ferejohn 1994: 74-75.)

The difference between radical and moderate externalism is that the latter accepts that mental entities exist and that they are causally efficacious. It differs from internalism by claiming that detailed accounts of these entities do not figure in the best rational choice explanations of human action. According to Satz and Ferejohn, the user of RCT does not need to commit herself to any particular theory of internal structure of intentional agency. Similarly, she is not committed to any substantial psychological theses about the desires or motives of the agent. This is in sharp contrast with Pettit's approach. For Satz and Ferejohn the most persuasive uses of RCT do not interpret it as describing human psychology. For example, in economics correspondences of supply and demand or the existence of competitive *equilibria* and their characteristics do not depend on any specific psychological make-up of the agents. The only required hypothesis is that agents' behavior can be understood as *as if* it was maximizing according to some suitably restricted utility function. This makes the relation of RCT to psychology very remote. (Satz and Ferejohn 1994: 76.)

For Satz and Ferejohn the most interesting *explananda* of RCT are structural. Typically, RCT is used to explain the relative stability of certain patterns of behavior. For example, neoclassical economics is interested in the existence of equilibria and the relations of these equilibria to various descriptive parameters of the economy. The hypothesis that individual behavior satisfies the externalist coherence tests is sufficient for this purpose. There is no need to go more deeply to the psychologies of the market agents. Typical explanatory statements in economics relate the changes in some costs to the choices made by agents. The observable changes of parameters are related to the changes in behavior patterns without making any specific psychological hypotheses about the agents. An equilibrium explanation does not point to the actual cause of any particular agent's action, but it describes the causal structure that is essential in accounting for the aggregate effects of individual behaviors. In this interpretation, the point of the theory is not to show that agents are rational or wealth-seekers, but to understand how shifts in parametric constraints affect the behavior at the aggregate level. The explanatory work is done by these constraints, not by hypotheses about the beliefs and preferences of individual agents. (Satz and Ferejohn 1994: 77-78.)

Satz and Ferejohn claim that in this interpretation of RCT the preferences are attributed to the individuals on the basis of their position within a surrounding structure. Instead of trying to look inside people's minds, the theorist reads their preferences from their external action situation. This looks very much like the interest approach, and Satz and Ferejohn notice this (Satz and Ferejohn 1994: 78-79). As an example of a 'structural theory of interests' they mention the neoclassical theory of

the firm. The neoclassical assumption that a firm acts to maximize its profits does not apply to the firm by virtue of any psychological assumptions about its managers. Instead, the environment of competitive capital markets acts as a selector of firms. It ensures that only firms that act to maximize their profits survive the competition. The firm's interests arise from demands of their respective environments and they have the central explanatory role, not the psychological make-up of the management of the firms. The competitive environment of the market externally determines the types of action patterns that will lead to survival and success. In this account, the hypothesis of profit maximization is compatible with various models of psychological processes. The 'preferences' describe the structure of the situation, not the psychological characteristics of agents. The psychological motives of the decision-makers in the firm can be whatever, as long as they are compatible with the assumption of profit maximization. This allows the same model to be applied in various historical and cultural settings. (Satz and Ferejohn 1994: 79.)

Satz and Ferejohn note some limitations of the applicability of the externalist interpretation. These limitations can be seen by comparing a theory of voting behavior with a theory of electoral competition. In a theory of electoral competition in a plurality-rule system, the results can be obtained in a theoretical manner. The theory gives strong predictions as to what the parties will do in equilibrium. The competition among parties for offices encourages them to be electorally motivated, since the non-electorally motivated parties will tend not to be elected and would thus be unable to reward their supporters. In the case of voters the situation is the opposite. The way an individual voter casts her vote has very little effect on her life, at least in major elections. Consequently, there are no strong constraints on how people will choose their candidate. Furthermore, there are no competitive forces that would shape the preferences of the individual citizens. This makes the theory of voting behavior very weak in terms of explanation and prediction. (Satz and Ferejohn 1994: 79-80.)

This observation leads Satz and Ferejohn to conclude that RCT is most credible under conditions where the choices are severely constrained. If strong constraints are lacking, the agents will not generally behave according to the predictions of rational choice theory. They also note that the social analogies of natural selection play a central role in the modest externalist account of rational choice. The selection mechanism ensures that the agents behave according to the theory, and if we cannot assume such a mechanism, we cannot use the externalist interpretation. In such cases the internalist interpretation is the only game in town. (Satz and Ferejohn 1994: 81.)

Some critical comments

Unfortunately Satz and Ferejohn do not further discuss the nature of the social selection processes presupposed by their interpretation of RCT. If RCT is interpreted as they suggest, the explanatory work is done mainly by these evolutionary mechanisms. RCT in itself does not say anything about these selection mechanisms and processes, they are just assumed to be in place. This fact leads to a situation where RCT is more or less reduced to the role of a model-building device that does not carry much explanatory weight. The same results could be obtained if the agents were assumed to behave *as if* they were satisfying agents instead of maximizing agents.

This situation makes the evaluation of the externalist interpretation quite difficult. Satz and Ferejohn leave the selection processes, which are the real explanatory mechanisms, unspecified. Consequently, it is impossible to judge how wide the scope of application of their interpretation of RCT is. In this situation, the critics seem to be justified in their skepticism (Green and Shapiro 1994: 22-23). Natural selection cannot be treated as an automatic saviour of RCT, as Satz and Ferejohn and a number of economists (see Vromen 1995) seem to assume. There are clear limits to the applicability of evolutionary explanations, and if the requirements of these explanations are not spelled out explicitly, the externalist interpretation of RCT is open to allegations of being mere wishful thinking. It is not enough that one can build a model, one should also be able to justify its assumptions.

What is the relevance of this discussion to our understanding of interest explanations? In the first place, the interest approach, as I have described it in Chapters 5 and 6, is not as psychologically unrealistic as the standard RCT. As noted earlier, it is intended to be compatible with an empirical account of our cognitive capacities. Consequently, the motivations for abstracting from the mental life of the agents are different. Satz and Ferejohn present their externalist interpretation because the internalist interpretation of RCT makes RCT empirically suspect. On the other hand, interest theorists are motivated by a wish to make their explanations stronger.

The second important difference is methodological. Satz and Ferejohn describe RCT as an approach that is oriented towards theoretical model building. This is not the approach of social studies of science, which are oriented toward empirical case studies of historical and contemporary episodes. It is possible to introduce the model-building approach to social studies of science by using ideas incorporated in the model of cycles of credibility, for example. As suggested in Chapter 6, one could attempt to explain some general characteristics of science by building models based on the entrepreneurial analogy. However, this has not been done yet.

On the positive side, the externalist interpretation of RCT provides an interesting way to interpret the model of cycles of credibility. In this interpretation, this model does not describe practical reasoning by scientists, but the general structure of scientific research efforts. The scientists might in fact be motivated by all sorts of interests and goals, the only relevant fact is that they act *as if* they were trying to maximize their credibility. This seems to be a more faithful reconstruction of the original ideas of Latour and Woolgar than the Pettit-inspired reconstruction presented earlier in this chapter. Why should we suppose that scientists fulfill this behavioral assumption? The selection mechanism takes the responsibility here: those scientists whose research does not contribute to their credibility do not succeed in the scientific marketplace. They are either marginalized and driven away due to lack of resources or they learn the trick and start to do research that contributes to their scientific capital. The analogy with the neoclassical theory of the firm is strong. The model is not about the psychological motives of scientists, just as the neoclassical theory of the firm is not about the psychologies of managers.

The explanatory power of this interpretation depends strongly on the assumption of selection. The model works only if the selection mechanism is strong enough. As far as possible *explananda* of the model interpreted this way are concerned the analogy with the neoclassical theory of the firm is again relevant. Just as the neoclassical theory of the firm is not intended to explain the behavior of any particular firm, the model of professional interests is not intended to explain the behavior of individual scientists or research groups. Both aim to explain more general features. The neoclassical theory of the firm aims to explain some general characteristics of the economy, and the theory of scientific capitalism some generic features of the research system and its workings. In both of these tasks most of the idiosyncratic characteristics of firms/scientists can be safely abstracted away. So, if we are to develop the model of cycles of credibility into an explanatory theory of some general structures of academic science, the analogy with externally interpreted rational choice theory seems to be heuristically promising.

I have only considered the relevance of Satz and Ferejohn's ideas to interest explanations. There might be also some relevance in the other direction. The interest approach might contribute to a better understanding of RCT. Satz and Ferejohn give a somewhat simplistic picture of the attribution of interests. The agents' interests or preferences are not simply inferred from the description of the structural parameters of their situation, as they suggest (Satz and Ferejohn 1994: 78, 86). The application of interest concepts always presupposes the attribution of some basic goals. Only against the background of these goals is it possible to determine what is in the interest of the agent. Similarly, indi-

vidual preferences cannot be simply constructed "out of the social environment". There has to be at least some contribution from the agents' goals and desires.

3. Elster against functional explanation

Interest explanations based on non-intentional filtering mechanisms and the externalist interpretation of RCT face a similar challenge. They should be able to show that non-intentional filtering explanations are legitimate in the social sciences. The standard argument against their legitimacy was advanced by Jon Elster. Elster assimilates non-intentional filtering explanations into a broader category of functional explanation. In this section I will take a close look at Elster's argument. My intention is not to defend all functional explanations in the social sciences or to advocate any sort of functionalism as a social science paradigm. My primary aim is to illuminate and legitimate non-intentional filtering explanations by offering a critical account of Elster's argumentation.

My basic claim is that Elster's analysis of functional explanation is not adequate and that his conclusions are broader than his argument supports. I will proceed as follows. First, I will briefly draw attention to a distinction between *functional analysis* and *functional explanation* that is overlooked by Elster. I do this in order to make the idea of functional explanation better understood and also to show that Elster's argument applies only to some uses of functional concepts in the social sciences. Then I will proceed to consider the requirements Elster sets for acceptable functional explanation. My discussion will concentrate on the distinction between manifest and latent functions that Elster adopts from Robert Merton. I argue that Elster's analysis is weakened by some problems with this distinction. I also offer an alternative conceptualization. The final part of this section will discuss Elster's assumptions about functional explanations in biology and raise some related issues about his preferred methodological positions.

Functional analysis and functional explanation

The background for Elster's argument is the prevalence of functional language in the social social sciences. He seems to assume that his critique of functional explanation implies a refutation of the whole functional way of thinking in the social sciences. This interpretation is not warranted. Functional language is widely used in the social sciences, but there is no general agreement concerning the use of terms and concepts. Most of the functional talk in the social sciences is not intended to be explanatory at all (Boudon 1990: 137-139). As Elster's arguments

concentrate on explanation, they are not a refutation of non-explanatory uses of functionalist concepts and ideas. In the following I will only concentrate on functional explanation and leave out all discussion about functionalism as a social science paradigm. (For a more general discussion on functional explanation, see van Parijs (1981) and Kincaid (1996: Chapter 4).)

There is no consensus among social scientists concerning the proper use of the concept of function. The situation is not better among philosophers of science. However, a recent discussion in philosophy of biology has produced two basic accounts of functions (Millikan 1989; Godfrey-Smith 1993). These accounts, causal role functions (or Cummins-functions, named after Robert Cummins) and etiological functions (or Wright-functions, named after Larry Wright), illuminate two different explanatory uses of functional concepts. Both are compatible with the idea that functional explanations are in the final analysis causal explanations. I will call the first account *functional analysis* and the second *functional explanation*.

In functional analysis[4] one tries to find out the function, understood as the causal role, that an entity (a component) has in some larger whole (a system). What a particular component (*X*) does, or can do, in system (*S*) is its function (*Y*) in it. The *explanandum* of functional analysis is some property or capacity of the system. It is explained by analyzing the system to its components and by showing how the organization of these components produces the property to be explained. Functional analysis is clearly an example of explanation of properties mentioned in Chapter 2. (For an exposition of the use of functional analysis in psychology and the biosciences, see Cummins 1983, 2000; Bechtel & Richardson 1993; Amundson & Lauder 1994.)

It is important to notice that *the explanandum is not the existence of X*. Functional analysis is not trying to explain how or why *X* is there. In functional analysis it is meaningful to say that one function of a nose is to support the spectacles, although this is not the reason why the nose exists. Functional analysis is completely ahistorical: it does not explain why a given component exists; the relevant question is instead '*what does X do?*', or '*how does X contribute to the system?*'. The *explanandum* is a property, not an event. Furthermore, this property is not a property of *X*, but of the system. One can ask 'what is the function of a spark plug in an engine?' The answer would show how a spark plug contributes to the proper working of an engine. Analogously one can ask how the institution of the family contributes to the stability of a society.

A functional explanation answers a question left open by functional analysis. In functional explanation the question to be answered is: '*why is X there?*'. The *explanandum* is the existence (or persistence) of *X*. The explanation is given in terms of *X*'s earlier beneficial effects (*Y*). Func-

tional explanation is *historical explanation*: only past beneficial effects are explanatory. Beneficial effects in general or in the future are not explanatory in functional explanation.

A functional explanation has two parts. The first part claims that Y is an effect of *X*. The second part states that *X* exists because it causes *Y*.[5] The second claim presupposes that there is a mechanism that ensures that *X* exists when it has consequence *Y*. The relevant counterfactual here is: were *X* not producing *Y*, *X* would not exist. In evolutionary biology, functional explanations are backed up by the mechanism of natural selection. In the social sciences intentional human action can work as such a mechanism. The question to be discussed later in this section is whether there are any non-intentional mechanisms.

The *explanandum* of a functional explanation is not a singular event; neither does it address questions concerning *X*'s origin. The fact to be explained is the *persistence* of *X*. For example, theory of natural selection does not explain why a particular variant (mutation) was there, but it explains why a particular variant increases its frequency in population and why it persists in the population. As argued in Chapter 4, an evolutionary explanation is contrastive: it explains why a variant *A* was more successful than variants *B*, *B*', …, or *B** in some particular environment. If two variants (A and B) have equal fitnesses, an evolutionary explanation cannot make a difference between them. It cannot explain why there is *A* rather than *B*. In such a situation the explanation is to be found from the contingent historical facts about the population. Naturally, such an explanation would not be a functional explanation.

There are some interconnections between functional explanation and functional analysis. For example, one could claim that most functional explanations presuppose functional analysis because the entities or properties to be explained (for example, organs or other characteristics) are parts of larger systems (organisms), and their success is determined by their contribution to these larger systems. The reverse is not true: a functional analysis of a component is possible even if there is no hope of explaining the presence of a component functionally.

One noteworthy difference between functional explanation and analysis is the concept of *dysfunction*. In functional analysis it is meaningful to speak of dysfunctions of a component. In principle, there is no asymmetry between functions and dysfunctions. One can explain why a system does not have a certain property or why it does not work by referring to dysfunctions of its components. And in general one can list both functions and dysfunctions of an item when making a functional description of the components of a system. No such symmetry is to be found in functional explanation. A functional explanation can mention *X*'s dysfunctions, but the dysfunctions can never explain the presence of *X*. Quite the contrary, they make the explanatory task even harder. (For the concept of dysfunction, see Merton 1967: 105.)

Most 'functional explanations' in the social sciences are closer to functional analysis than to functional explanation. Note the frequency of references to dysfunctions in the social sciences. This is natural, as the original inspiration for functionalist thinking in sociology came from physiology. Consequently, Elster's broad critique of functionalism in the social sciences does not hit its target.[6]

The fact that Elster's discussion does not apply to functional analysis is not yet a defense of non-intentional filtering explanations. These explanations are clearly functional explanations in the above sense. We have to consider next whether Elster is successful in showing that functional explanation in a more narrow sense is illegitimate in the social sciences.

Elster's criteria for acceptable functional explanation

According to Elster (1979: 28) an institution or a behavioral pattern *X* is explained by its function *Y* for group *Z* if and only if:

(1) *Y* is an effect of *X*;

(2) *Y* is beneficial for *Z*;

(3) *Y* is unintended by the actors producing *X*;

(4) *Y* (or at least a causal relationship between *X* and *Y*) is unrecognized by the actors in *Z*;

(5) *Y* maintains *X* by a causal feedback loop passing through *Z*.

According to Elster, functional explanation has to satisfy these five requirements in order to be acceptable. He draws requirements (1)-(4) from Robert Merton's writings and requirement (5) from Arthur Stinchcombe. Elster claims that most functional explanations fail by presuming rather than showing that requirement (5) is satisfied. (Elster 1979: 29-30; 1983: 58-59.)

As Elster himself points out, (1) is quite unproblematic. There is no sense in saying that the function of *X* is to *Y* if *Y* is not an effect of *X* (Elster 1979: 29). The sole exceptions are broken artifacts. One might say that the (designed) function of *X* is to *Y*, but it does not cause *Y*, because *X* does not work for some reason. Non-functional hearts and kidneys are also borderline cases that are not relevant in this context. From the point of view of my discussion, (1) can stand as it is. In line with the position developed in Chapter 2, I will also take the basic idea behind (5) for granted. All explanations require a mechanism, and functional explanations should not be exceptions. In the following discussion I will concentrate on Elster's requirements (2), (3) and (4).

Who will benefit?

Let us start with requirement (2). As a clarification Elster defines a beneficial effect as follows:

> *Y* is a local maximum of some state variable of which the actors in *Z* always want more than less (Elster 1979: 29).

Elster does not discuss this issue further, so there might not be so much to criticize. However, I think something should be said. First, it is not clear why we should suppose that the 'beneficiaries' are always persons or groups. G. A. Cohen gives the following example:

> Suppose some practice *X* has an effect *Y* which keeps an institution *Z* stable, as a result of which *X* is reproduced. [...] There could be an air force devoted to suicide flights which inspire reckless civilian youths who consequently enlist and perpetuate the suicide flying – but not in order to attract new recruits and thereby stabilize the air force. This seems to yield an Elster loop – *X* = suicide flying, *Y* = youths being inspired, *Z* = the air force – and not because anyone benefits from suicide flights, or from the stability of the air force (Cohen 1982: 44).

This example is extreme, but I think that it shows that one should not *a priori* decide that beneficiaries should be persons or groups. Cohen's example is analogous to many examples in the literature of cultural evolution (Sperber 1996). In these models it is not assumed that the people having the beliefs (or memes) benefit from these beliefs in any obvious sense. There is an analogy in evolutionary biology: not all consequences of natural selection are to the benefit of an organism (nor for the benefit of other organisms, as is in the case of parasitism). Consider for example viruses, or so-called selfish DNA. Their effect is negative or neutral from the point of view of the organism. The real 'beneficiaries' are the viruses or DNA segments themselves. The benefit is understood here as success in reproduction. This shows that it is illegitimate to limit the range of beneficiaries to individuals without further argument. In some contexts the benefit just means that *Y* contributes to *X*'s persistence. In these cases there is no need for any specific *Z*.

Elster's use of the notion of want in the above quotation creates some problems. This is the case even if we restrict ourselves to persons and groups as beneficiaries. The first problem is the attribution of wants to groups. Groups can have goals, but do they have desires or wants? Secondly, no matter how wants are analyzed, they must include some intentional and conscious components. And this is a problem, since we are talking about latent functions. In the case of latent function, It is not required that agents know the function or that they recognize the benefit it brings. These points can be generalized to a question: *why*

import psychological concepts in this context at all? Functionalists themselves sometimes speak of systemic needs, requirements or prerequisites. These concepts may be as problematic as the concept of want, but that is not the issue here. The point is that these are not psychological concepts as Elster's 'want'.

The third major problem concerns the concept of optimality. I agree that an assumption of (local or global) optimality is quite often made, but I do not think that it is necessary for functional explanation to be interesting. As argued in Chapter 4, the *explananda* of evolutionary explanations are always contrastive. One does not simply explain the presence of *X*. Rather, one explains why there is *X* rather than *X'* or *X''*. The explanation presupposes that *X* is more optimal than *X'* or *X''*, but it permits that there might be a further variant *X** that is more optimal than *X*. *X** might have not yet emerged because of the shortness of time or it may never be available for a selection process. It might also be that for some accidental reason *X**'s became extinct early on.[7] The important point here is that an evolutionary explanation always contrasts actual alternatives, not possible alternatives. The same point applies to evolutionary explanations in the social sciences. There is no reason to demand that social scientists should give more ambitious explanations than evolutionary biologists. Because of this, social scientists should be understood as answering questions like 1), not questions like 2):

[1] Why *X* rather than any of the *specified* alternatives to *X*?

[2] Why *X* rather than any of the *imaginable* alternatives to *X*?

An evolutionary answer to question [1] requires showing that *X* is more optimal than its specified alternatives. Answers to questions like this are informative, and constructing them is hard enough. On the other hand, an evolutionary answer to question [2] requires showing that *X* is a local or global optimum, which is a much more demanding *explanandum*.

We can try to create more ambitious explanations showing that *Y* is a local or global optimum and that it would have won against every possible alternative. But this explanation is very ambitious, and I do not think that social scientists or biologists usually aim that high. An optimality explanation presupposes a long and stable period of time for selection to work. One has to suppose that there has been enough time for all possible variants to emerge and for selection to work on them. In a more modest explanation one only supposes that there has been enough time for selection to work on a given set of variants. To my mind this makes modest *explananda* much more realistic than more ambitious optimum explanations. Furthermore, by constructing one's *explananda* more modestly, one can more easily avoid charges of adaptationism. One would only claim relative optimality, not absolute. Similarly, one would only be claiming that selection is an important

force in evolution, not that it is omnipotent.

Functions: latent and manifest

Elster's analysis is based on Robert Merton's distinction between manifest and latent functions. Merton introduced this distinction to distinguish between "… those objective consequences for a specified unit (person, subgroup, social or cultural system) which contribute to its adjustment or adaptation and were so intended" and "… unintended and unrecognized consequences of the same order" (Merton 1967: 117).

Merton's use of this distinction is not uniform. For example, sometimes the notion of manifest function is used to refer to the intended consequences of an action, without any reference to its actual consequences. The distinction has many different uses even in Merton's article and in the functionalist literature. It has been used to make at least three distinctions:

i) the distinction between the public (or official) functions of some practice and its unofficial (or illegitimate) functions;

ii) the distinction between functions that the informants attribute to a practice and the functions that are inferred or observed by a social scientists;

iii) the distinction between common sense accounts and sociological descriptions.[8]

These distinctions do not match easily with the 'official' definition of manifest and latent functions. The ambiguity does not spring merely from Merton's conceptual sloppiness. One important factor is the ambiguity in the basic concepts used in the definition. By Merton's definition, latent functions are both unintended and unrecognized, and manifest functions are both intended and recognized. The definition makes two presuppositions:

1) there is a close connection between intending and recognizing;

2) there is a close connection between not intending and not recognizing.

In the light of my discussion in Chapter 6, both of these presuppositions are problematic. I distinguished three kinds of unintended consequences of action:

UIC 1) consequences that were not the agent's reasons for her actions, but were taken into account in her practical reasoning

UIC 2) consequences the agent anticipated, but which were disregarded in her practical reasoning

UIC 3) consequences the agent did not anticipate.

Notice that the agent's later recognition of the consequences of her action is not relevant to this classification. The distinction between recognized and unrecognized consequences crosscuts all the above categories. First, there can be intended consequences that are not later recognized by the agent. For example, an agent can try to bring about some state of affairs, but still be unable to check later whether she was successful. The same goes for consequences that he anticipated but did not intend to bring about. On the other hand, an agent can also learn about the consequences of her action that she neither intended nor anticipated. These distinctions show that greater care should be taken in defining manifest and latent functions.

From the point of view of functional explanation, cases where something is intended but not later recognized are quite uninteresting, since the interest lies in regular behavioral patterns, not in singular actions. However, this shows that Merton's typology is not exhaustive. It is an unhappy combination of two completely different distinctions: 1) between intended and unintended consequences; and 2) between recognized and unrecognized consequences.

Elster's characterization of functional explanation is also based on this unhappy dichotomy between manifest and latent functions. The following explanation of the growth of the American national bureaucracy by Morris Fiorina shows how the problems in defining these two concepts are reflected in the difficulties with Elster's requirements (3) and (4). According to Fiorina the careers of the members of Congress benefit from the unintended growth of bureaucracy. The following is Russell Hardin's reconstruction of Fiorina's argument:

> X: growth of bureaus (which get their budgets from Congress and which are therefore responsive to congressional requests for constituency assistance);
>
> Y: re-election of members of Congress (who indulge their constituents by intervening in bureaus);
>
> Z: members of Congress.
>
> Because they spend more time servicing constituents, members of Congress delegate more power of decision and resources to bureaus so that, although this is not intended, constituents have increased dealings with bureaus.
>
> There is feedback from congressional careers to growth of the bureaucracy in at least two ways. (1) The growth of bureaucracy produces more demands by constituents and therefore more occasion for members of Congress to seek the ombudsman role. (2) Playing ombudsman distracts members of Congress from legislative and oversight roles, so that they devolve more discretion to administrative agencies. The result is the

> selected survival of the fittest members of Congress – those whose constituency and interest group service wins them enough additional votes to lift them above the status of marginal re-electability (Hardin 1980: 758).

Clearly, it is irrelevant from the point of view of this explanation whether members of Congress recognize the feedback mechanism between the growth of bureaucracy and their own reelection. The mechanism works without regard to whether or not members of Congress recognize it, because all representatives are tied to the same practice to ensure their reelection.

The phenomenon is quite general: the mere recognition of certain consequences of action is irrelevant from the point of view of its causal explanation (Grimen 1994: 119). Requirement (4) is redundant. The only way the recognition of an effect can be causally relevant in explanation is when it is fed back into agents' intentions. That is, agents are trying to achieve it or trying to avoid it. However, when this happens, (3) applies and (4) is again redundant. *Elster's requirement (4) can de dropped altogether*. It does not affect the explanation in any way.

This example can also be used to point out some problems with Elster's requirement (3). What happens when the members of Congress recognize this mechanism and start to use it intentionally to ensure their reelection? I would say that there are no drastic effects on our explanation for the growth of bureaucracy. The same causal mechanism is at work as before. The only change might be the accelerated rate of growth of the bureaucracy that was caused by the attempts of members of Congress to speed up the process to their advantage. Of course, a full explanation would *mention* that the members of Congress are fully aware of the situation and they act on the basis of this knowledge. However, the basic pattern of the explanation would remain the same. What is important from the point of view of the explanation is the feedback loop from the services rendered to the constituency to the growth of bureaucracy. A counterfactual test shows its explanatory relevance. If the feedback mechanism exists, the result will follow irrespective of the intentions of the members of Congress. However, if the feedback mechanism is not in place, the member of Congress cannot improve their chances of reelection this way, no matter how hard they try.[9]

This example shows that sometimes a functional explanation can work irrespective of agents' intentions. The later recognition, and the intentional action based on it, does not affect the functional nature of explanation before this recognition occurs. And when the recognition takes place, the essential elements of the explanation remain the same. If these points are correct, the satisfaction of requirement (3) is not a necessary feature of all functional explanations.

There are also other problems with requirement (3). Consider a case of recurring wayward causation. Let us suppose that the members of *Z*

intend to bring about *Y* by way of *X*, and although *Y* actually occurs, it is not brought about in the way it was intended. One cannot explain such a case solely by referring to the agents' intentions. If the agents intend to bring about *Y*, but not by way of *X*, these intentions do not affect the potential functional explanation of *X* by *Y*. This means that requirement (3) must be changed to rule out wayward causation. It should read:

(3'): The actors do not intend to bring about *Y* by way of *X*.

This reformulation is, however, only a minor improvement and it does not save Elster's analysis because neither (3) nor (3') can be a necessary requirement, as argued above.

There is also some ambiguity concerning the agents in question. The above discussion supposes for the sake of simplicity that the agents practicing *X* and the agents in group *Z* are the same. However, Elster claims that the agents engaged in *X* need not be the same agents that constitute the group *Z*. According to him these two groups can be overlapping, inclusive, and disjointed (Elster 1979: 29; Elster 1990: 131). This is all right, but if the group acting and the group recognizing are different, then one may wonder what is meant by the recognition?

To see the problem, consider Elster's account of filter explanation.[10] According to Elster in filter explanation his requirements (1)-(3) and (5) are satisfied, but not requirement (4). He gives the following as an example of filter explanation:

> Militarily financed research may be analyzed by a filter-explanation. If academic personnel apply for military funds in order to be able to conduct the research that they would have done in any case (i.e. with money from any other source), the Department of Defence may serve as a filter that selects some applications and rejects others. The resulting composition of research will be beneficial to the military interests, while wholly unintended by the individual scientist, who can argue truthfully that no one has told him what to do. This may also be called a case of *artificial selection*, where the feedback loop operates through the *recognized* effects of the structure whose persistence is to be explained (Elster 1979: 30).

As I interpret this passage, scientific research is meant to be *X*, the agents involved in *X* are scientists, *Y* is the composition of research, and the *Z* is the Department of Defense (or an equivalent military interest group). The basic idea is that the scientists do not intend to promote military science, but as a consequence of the funding structure they actually do it.

The first problem is that the Department of Defense does not merely recognize or observe the process, it actively promotes it. The feedback loop demanded by (5) presupposes this intervention. Consequently, as this explanation satisfies neither (3) nor (4), this cannot be an example

of filter explanation, as Elster defines it.

Second, scientists' intentions are irrelevant from the point of view of explanation, because the intentional selection by the Department of Defense explains the composition of research. This is an empirically interesting fact if the scientists do not recognize the consequences of their actions, but from the point of view of the basic explanation pattern, this issue is irrelevant. As long as the Department of Defense controls the funding as it does, the same composition of research will result whether or not scientists intentionally promote it. As I see the situation, this is basically a simple *intentional filtering explanation*. The Department of Defense selects research projects it believes to be conducive to its goals and the scientists' actions serve the interests of the Department of Defense even if they do not share its goals.

Can we call this an example of artificial selection as Elster does? I think we should not because the Department of Defense does not select projects by their actual results. It makes its funding decisions based on its *beliefs* about the anticipated results. In contrast, artificial selection works by selection of actual results, not by selection of anticipated results. This observation suggests a distinction within the category of intentional filtering explanations. An intentional filtering can work either by the anticipated results or by the actual results. Both examples of intentional filtering explanation in the previous chapter were examples of the former. So is Elster's military example.

These seem to be good reasons to reconsider Elster's account of filter explanations. Elster encounters these problems for two reasons. First, he starts with the distinction between manifest and latent functions. Second, he limits the functionalist program to latent functions and takes manifest functions to be unproblematically explainable by intentional means. Let us call this latter assumption *intentional chauvinism*. It is based on the presumption that recognition implies an intention and that intention implies the irrelevance of a non-intentional mechanism. This presumption is not legitimate.

My proposal is that we avoid intentional chauvinism by not building our taxonomy of explanations upon Merton's distinction. One should start from the fact that functional explanation is explanation by *actual* effects. This distinguishes it from intentional explanation, which is explanation by *anticipated* effects. The class of intentional filtering explanations cuts across these two categories. Intentional filtering explanations that are based on anticipated results are basically intentional explanations. Intentional filtering explanations that refer to recognized actual effects that inform the later intentions of agents are both functional and intentional explanations. Consequently, the categories of intentional and functional explanation are not exclusive. On the other hand, explanations based on non-intentional filtering mechanisms are clearly functional explanations, since they explain by actual results.

Consider now the classic example of Hopi indian rain dance. As the story goes, the Hopis claim that they dance in order to create rain. Anthropologists explain the ritual by its effects on group cohesiveness because there is no causal connection between dancing and raining. We can have alternative interpretations of the anthropologists' explanation for the persistence of this ritual. An anti-functionalist would try to transform the explanation to an intentional filter explanation (if he does not abandon it entirely). According to this construction, the agents recognize the beneficial effects of the ritual and this recognition informs their intentions. The feedback mechanism is intentional. On the other hand, a functionalist would try to construe some kind of non-intentional selection mechanism as a feedback loop. We need not go into these details here, as the point I want to make is that the difference between these two explanations is not that great. Basically we explain this institution in both cases by its actual effects, that is, functionally. The actual effects make up the main part of explanation, while the specific nature of the feedback mechanism is an auxiliary question (as long as there is a feedback mechanism).

The difference is blurred even more if one considers the example more closely. First, if the ritual has persisted for a very long time, it might be that more than one feedback mechanism has been working. These mechanisms might have worked at different times or at the same time. It might be that the social functions of the rain dance are sometimes recognized and that sometimes this knowledge is forgotten. Although some of the agents know about the beneficial effects of the ritual, nothing ensures that this knowledge will persist as long as the institution.

Consider next the number and status of people having this knowledge. When proposing intentional filtering explanations one should ask two questions. First, how many persons have to know the effects of X for the filtering mechanism to be effective? Second, what kind of social status should these persons have to have in the society? Sometimes the recognition may be without any real effect in the process of persistence. This can happen when only few persons have the knowledge or if their status in society does not allow them to affect the way the institution functions. More than recognition by someone is required. In such cases, the reference to the social effects of the ritual may be the safest part of the explanation, the details of the feedback mechanism being more messy and variable.

The same standards for all

The most fundamental criticism Elster raises against functional explanation is that:

> …many purported cases of functional explanation fail because the feedback loop criterion (5) is postulated rather than dem-

> onstrated. Or perhaps 'postulated' is too strong, a better term being 'tacitly presupposed'. Functionalist sociologists argue as if (which is not to argue that) criterion (5) is automatically fulfilled whenever the other criteria are (Elster 1983: 58-59).

First, a functional explanation can succeed only if there are reasons for believing in a feedback loop from the consequences to the phenomenon to be explained. Secondly, these reasons can only be the exhibition of a specific feedback mechanism in each particular case. The second premise is not needed in the case of functional explanation in biology, for here we have general knowledge - the theory of evolution through natural selection - that ensures the existence of some feedback mechanism, even though in a given case we may be unable to exhibit it (Elster 1983: 61).

I do not want to argue against requirement (5), which demands that there has to be a causal feedback mechanism. If someone denies this requirement, Elster is right in raising this issue. My concern is that Elster is not totally consistent in this demand. It seems to me that his standards for functional explanation in the social sciences are stricter than for other kinds of causal explanation and especially for functional explanation in biology.

In the case of functional explanation in the social sciences, Elster demands that the feedback mechanism should always be specified for each particular case. However, he does not demand the same from evolutionary explanations in biology. This cannot be right. Not everything in biology is explained by natural selection. Not all traits are functional or selected for these functions. One should not infer the commitment to unqualified adaptationism from the importance of natural selection. As I see it, both kinds of evolutionary explanation are in the same boat. They both require that the explanatory mechanism be specifiable. Similarly, Elster does not demand that the causal mechanism should always be specified for ordinary causal explanations. One should ask why? There are spurious causes as there are spurious functions. The requirement of the mechanism is as relevant here as it is in the case of functional explanation.

Elster's selective demands on the causal mechanism correlate with his methodological preferences. For example, he argues against non-individualist social science by demanding that causal mechanisms be specified, but does not demand the same of individualistic explanations. However, there are no convincing reasons for this asymmetry. One should consider all demands made on causal mechanisms similarly. There is no justification for preferential treatment for some forms of explanation. There is no reason to set functional explanations apart from other explanations. It might be that acceptable functional explanations are not very common, but their infrequency should not be an argument against their legitimacy.

The implications of my argument

Now it is time to summarize the implications of my critique of Elster's account of functional explanation. I have argued that Elster's requirements (2), (3), and (4) misrepresent the nature of functional explanation. Requirement (2) constrains in an arbitrary fashion the *explananda* a functional explanation could have. There is no reason to suppose that the functional benefit should be for persons or groups, or that it should be understood in terms of people's wants. Secondly, the assumption that the item to be explained is optimal puts excessively strong demands on functional explanation. Against requirement (3) I argued that it assumes illegitimately that only latent functions need functional explanation. If the relationship between the intended, unintended, and recognized consequences of an action is understood correctly, the categories of intentional and functional explanation are not exclusive. Finally, I argued that requirement (4) is completely irrelevant from the point of view of explanation.

My conclusion is that Elster's analysis does not provide legitimate grounds for criticizing non-intentional filtering explanations. Elster's central point about the importance of a plausible causal mechanism for the legitimacy of an explanation is valid, but it does not distinguish non-intentional filtering explanations from other forms of explanation. All explanations face the same challenge.

The failure of Elster's analysis does not show that there is no need for a explication of the structure of functional explanation. To the contrary, such explication is badly needed. However, Elster's failure might indicate that a general account of functional explanation is not the best way to proceed. There seems to be too much variation in patterns of functional explanation for an informative general account to be possible. For example, the cultural selection explanations discussed in previous chapter seem to be quite dissimilar to Fiorina's explanation of the growth of bureaucracy discussed in this chapter. It seems plausible that the analysis of functional explanations should proceed in a less ambitious manner. It should explicate various forms of functional explanation without assuming an underlying unity. It might also be profitable to change the focal point of attention away from the concept of function. It seems that linguistic intuitions about this concept have often diverted attention from the true nature of these explanations.

Notes to Chapter 7

1 This can be easily seen, for example, in the case of political science by looking at various accounts of rational choice found in Friedman 1996. For a review of the use of RCT in sociology see Hechter and Kanazawa 1997 and Goldthorpe 1998.

2 For example, sociologists might find the psychological theory developed by Howard Margolis (1988, 1993) highly interesting. This is surprising since Margolis tries to give impression that his theory provides an alternative, not a building block, to the accounts of science provided by sociologists of scientific knowledge.

3 The externalist interpretation by Satz and Ferejohn has similarities to Karl Popper's discussions of 'the logic of situation'. Popper, who seems to have picked up the basic idea from Max Weber, never developed this idea very far. For discussion, see Latsis 1972 and Hedström, Swedberg & Udehn 1998.

4 Here the term 'functional analysis' is used in the sense Cummins (1983) uses it. It should not be confused, for example, with Hempel's (1965: 297-330) famous account of functional analysis.

5 This analysis is simplified and not sufficient. (See Wright 1976; Boorse 1984; Millikan 1984; Davies 1994; Godfrey-Smith 1994.) However, a more detailed analysis is not needed to show the difference between functional analysis and functional explanation.

6 For example, Elster claims that Robert Merton is the leading exponent of what he calls the Main Functional Paradigm, which holds that "latent functions of institutions also explain those institutions" (Elster 1982: 455). Elster states: "personally I tend to get the impression that Merton thinks of functional analysis as providing an explanation of the phenomenon to which these functions are imputed" (1979: 31) Elster's evidence is quite shallow. A more adequate interpretation is that the few ambiguous instances that Elster is able to mention are unfortunate slips that have nothing to do with the basic ideas of Merton's functional analysis. (For a similar evaluation, see Sztompka 1986: 140-141, Elster 1990 seems to agree.)

7 Recall from Chapter 4 that in evolutionary theory, a better relative adaptedness or fitness raises the chance of success, but it does not guarantee it. Natural selection is not the only force of evolution.

8 For a critical discussion of Merton's use of the distinction between manifest and latent functions see Helm (1971) and Campbell (1982).

9 The above reasoning presupposes that the members of the Congress do not try to act against the mechanism. But this is secured by the game-theoretical structure of the electoral competition. Note an interesting feature of this example. In it an explanation by RCT and functional explanation are combined (Hardin 1980: 758-760). This shows that functional explanations and the use of game-theoretical analyses are not alternatives to each other, as Elster (1982) seems to suggest. This observation is not surprising when one recalls that the externalist interpretation of RCT advocated by Satz and Ferejohn gave an important explanatory role to evolutionary considerations. For a discussion of the role of evolutionary considerations in economics, see Vromen 1995.

10 My account of intentional and non-intentional filtering explanations not based on Elster's definition of filter explanation.

Bibliography

Achinstein, Peter 1983: *The Nature of Explanation*. Oxford University Press. Oxford.

Ahn, Woo-kyoung & Kalish, Charles W. 2000: 'The Role of Mechanism Beliefs in Causal Reasoning', in Keil & Wilson (eds.): *Explanation and Cognition*. MIT Press. Cambridge: 199-226.

Amundson, Ron & Lauder, George V. 1994: 'Function without Purpose: The Uses of Causal Role Function in Evolutionary Biology', *Biology & Philosophy 9*: 443-470.

Anderson, John 1938: 'The Problem of Causality', *Australasian Journal of Psychology and Philosophy 16*: 127-142.

Baker, Lynne Rudder 1995: *Explaining Attitudes. A Practical Approach to the Mind*. Cambridge University Press. Cambridge.

Barnes, Barry 1974: *Scientific Knowledge and Sociological Theory*. Routledge & Kegan Paul. London.

Barnes, Barry 1976: 'Natural Rationality: A Neglegted Concept in the Social Sciences', *Philosophy of the Social Sciences 6*: 115-126.

Barnes, Barry 1977: *Interests and the Growth of Knowledge*. Routledge & Kegan Paul. London.

Barnes, Barry 1981: 'On the 'Hows' and 'Whys' of Cultural Change', *Social Studies of Science 11*: 481-98.

Barnes, Barry 1982: *T. S. Kuhn and Social Science*. Columbia University Press. New York.

Barnes, Barry 1985: *About Science*. Basil Blackwell. Oxford.

Barnes, Barry 1988: *The Nature of Power*. Polity Press. Cambridge.

Barnes, Barry 1995: *The Elements of Social Theory*. UCL Press. London.

Barnes, Barry & Bloor, David 1982: 'Relativism, Rationalism and the Sociology of Scientific Knowledge', in Hollis & Lukes (eds.) *Rationality and Relativism*. Basil Blackwell. Oxford: 21-47.

Barnes, Barry, Bloor, David & Henry, John 1996: *Scientific Knowledge. A Sociological Analysis*. Athlone Press. London.

Barnes, Barry & MacKenzie, Donald 1979: 'On the Role of Interest in Scientific Change', *Sociological Review Monograph 27*: 49-66.

Barnes, Eric 1994: 'Why P Rather Than Q? The Curiosities of Fact and Foil', *Philosophical Studies 73*: 35-53.

Beatty, John 1980: 'Optimal-Design Models and the Strategy of Model Building in Evolutionary Biology'. *Philosophy of Science 47*: 532-561.

Beatty, John 1984: 'Chance and Natural Selection'. *Philosophy of Science 51*: 183-211.

Beatty, John & Finsen, Susan 1989: 'Rethinking the Propensity Interpretation: A Peek Inside Pandora's Box', in Ruse (ed.): *What the Philosophy of Biology Is*. Kluwer. Dordrecht: 17-30.

Bechtel, William and Richardson, Robert C. 1993: *Discovering Complexity. Decomposition and localization as strategies in scientific research*. Princeton University Press. Princeton.

Biagioli, Mario 1993: *Galileo Courtier. The Practice of Science in the Culture of Absolutism*. The University of Chicago Press. Chicago.

Blackburn, Simon 1991: ' Losing your mind: physics, identity, and folk burglar prevention', in Greenwood (ed.) *The future of folk psychology*. Cambridge University Press. Cambridge: 196-225.

Block, Ned 1990: 'Can the mind change the world?', in Boolos (ed.): *Meaning and Method: Essays in Honor of Hilary Putnam*. Cambridge University Press. Cambridge: 137-170.

Bloor, David 1983: *Wittgenstein. A Social Theory of Knowledge*. Macmillan. London.

Bloor, David 1991: *Knowledge and Social Imagery* (2nd ed). The University of Chicago Press. Chicago.

Bloor, David 1997: *Wittgenstein, Rules and Institutions*. Routledge. London.

Bloor, David 1999: 'Anti-Latour', *Studies in the History and Philosophy of Science 30*: 81-112.

Bohman, James 1991: *New Philosophy of Social Science*. Polity Press. Cambridge.

Boorse, Christopher 1984: 'Wright on Functions', in Sober (ed.): *Conceptual Issues in Evolutionary Biology*. MIT Press. Cambridge: 369-385.

Boudon, Raymond 1990: 'Boudon Replies to Elster', in Clark, Modgil & Modgil (eds.): *Robert K. Merton. Consensus and Controversy*. Falmer Press. London: 136-137.

Bourdieu, Pierre 1975: 'The Specificity of the Scientific Field and the Social Conditions of the Progress of Reason', *Social Science Information 14*: 19-47.

Braddon-Mitchell, David 1993: 'The Microstructural Causation Hypothesis', *Erkenntnis 39*: 257-283.

Brandon, Robert 1978: 'Adaptation and Evolutionary Theory', *Studies in History and Philosophy of Science 9*: 181-206.

Brandon, Robert 1990: *Adaptation and Environment*. Princeton University Press. Princeton.

Brandon, Robert & Carson, Scott 1996: 'The Indeterministic Character of Evolutionary Theory: No "Hidden Variables Proof", but No Room for Determinism Either', *Philosophy of Science 63*: 315-337.

Brante, Thomas 1994: *Vetenskapens sociala grunder - en studie av konflikter i forskarvärlden*. Rabén & Sjögren. Kristianstad.

Bratman, Michael E. 1987: *Intention, Plans, and Practical Reason*. Harvard University Press. Cambridge.

Brown, James Robert 1989: *The Rational and The Social*. Routledge. London.

Bunzl, Martin 1979: 'Causal Overdetermination', *The Journal of Philosophy 76*: 134-150.

Burian, Richard 1983: '"Adaptation"', in Grene (ed.) *Dimensions of Darwinism*. Cambridge University Press. Cambridge: 287-314.

Callon, Michel 1995: 'Four Models for the Dynamics of Science', in Jasanoff, Markle, Petersen & Pinch (eds.): *Handbook of Science and Technology Studies*. Sage. Albany: 29-63.

Callon, Michel and Law, John 1982: 'On Interests and their Transformation: Enrolment and Counter-Enrolment', *Social Studies of Science 12*: 615-625.

Campbell, Colin 1982: 'A Dubious Distinction? An Inquiry into the Value and Use of Merton's Concepts of Manifest and Latent Function', *American Sociological Review 47*: 29-44.

Carroll, John W. 1997: 'Lipton on compatible contrasts', *Analysis 57*: 170-178.

Carroll, John W. 1999: 'The Two Dams and That Damned Paresis', *The British Journal for the Philosophy of Science 50*: 65-81.

Chalmers, Alan 1990: *Science and its Fabrication*. Open University Press. Milton Keynes.

Churchland, Paul M. 1991: 'Folk psychology and the explanation of human behavior', in Greenwood (ed.) *The future of folk psychology*. Cambridge University Press. Cambridge: 51-69.

Clark, Andy 1997: 'Economic Reason: The Interplay of Individual Learning and External Structure', in Drobak & Nye (eds.): *The Frontiers of the New Institutional Economics*. Academic Press. New York: 269-290.

Coady, C. A. J. 1992: *Testimony. A Philosophical Study*. Claredon Press. Oxford.

Cohen, G. A. 1978: *Karl Marx's Theory of History: A Defence*. Princeton University Press. Princeton.

Cohen, G. A. 1982: 'Functional Explanation, Consequence Explanation, and Marxism', *Inquiry 25*: 27-56.

Collins, H. M. 1981: 'What is TRASP? The Radical Programme as a Methodological Imperative', *Philosophy of the Social Sciences 11*: 215-224.

Collins, H. M. 1998: 'The Meaning of Data: Open and Closed Evidential Cultures in the Search for Gravitational Waves', *American Journal of Sociology 104*: 293-338.

Collins, H. M. 1999: 'Tantalus and the Aliens: Publications, Audiences and the Search for Gravitational Waves', *Social Studies of Science 29*: 163-197.

Conlisk, John 1996: 'Why Bounded Rationality?', *Journal of Economic Literature 34*: 669-701.

Crane, T. & Mellor D. H. 1990: 'There is no Question of Physicalism', *Mind 99*: 185-206.

Cross, Charles B. 1991: 'Explanation and the Theory of Questions', *Erkenntnis 34*: 237-260.

Cummins, Robert 1983: *The Nature of Psychological Explanation*. MIT Press. Cambridge.

Cummins, Robert 2000: '"How Does It Work?" versus "What Are the Laws?": Two Conceptions of Psychological Explanation', in Keil & Wilson (eds.): *Explanation and Cognition*. MIT Press. Cambridge: 117-144.

Davies, Paul Sheldon 1994: 'Troubles For Direct Proper Functions', *Noûs 28*: 363-381.

Davis, Wayne 1988: 'Probabilistic Theories of Causation', in Fetzer (ed.): *Probability and Causality*. D. Reidel. Dordrecht: 133-160.

Diamond, Arthur M. Jr. 1996: 'The Economics of Science', *Knowledge & Policy 9* (Issue 2/3): 6-51.

Dowe, Phil 2000: *Physical Causation*. Cambridge University Press. Cambridge.

Earl, Peter E. 1983: 'A Behavioral Theory of Economists' Behavior', in Eichner (ed.): *Why Economics is not yet a Science*. Macmillan. London: 90-125.

Edge, David 1990: 'Competition in Modern Science', in Frängsmyr (ed.): *Solomon's House Revisited*. Science History Publications. Canton MA: 208-232.

Egan, Frances 1995: 'Folk Psychology and Cognitive Architecture', *Philosophy of Science 62*: 179-196.

Eells, Ellery 1991: *Probabilistic Causality*. Cambridge University Press. Cambridge.

Elster, Jon 1979: *Ulysses and the Sirens*. Cambridge University Press. Cambridge.

Elster, Jon 1982: 'Marxism, Functionalism, and Game Theory', *Theory and Society 11*: 453-482.

Elster, Jon 1983: *Explaining Technical Change*. Cambridge University Press. Cambridge.

Elster, Jon 1989: *Nuts and Bolts for the Social Sciences*. Cambridge University Press. Cambridge.

Elster, Jon 1990: 'Merton's Functionalism and the Unintended Consequences of Action', in Clark, Modgil & Modgil (eds.): *Robert K. Merton. Consensus and Controversy*. Falmer Press. London: 129-135.

Elster, Jon 1999: *Alchemies of the Mind*. Cambridge University Press. Cambridge.

Endler, John A. 1984: *Natural Selection in the Wild*. Princeton University Press. Princeton.

Engel, Pascal 1998: 'Believing, holding true, and accepting', *Philosophical Explorations 2*: 140-151.

Ettinger, Lia, Jablonka, Eva & McLaughlin Peter 1990: 'On the Adaptations of Organisms and the Fitness of Types', *Philosophy of Science 57*: 499-513.

Faust, David 1984: *The Limits of Scientific Reasoning*. University of Minnesota Press. Minneapolis.

Faye, Jan 1999: 'Explanation Explained', *Synthese 120*: 61-75.

Feinberg, Joel 1984: *Harm to Others*. Oxford University Press. Oxford.

Friedman, Michael 1974: 'Explanation and Scientific Understanding', *The Journal of Philosophy 71*: 5-19.

Friedman (ed.)1996 : *The Rational Choice Controversy*. Yale University Press. New Haven.

Fuchs, Stephan 1992: *The Professional Quest for Truth. A Social Theory of Science and Knowledge*. SUNY. New York.

Fuller, Steve 1993: 'Critical notice: David Bloor's Knowledge and Social Imagery', *Philosophy of Science 60*: 158-170.

Garfinkel, Alan 1981: *Forms of Explanation*. Yale University Press. New Haven.

Geison, Gerald L. 1995: *The Private Science of Louis Pasteur*. Princeton University Press. Princeton.

Giere, Ronald N. 1988: *Explaining Science. A Cognitive Approach*. The University of Chicago Press. Chicago.

Gieryn, Thomas P. 1999: *Cultural Boundaries of Science. Credibility on the line. The University of Chicago Press*. Chicago.

Gilbert, Nigel 1976: 'The Development of Science and Scientific Knowledge: the Case of Radar Meteor Research', in Lemaine, MacLeod, Mulkay & Weingart (eds.): *Perspectives on the Emergence of Scientific Disciplines*. Mouton. The Hague: 187-204.

Gilbert, Nigel 1977: 'Competition, differentiation and careers in science', *Social Science Information 16*: 103-123.

Gilbert, Nigel & Mulkay, Michael 1984: *Opening Pandora's Box. A sociological analysis of scientists' discourse*. Cambridge University Press. Cambridge.

Glennan, Stuart 1996: 'Mechanisms, and the Nature of Causation', *Erkenntnis 44*: 49-71.

Glymour, Bruce 1998: 'Contrastive, Non-Probabilistic Statistical Explanations', *Philosophy of Science 65*: 448-471.

Godfrey-Smith, Peter 1993: 'Functions: Consensus Without Unity', *Pacific Philosophical Quarterly 74*: 196-208.

Godfrey-Smith, Peter 1994: 'A Modern History Theory of Functions', *Noûs 28*: 344-362.

Goldman, Alvin 1999: *Knowledge in a Social World*. Oxford University Press. Oxford.

Goldthorpe, John H. 1998: 'Rational action theory for sociology', *British Journal of Sociology 49*: 167-192.

Golinski, Jan 1998: *Making Natural Knowledge. Constructivism and the History of Science*. Cambridge University Press. Cambridge.

Graham, George & Horgan, Terence 1988: 'How to be Realistic about Folk Psychology', *Philosophical Psychology 1*: 69-81.

Green, Donald & Shapiro, Ian 1994: *Pathologies of Rational Choice Theory*. Yale University Press. New Haven.

Green, Donald & Shapiro, Ian 1995: 'Pathologies Revisited: Reflections on Our Critics', in Friedman (ed.): *The Rational Choice Controversy*. Yale University Press. New Haven: 235-276.

Greenwood, John D. (ed.) 1991: *The future of folk psychology*. Cambridge University Press. Cambridge.

Grimen, Harald 1994: 'Causally inefficient knowledge and functional explanation', *Social Science Information 33*: 117-127.

Grimes, Thomas R. 1993: 'Explanatory Understanding and Contrastive Facts', *Philosophica 51*: 21-38.

Gärdenfors, Peter 1980: 'A Pragmatic Approach to Explanations', *Philosophy of Science 47*: 404-423.

Habermas, Jürgen 1972: *Knowledge and Human Interests*. Heinemann. London.

Hacking, Ian 1999: *The Social Construction of What?*. Harvard University Press. Cambridge.

Hansson, Bengt 1975: 'Explanations – Of What?', unpublished manuscript.

Hardin, Russell 1980: 'Rationality, irrationality and functionalist explanation', *Social Science Information 19*: 755-772.

Hardwig, John 1985: 'Epistemic Dependence', *The Journal of Philosophy 82*: 335-349.

Hardwig, John 1991: 'The Role of Trust in Knowledge', *The Journal of Philosophy 88*: 693-708.

Hart, H. L. A. & Honoré, A. M. 1959: *Causation in the Law*. Claredon Press. Oxford.

Hausman, Daniel 1998: *Causal Asymmetries*. Cambridge University Press. Cambridge.

Hechter, Michael & Kanazawa, Satoshi 1997: 'Sociological Rational Choice Theory', *Annual Review of Sociology 23*: 191-214.

Hedström, Peter, Swedberg, Richard & Udehn, Lars 1998: 'Popper's Situational Analysis and Contemporary Sociology', *Philosophy of the Social Sciences 28*: 339-365.

Hedström, Peter & Swedberg, Richard (eds.) 1998: *Social Mechanisms*. Cambridge University Press.

Heil, John & Mele, Alfred (eds.) 1993: *Mental Causation*. Claredon Press. Oxford.

Helm, Paul 1971: 'Manifest and Latent Functions', *Philosophical Quarterly 21*: 51-60.

Hempel, Carl 1965: *Aspects of Scientific Explanation*. The Free Press. New York.

Henderson, David K. 1993: *Interpretation and Explanation in the Human Sciences*. SUNY. New York.

Hesslow, Germund 1983: 'Explaining differences and weighting causes', *Theoria 49*: 87-111.

Hilton, Denis J. 1995: 'Logic and language in causal explanation', in Sperber, Premack & Premack (eds.): *Causal Cognition*. Oxford University Press, Oxford: 495-525.

Hindess, Barry 1986: ''Interests' in political analysis', *Sociological Review Monograph 32*: 112-131.

Hintikka, Jaakko & Halonen, Ilpo 1995: 'Semantics and Pragmatics for Why-Questions', *The Journal of Philosophy 92*: 636-657.

Hintikka, Jaakko & Halonen, Ilpo 2001: 'Toward a Theory of the Process of Explanation', *Synthese*, forthcoming.

Hirschman, Albert O. 1977: *The Passions and the Interests. Political Arguments for Capitalism before Its Triumph*. Princeton University Press. Princeton.

Hitchcock, Christopher C. 1995: ' Discussion: Salmon on Explanatory Relevance', *Philosophy of Science 62*: 304-320.

Hitchcock, Christopher C. 1996: 'The Role of Contrast in Causal and Explanatory Claims', *Synthese 107*: 395-419.

Hitchcock, Christopher C. 1999: 'Contrastive Explanation and the Demons of Determinism', *The British Journal for the Philosophy of Science 50*: 585-612.

Horan, Barbara 1994: 'The Statistical Character of Evolutionary Theory'. *Philosophy of Science 61*: 76-95.

Horgan, Terence 1989: 'Mental Quasation', *Philosophical Perspectives 3*: 47-76.

Horgan, Terence 1993: 'From Supervenience to Superdupervenience: Meeting the Demands of a Material World', *Mind 102*: 555-586.

Horgan, Terence & Woodward, James 1991: 'Folk psychology is here to stay', in Greenwood (ed.) *The future of folk psychology*. Cambridge University Press. Cambridge: 149-175.

Hull, David L. 1988: *Science as a Process. An Evolutionary Account of the Social and Conceptual Development of Science*. The University of Chicago Press. Chicago.

Humphreys, Paul 1989: *The Chances of Explanation*. Princeton University Press. Princeton.

Hållsten, Henrik 1999: 'Deductive Chauvinism', *Synthese 120*: 49-59.

Jackson, Frank & Pettit, Philip 1988: 'Functionalism and Broad Content', *Mind* 97: 381-400.

Jackson, Frank & Pettit, Philip 1990a: 'Program Explanation: A General Perspective', *Analysis* 50: 107-117.

Jackson, Frank & Pettit, Philip 1990b: 'Causation in the Philosophy of Mind', *Philosophy and Phenomenological Research 50*: 195-214.

Jackson, Frank & Pettit, Philip 1990c: 'In Defence of Folk Psychology', *Philosophical Studies 59*: 31-54.

Jackson, Frank & Pettit, Philip 1992a: 'Structural Explanation and Social Theory', in Charles & Lennon (eds.): *Reductionism and Anti-reductionism*. Oxford University Press. Oxford: 97-131.

Jackson, Frank & Pettit, Philip 1992b: 'In Defense of Explanatory Ecumenism', *Economics and Philosophy 8*: 1-21.

Jardine, Nicholas 1991: *The Scenes of Inquiry. On the Reality of Questions in the Sciences*. Claredon Press. Oxford.

Kahneman, D., Slovic, P. & Tversky, A. (eds.) 1982: *Judgement under uncertainty: Heuristics and biases*. Cambridge University Press. Cambridge.

Kim, Jaegwon 1973: 'Causes and Counterfactuals', *The Journal of Philosophy 70*: 570-572.

Kim, Jaegwon 1993: *Supervenience and Mind*. Cambridge University Press. Cambridge.

Kim, Jaegwon 1998: *Mind in a Physical World*. MIT Press. Cambridge.

Kim, Kyung-Man 1994: *Explaining Scientific Consensus. The Case of Mendelian Genetics*. The Guilford Press. New York.

Kincaid, Harold 1996: *Philosophical Foundations of the Social Sciences. Analyzing Controversies in Social Research*. Cambridge University Press. Cambridge.

Kitcher, Philip 1989: 'Explanatory Unification and the Causal Structure of the World', in Kitcher & Salmon (eds.): *Scientific Explanation. Minnesota Studies in the Philosophy of Science vol XIII*. University of Minnesota Press. Minneapolis: 410-505.

Kitcher, Philip 1993: *The Advancement of Science. Science without Legend, Objectivity without Illusions*. Oxford University Press. Oxford.

Knorr–Cetina, Karin 1981: *The Manufacture of Knowledge. An Essay on the Constructivist and Contextual Nature of Science*. Pergamon Press. Oxford.

Knorr-Cetina, Karin 1982: 'Scientific Communities or Transepistemic Arenas of Research? A Critique of Quasi-Economic Models of Science', *Social Studies of Science 12*: 101-130.

Knorr-Cetina, Karin 1987: 'Evolutionary Epistemology and Sociology of Science', in Callebaut & Pinxten (eds.) *Evolutionary Epistemology*. D. Reidel. Dordrecht: 179-201.

Kohler, Robert E. 1994: *Lords of the Fly. Drosophila Genetics and the Experimental Life*. The University of Chicago Press. Chicago.

Koura, Antti 1988: 'An Approach to Why-Questions', *Synthese 74*: 191-206.

Kripke, Saul 1982: *Wittgenstein on Rules and Private Language*. Oxford University Press. Oxford.

Latour, Bruno 1987: *Science in Action*. Open University Press. Milton Keynes.

Latour, Bruno 1988a: *The Pasteurization of France*. Harvard University Press. Cambridge.

Latour, Bruno 1988b: 'The Politics of Explanation: An Alternative', in Woolgar (ed.): *Knowledge and Reflexivity. New Frontiers in the Sociology of Knowledge*. Sage. Albany: 155-176.

Latour, Bruno 1993: *La clef de berlin et autres leçons d'un amateur de sciences*. Éditions la Découverte. Paris.

Latour, Bruno & Woolgar, Steve 1986: *Laboratory Life. The Construction of Scientific Facts*. (2 nd ed). Princeton University Press. Princeton.

Latsis, Spiro 1972: 'Situational Determinism in Economics', *The British Journal for the Philosophy of Science 23*: 207-245.

Laudan, Larry 1977: *Progress and Its Problems*. Routledge & Kegan Paul. London.

Laudan, Larry 1981: 'The Pseudo-Science of Science?', *Philosophy of the Social Sciences 11*: 173-198.

LeGrand, H. E. 1988: *Drifting continents and shifting theories*. Cambridge University Press. Cambridge.

Lenoir, Timothy 1997: *Instituting Science. The Cultural Production of Scientific Disciplines*. Stanford University Press. Stanford.

Lewis, David 1969: *Convention*. Harvard University Press. Cambridge.

Lewis, David 1983: *Philosophical Papers vol I*. Oxford University Press. Oxford.

Lewis, David 1986: *Philosophical Papers vol II*. Oxford University Press. Oxford.

Lewis, David 2000: 'Causation as Influence', *The Journal of Philosophy 97*: 182-197.

Lipton, Peter 1990: 'Contrastive Explanations', in Knowles (ed.): *Explanation and its Limits*. Cambridge University Press. Cambridge: 247-266.

Lipton, Peter 1991: *Inference to the Best Explanation*. Routledge. London.

Lipton, Peter 1993: 'Making a Difference', *Philosophica 51*: 39-54.

Lukes, Steven 1974: *Power: A Radical View*. Macmillan. London.

Lynch, Michael 1993: *Scientific practice and ordinary action. Ethnomethodology and social studies of science*. Cambridge University Press. Cambridge.

McDermott, Michael 1995: 'Redundant Causation', *British Journal for the Philosophy of Science 46*: 523-544.

McIntyre, Lee C. 1996: *Laws and Explanation in the Social Sciences*. Westview Press. Boulder.

Machamer, P., Darden, L. & Craver, C. F. 2000: 'Thinking About Mechanisms', *Philosophy of Science 67*: 1-25.

MacKenzie, Donald 1978: 'Statistical Theory and Social Interests: A Case Study', *Social Studies of Science 8*: 35-84.

MacKenzie, Donald 1981: *Statistics in Britain 1865-1930. The Social Construction of Scientific Knowledge*. Edinburgh University Press. Edinburgh.

MacKenzie, Donald 1990: *Inventing Accuracy. A Historical Sociology of Nuclear Missile Guidance*. MIT Press. Cambridge.

MacKenzie, Donald 1999: 'The Zero-Sum Assumption', *Social Studies of Science 29*: 223-234.

MacKenzie, Donald & Barnes, Barry 1979: 'Scientific Judgment: The Biometry-Mendelism Controversy', in Barnes & Shapin (eds.): *Natural Order. Historical Studies of Scientific Culture*. Sage. Beverly Hills: 191-210.

Mackie, J. L. 1974: *The Cement of the Universe*. Claredon Press. Oxford.

Margolis, Howard 1988: *Patterns, Thinking, and Cognition : A Theory of Judgment*. The University of Chicago Press. Chicago.

Margolis, Howard 1993: *Paradigms & Barriers. How Habits of Mind Govern Scientific Beliefs*. The University of Chicago Press. Chicago.

Markwick, P. 1999: 'Interrogatives and Contrasts in Explanation Theory', *Philosophical Studies 96*: 183-204.

McCarthy, Timothy 1993: 'On an Aristotelian Model of Scientific Explanation', in Ruben (ed.): *Explanation*. Oxford University Press. Oxford: 128-135.

Mellor, Hugh 1995: *The Facts of Causation*. Routledge. London.

Mendelsohn, Everett 1987: 'The political anatomy of controversy in the sciences', in Engelhardt & Caplan (eds.): *Scientific controversies*. Cambridge University Press. Cambridge: 93-125.

Menzies, Peter 1988: 'Against Causal Reductionism', *Mind 97*: 551-574.

Menzies, Peter & Price, Huw 1993: 'Causation as a Secondary Quality', *British Journal for the Philosophy of Science 44*: 187-203.

Merton, Robert K. 1967: 'Manifest and Latent Functions', *On Theoretical Sociology*. The Free Press. New York: 73-138.

Mill, John S. 1906: *System of Logic*. (8 th. ed.) Longmans, Green & Co. London.

Miller, Richard W. 1978: 'Methodological Individualism and Social Explanation', *Philosophy of Science 45*: 387-414.

Miller, Richard W. 1987: *Fact and Method. Explanation, Confirmation and Reality in the Natural and the Social Sciences*. Princeton University Press. Princeton.

Millikan, Ruth Garrett 1989: 'An Ambiguity in the Notion "Function"', *Biology and Philosophy 4*: 172-176.

Mills, Susan & Beatty, John 1994: 'The Propensity Interpretation of Fitness', in Sober (ed.): *Conceptual Issues in Evolutionary Biology. 2nd ed*. MIT Press. Cambridge: 3-27.

Myers, Milton L. 1983: *The Soul of Modern Economic Man. Ideas of Self-Interest Thomas Hobbes to Adam Smith*. The University of Chicago Press. Chicago.

Mäki, Uskali 1993: 'Social Theories of Science and the Fate of Institutionalism in Economics', in Mäki, Gustafsson & Knudsen (eds.) *Rationality, Institutions and Economic Methodology*. Routledge. London: 76-109.

Newton-Smith, William 1985: 'The Role of Interests in Science', in Griffiths (ed.): *Philosophy and Practice*. Cambridge University Press. Cambridge: 59-73.

Niiniluoto, Ilkka 1991: 'Realism, Relativism, and Comstructivism', *Synthese 89*: 135-162.

Nozick, Robert 1974: *Anarchy, State, and Utopia*. Basil Blackwell. Oxford.

Peillon, Michel 1990: *The Concept of Interest in Social Theory*. The Edwin Mellen Press. Lewinston.

Percival, Philip 2000: 'Lewis's Dilemma of Explanation under Indeterminism Exposed and Resolved', *Mind 109*: 39-66.

Pettit, Philip 1992: 'The Nature of Naturalism II'. *The Aristotelian Society. Supplementary Volume* LXVI. The Aristotelian Society: 245-266.

Pettit, Philip 1993a: *The Common Mind. An Essay on Psychology, Society and Politics*. Oxford University Press. Oxford.

Pettit, Philip 1993b: 'A definition of physicalism', *Analysis* 53: 213-223.

Pettit, Philip 1995a: 'The Virtual Reality of Homo Economicus', *The Monist 78*: 308-329.

Pettit, Philip 1995b: 'Causality at higher levels', in Sperber, Premack & Premack (eds.): *Causal Cognition*. Oxford University Press, Oxford: 399-421.

Pettit, Philip 2000: 'Rational choice, functional selection and empty black boxes', *Journal of Economic Methodology 7*: 33-57.

Pickering, Andrew 1980: 'The Role of Interests in High-Energy Physics: The Choice Between Charm And Colour', in Knorr, Krohn & Whitley (eds.) *The Social Process of Scientific Investigation*. D. Reidel. Dordrecht: 107-138.

Pickering, Andrew 1984: *Constructing Quarks. A Sociological History of Particle Physics*. Edinburgh University Press. Edinburgh.

Pickering, Andrew 1995: *The Mangle of Practice. Time, Agency & Science*. The University of Chicago Press. Chicago.

Pollner, Melvin 1974: 'Mundane Reasoning', *Philosophy of the Social Sciences 4*: 35-54.

Proctor, Robert N. 1995: *Cancer Wars*. Basic Books. New York.

Railton, Peter 1978: 'A Deductive-Nomological Model of Probabilistic Explanation', *Philosophy of Science 45*: 206-226.

Railton, Peter 1980: *Explaining Explanation: A Realist Account of Scientific Explanation*. Ph.D. thesis. Princeton University.

Railton, Peter 1981: 'Probability, Explanation, and Information', *Synthese 48*: 233-256.

Reeve, Andrew & Ware, Alan 1984: 'Interests in Political Theory', *British Journal of Political Science 13*: 379-400.

Resnik, David 1998: 'Conflicts of Interest in Science', *Perspectives on Science 6*: 381-408.

Richardson, Robert & Burian, Richard 1992: 'A Defence of Propensity Interpretations of Fitness', in Hull, Forber & Okruhlik (eds.) *PSA 1992 vol. 1*: 349-362.

Risjord, Mark W. 2000: *Woodcutters and Witchcraft. Rationality and Interpretive Change in the Social Sciences*. SUNY. Albany.

Rosenberg, Alexander 1978: 'The supervience of biological concepts'. *Philosophy of Science 45*: 368-386.

Rosenberg, Alexander 1985: *The Structure of Biological Science*. Cambridge University Press. Cambridge.

Rosenberg, Alexander 1994: *Instrumental Biology or the Disunity of Science*. The University of Chicago Press. Chicago.

Roth, Paul A. 1987: *Meaning and Method in the Social Sciences. A Case for Methodological Pluralism*. Cornell University Press. Ithaca.

Roth, Paul A. 1996: 'Will the Real Scientists Please Stand Up? Dead Ends and Live Issues in the Explanation of Scientific Knowledge', *Studies in the History and Philosophy of Science 27*: 43-68.

Ruben, David-Hillel 1985: *The Metaphysics of the Social World*. Routledge. London.

Ruben, David-Hillel 1990: *Explaining Explanation*. Routledge. London.

Ruben, David-Hillel 1994: 'A Counterfactual Theory of Causal Explanation', *Noûs 28*: 465-481.

Rudwick, Martin 1985: *The Great Devonian Controversy*. The University of Chicago Press. Chicago.

Salmon, Wesley 1984: *Scientific Explanation and the Causal Structure of the World*. Princeton University Press. Princeton.

Salmon, Wesley 1989: *Four Decades of Scientific Explanation*. University of Minnesota Press. Minneapolis.

Salmon, Wesley 1997: 'Causality and Explanation: A Reply to Two Critique', *Philosophy of Science 64*: 461-477.

Salmon, Wesley 1998: *Causality and Explanation*. Oxford University Press. Oxford.

Sapp, Jan 1987: *Beyond the Gene. Cytoplasmic Inheritance and the Struggle for Authority in Genetics*. Oxford University Press. Oxford.

Satz, Debra & Ferejohn, John 1994: 'Rational Choice and Social Theory', *The Journal of Philosophy 91*: 71-87.

Schiffer, Stephen 1991: 'Ceteris Paribus Laws', *Mind 100*: 1-17.

Schmaus, Warren 1992: 'Research programs as intellectual niches', *Social Epistemology 6*: 13-22.

Scriven, Michael 1959: 'Truisms as the Grounds fo Historical Explanations', in Gardiner (ed.): *Theories of History*. The Free Press. New York: 443-475.

Scriven, Michael 1962: 'Explanations, Predictions, and Laws', in Feigl & Maxwell (eds.): *Scientific Explanation, Space, and Time. Minnesota Studies in the Philosophy of Science vol III*. University of Minnesota Press. Minneapolis: 170-230.

Scriven, Michael 1975: 'Causation as Explanation', *Noûs 9*: 3-16.

Searle, John R. 1995: *The Construction of Social Reality*. Penguin Books. New York.

Segerstråle, Ullica 2000: *Defenders of the Truth*. Oxford University Press. Oxford.

Settle, Tom 1993: ''Fitness' and 'Altruism': Traps for the Unwary, Bystander and Biologist Alike', *Biology & Philosophy 8*: 61-83.

Schaffer, Jonathan 2000: 'Causation by Disconnection', *Philosophy of Science 67*: 285-300.

Shapin, Steven 1979a: 'Homo Phrenologicus: Anthropological Perspectives on an Historical Problem', in Barnes & Shapin (eds.): *Natural Order. Historical Studies of Scientific Culture*. Sage. Beverly Hills: 41-71.

Shapin, Steven 1979b: 'The Politics of Observation: Cerebral Anatomy and Social Interests in the Edinburgh Phrenology Disputes', *Sociological Review Monograph 27*: 139-178.

Shapin, Steven 1986: 'History of Science and Its Sociological Reconstructions', in Cohen & Schnelle (eds.): *Cognition and Fact. Materials on Ludwik Fleck*. D. Reidel. Dordrecht: 325-386.

Shapin, Steven 1988: 'Following Scientists Around', *Social Studies of Science 18*: 533-550.

Shapin, Steven 1992: 'Discipline and Bounding: The History and Sociology of Science as Seen Through the Externalism-Internalism Debate', *History of Science 30*: 333-369.

Shapin, Steven 1994: *Social History of Truth. Civility and Science in Seventeenth-Century England*. The University of Chicago Press. Chicago.

Shapin, Steven 1995a: 'Cordelia's Love: Credibility and the Social Studies of Science', *Perspectives on Science 3*: 255-275.

Shapin, Steven 1995b: 'Here and Everywhere: Sociology of Scientific Knowledge', *Annual Review of Sociology 21*: 289-321.

Shapin, Steven & Schaffer, Simon 1985: *Leviathan and the Air-Pump. Hobbes, Boyle, and the Experimental Life*. Princeton University Press. Princeton.

Simon, Herbert A. 1982: *Models of Bounded Rationality, vol. 2*. MIT Press. Cambridge.

Sintonen, Matti 1984: *The Pragmatics of Scientific Explanation*. Acta Philosophica Fennica 37. Societas Philosophica Fennica. Helsinki.

Sintonen, Matti 1993: 'In Search of Explanations: from Why-questions to Shakespearean Questions', *Philosophica 51*: 55-82.

Sintonen, Matti & Kiikeri, Mika 1994: 'Idealization in Evolutionary Biology', in Kuokkanen (ed.): *Idealization VII: Structuralism, Idealization and Approximation. Poznan Studies in the Philosophy of the Sciences and the Humanities 42*: 201-216.

Sober, Elliot 1984: *The Nature of Selection*. MIT Press. Cambridge.

Sober, Elliot 1993: *Philosophy of Biology*. Oxford University Press. Oxford.

Sober, Elliot 1994: *From a Biological Point of View*. Cambridge University Press. Cambridge.

Sober, Elliot 1999: 'The Multiple Realizability Argument Against Reductionism', *Philosophy of Science 66*: 542-564.

Sperber, Dan 1996: *Explaining Culture. A Naturalist Approach*. Blackwell. Oxford.

Stephan, Paula E. 1996: 'The Economics of Science', *Journal of Economic Literature 34*: 1199-1235.

Stewart, John A. 1990: *Drifting Continents and Colliding Paradigms: Perspectives on the Geoscience Revolution*. Indiana University Press. Bloomington.

Strawson, Galen 1989: *The Secret Connexion. Causation, Realism, and David Hume*. Claredon Press. Oxford.

Strawson, P. F. 1985: 'Causation and Explanation', in Vermazen & Hintikka (eds.): *Essays on Davidson: Actions and Events*. Claredon Press. Oxford: 115-136.

Sztompka, Piotr 1986: *Robert K. Merton. An intellectual profile*. Macmillan. London.

Tammi, Timo 1999: 'On Experimental Discourse in Economics', *Philosophy of the Social Sciences 29*: 62-89.

Temple, Dennis 1988: 'The Contrast Theory of Why-Questions', *Philosophy of Science 55*: 141-151.

Thagard, Paul 1999: *How Scientists Explain Disease*. Princeton University Press. Princeton.

Toulmin, Stephen 1961: *Foresight and Understanding*. Hutchinson. London.

Tuomela, Raimo 1977: *Human Action and its Explanation*. Kluwer. Dordrecht.

Tuomela, Raimo 1980: 'Explaining Explaining', *Erkenntnis 15*: 211-243.

Tuomela, Raimo 1994: 'In Search for the Common Mind: A Critical Notice of Philip Pettit's *The Common Mind*', *International Journal of Philosophical Studies 2*: 306-321.

Tuomela, Raimo 1995: *The Importance of Us*. Stanford University Press. Stanford.

Tuomela, Raimo 1998: 'A Defence of Mental Causation', *Philosophical Studies 90*: 1-34.

Turner, Stephen 1994: *The Social Theory of Practices*. Polity Press. Cambridge.

Turner, Stephen 1996: 'Directions for Future Research', *Knowledge & Policy 9* (Issue 2/3): 98-106.

van Fraassen, Bas 1980: *Scientific Image*. Oxford University Press. Oxford.

van Parijs, Philippe 1981: *Evolutionary Explanation in the Social Sciences. An Emerging Paradigm*. Rowman & Littlefield. Totowa.

von Wright, Georg Henrik 1971: *Explanation and Understanding*. Routledge & Kegan Paul. London.

Vromen, Jack 1995: *Economic Evolution*. Routledge. London.

Watkins, John 1984: *Science and Scepticism*. Princeton University Press. Princeton.

Whitley, Richard 1984: *The Intellectual and Social Organization of the Sciences*. Claredon Press. Oxford.

Williams, Rob & Law, John 1980: 'Beyond the Bounds of Credibility', *Fundamentae Scientiae 1*: 295-315.

Winch, Peter 1990: *The Idea of a Social Science and its Relation to Philosophy (2nd. ed.)*. Humanities Press. New Jersey.

Wittgenstein, Ludwig 1953: *Philosophical Investigations*. Blackwell. Oxford.

Woodward, James 1979: 'Scientific Explanation', *The British Journal for the Philosophy of Science 30*: 41-67.

Woodward, James 1986: 'Are Singular Explanations Implicit Covering-Law Explanations', *Canadian Journal of Philosophy 16*: 253-280.

Woodward, James 1989: 'The Causal Mechanical Model of Explanation', in Kitcher & Salmon (eds.): *Scientific Explanation. Minnesota Studies in the Philosophy of Science vol XIII*. University of Minnesota Press. Minneapolis: 357-383.

Woodward, James 1993: 'A Theory of Singular Causal Explanation', in Ruben (ed.): *Explanation*. Oxford University Press. Oxford: 246-274.

Woodward, James 2000: 'Explanation and Invariance in the Special Sciences', *British Journal for the Philosophy of Science 51*: 197-254.

Woolgar, Steve 1981: 'Interests and Explanation in the Social Study of Science', *Social Studies of Science 11*: 365-94.

Wright, Larry 1976: *Teleological Explanations. An Etiological Analysis of Goals and Functions*. California University Press. Berkeley.

Yearley, Steven 1982: 'The Relationship Between Epistemological and Sociological Cognitive Interests: Some Ambiguities Underlying the Use of Interest Theory in the Study of Scientific Knowledge', *Studies in the History and Philosophy of Science 13*: 353-388.

Yearley, Steven 1984: *Science and Sociological Practice*. Open University Press. Milton Keynes.

Ylikoski, Petri 1995: 'The Invisible Hand and Science', *Science Studies 2/1995*: 32-43.

Ylikoski, Petri 1997: *Dispositioiden ontologia ja selittäminen*. Licenciate thesis. August 1997. (URN:NBN:fi-fe20001094) [http://ethesis.helsinki.fi/julkaisut/val/kayta/lt/ylikoski/]

Ylikoski, Petri 1998: 'Does the Invisible Hand Save the Legend?', in Lewicka-Strzalecka & Loukola (eds.): *Science in Society: Science Policy and Ethics*. IFiS Publishers. Warsaw: 97-118.