99
9

Prospects for a Cognitive Science of Science

by

Stephen Matthew Downes

Dissertation submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

PhD

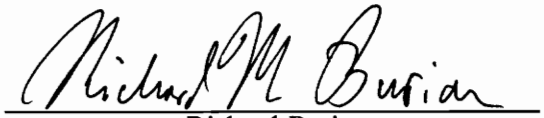
in

Science and Technology Studies

APPROVED:

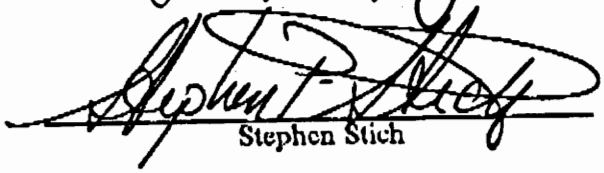

Steve Fuller, Chairman


Peter Barker


Richard Burian


Gary Downey


Robert Paterson


Stephen Stich

July, 1990

Blacksburg, Virginia

Prospects for a Cognitive Science of Science

by

Stephen Matthew Downes

Steve Fuller, Chairman

Science and Technology Studies

(ABSTRACT)

Cognitive science of science attempts to explain a range of phenomena familiar to philosophers of science, such as theory choice and scientific discovery. The appeal to cognitive science may be seen as an attempt to naturalize the philosophy of science. I examine and criticize several of the most important contributions to this new field. I argue that an unrecognized common defect of this work is its reliance on an explanatory approach that takes individuals' cognitive capacities as its units of analysis. I introduce the term "cognitive individualism" to identify this position, and conclude by examining the position in detail and sketching alternative approaches to naturalizing philosophy stressing the social dimensions of science.

In Chapter 1 I briefly describe the field of cognitive science, and outline the empirical resources it can provide a philosopher of science. I then outline key themes of current cognitive science of science. In the next four chapters I critically examine the work of four prominent cognitive scientists of science: Herbert Simon, Paul Thagard, Ronald Giere, and Paul Churchland. All share the same goals of naturalizing philosophy of science by using the empirical resources of cognitive science. I show that all four accounts ignore the important social nature of science, and share an adherence to cognitive individualism.

In the final chapter I develop the notion of cognitive individualism in detail. I show that relying on empirical evidence from work in cognitive psychology on human judgment, and work in the sociology of science is more fruitful for explaining science than the current cognitive individualist approach. I conclude with several theses that act as guidelines for future research in naturalized philosophy of science.

Acknowledgements

I want to thank all those people who in one way or another helped me to put this dissertation together. I have worked with Steve Fuller for five years now and he has always been encouraging and enthusiastic about my ideas, fortunately he retained his enthusiasm while they made their tortuous way to this final written form. Four of my other committee members were from the Science and Technology Studies Center at VPI, which has provided a lively environment for my last two years of study. Peter Barker, Richard Burian, and Gary Downey have all provided lengthy comments on various drafts, and contributed much of their valuable time to discussing the issues raised. Robert Paterson provided both scholarly and financial support during my stay at VPI, and proved a good ally when I attempted to defend my views to members of other departments here. Stephen Stich has supported me in my efforts to become a professional philosopher for some time now, and discussions with him over the years have always been constructive and useful to achieving that end. Ronald Giere and Thomas Nickles both gave their support to my dissertation project from the outset. I have benefitted from the opportunity to discuss the views expressed in this dissertation with all of the following people: Skip Fuhrman, Joseph Pitt, Roger Ariew, Moti Feingold, Eric Dietrich, Howard Smokler, Michael Gorman, Jeffrey Shrager, William Lynch, Adam Serchuk, Andrea Burrows, Charles Wallis, and Georg Schwartz. Thanks go to my mother and father, my sister Jane and my brother Timothy for all their support and encouragement from afar.

Finally, thanks to Hilary Hart for being there throughout the whole process, and being the major contributor to my enjoyable life outside the academy.

Table of Contents

CHAPTER 1

GENERAL INTRODUCTION

1.Introduction.

In this dissertation I argue that cognitive science of science is able to provide neither adequate descriptions of scientific practice, nor an adequate naturalized philosophy of science. One important reason for this inadequacy is the adoption of what I call cognitive individualism. A cognitive individualist account seeks an explanation of a cognitive phenomenon purely in terms of an individual's psychological processes, understood as internal mechanisms. In an explanation of science this approach is crucially limiting, because it ignores the important social dimension of scientific cognitive activity. I propose that a descriptively adequate and yet normative naturalized philosophy of science should make use of empirical work from the sociology of science as well as cognitive science, and would consequently provide norms that would be applicable to science as it is practiced now.

The cognitive science of science applies techniques from the cognitive sciences to issues previously examined by the philosophy of science, for example, theory change, theory evaluation, the nature of scientific theories, and scientific discovery. In general, cognitive science of science researchers set their work apart from more traditional approaches to philosophy of science such as logical empiricism. They are naturalists, who hold that empirical results from current science should inform and constrain philosophical theories. Some naturalists go further suggesting that philosophy be abandoned and replaced by a particular science, so some philosophers suggest that cognitive science of science should replace philosophy of science.

I examine the work of four cognitive science of science researchers: Herbert Simon, Paul Thagard, Ronald Giere and Paul Churchland. All have made important contributions to cognitive science of science and each takes a different theoretical stance, yet all are cognitive individualists. In the following four Chapters I evaluate each of their approaches on its own merits before defending my more general claim that they are each committed to cognitive individualism. In the final Chapter I flesh out the notion of cognitive individualism and show that it is misguided to use it as the basis for a naturalized philosophy of science. Finally I make some proposals about the direction that an alternative naturalized philosophy of science could take.

One source of the cognitive individualism in cognitive science of science is the need to embody abstract theories, which until recently have been the sole focus of philosophy of science. Traditional philosophy of science gives an account of scientific theories and their explanatory capacities. This involves a concentration on the logical relations between elements of the theory, characterized as sentences. As a result of this rational reconstructivist approach most traditional philosophy of science was neither couched in terms of individual scientists nor social groups, but in terms of theories themselves. This approach is exemplified even in more naturalistic work such as Laudan's (Laudan 1977 Ch.3.). A crucial question in naturalizing the philosophy of science is where to situate theories, or how to embody them. Following this is the question of which current science should be used to assist the naturalized philosopher. Cognitive individualists answer simply that theories are in individual scientists' heads, and that cognitive science gives empirical support for this claim.

The naturalistic turn of the cognitive scientists of science has some parallels with the historical orientation in the philosophy of science. Advocates of the latter argue that rational reconstructivist accounts of science were mistaken, because the theories did not fit with historical data. Case studies in the history of science are presented that challenge various assumptions of rational reconstructivist's theories. For example if a case study demonstrates that the H-D model of explanation was not operative, this provides evidence that such a model is not generally applicable to science. Compare this with the approach of a cognitive scientist of science who may argue that a particular philosophical theory is not generalizable because of what we know about human psychological make-up. Instead of a particular type of argument structure being crucial in a particular case they would propose a certain psychological process.

It may seem a short step from a concern with reasoning, argument, and the application of theories, to an account of individual psychological processing, but this step misses an important set of concerns in philosophy of science. The rational reconstruction of science not only involves reconstructing the work of individual scientists, but also confronts issues such as scientific progress and the accumulation of scientific knowledge, which involve the whole of science, and all of its participants. For example philosophers are concerned with science's ability to converge on the truth. The rational reconstructivist provides an abstract, or idealized account of all scientific activity, not simply an account of the ideal individual scientist. A naturalized philosophy of science must in turn be able to address issues such as the success of science as a whole. I argue in the following Chapters that the move from abstract entities such as theories, to theories embodied in individual's psychological processes, the cognitive individualist move, is mistaken. The move is mistaken primarily as a result of the fact that in traditional philosophy of science the relevant abstractions try to reconstruct the state of a field in relation to external events, plus the relevant community standards of the field, and also some norms to govern reasoning processes. Cognitive individiualists attempt to reconstruct all these features within an individual's internal representational structure. Considerations such as these lead me to argue that the social dimension of science, ignored by the cognitive individualists, provides a key to addressing issues such as the nature of scientific theories.

Many philosophers argue that social factors, and psychological factors, are not their proper concern. One way of setting up this argument is to make a distinction between cognitive^{c} and non - cognitive^{c}[1] factors in science. A parallel distinction is made between internal and external factors in science. On this view, when an account of the success of science is required it should be given in terms of cognitive^{c} factors. An example would be that scientific theories can be determined as approximating the truth, and such an approximation provides the metric of the theory's success. A corresponding non - cognitive^{c} account may explain the success of a theory in terms of its broad acceptance in the scientific community. Opponents of this latter view argue that non - cognitive^{c} explanations should be reserved for use in cases of unsuccessful, or failed scientific theories, and that success in science can only be explained by cognitive^{c} factors (See e.g. Lakatos 1981, Laudan 1977).

The sense of the term "cognitive^{c}" in the above discussion has more in common with the logical empiricists' "cognitive significance" criteria, than with "cognitive" in cognitive science. Ironically, on the above view the cognitive science of science is concerned with non- cognitive^{c} factors in science, specifically psychological processes. One of the issues I confront directly is a further twist in this story. Cognitive science of science researchers have produced a new version of the distinction between cognitive^{c} and non - cognitive^{c} factors. They distinguish between cognitive processes as investigated by cognitive science, and social processes. A similar kind of explanatory hierarchy is set up as in the older version of the distinction, and is central to cognitive individualism.

I present two main lines of argument against cognitive individualism. The first points out the social nature of scientific practice, something that cannot be sufficiently captured by a cognitive individualist account. This is a stronger argument than one which points to distinct categories of social and cognitive^{c} factors in science, and calls for a division of labor between cognitive scientist and sociologists. I argue that attempts by cognitive scientists of science to explain phenomena explicitly involving social interaction solely in terms of individual scientists' psychological processes must fail. The second line of argument rests on empirical evidence that humans are deficient cognitive agents. This evidence, drawn from work in cognitive science on human judgment and

---

[1] The term "cognitive" is used in different senses,and I use the superscript "c" to denote those that are consistent with a more traditional use of "cognitive" such as in the logical empiricists' notion of cognitive significance. The superscript avoids some difficulties in reading the following passages.

reasoning skills, shows that individual human cognitive agents are deficient at relatively simple cognitive tasks. To claim that the explanatory burden for the successes of science rests on such deficient individual agents is misguided.

There are two sets of implications that can be drawn from my arguments. The first set is descriptive, and addresses the descriptive adequacy of cognitive science of science. The main thrust here is that we should not be content with any descriptive account of science that is couched purely in terms of individual scientists' putative psychological processes. The second set is normative and connects with the difficult problem of providing any naturalistic yet normative account. The argument here is centered around the maxim "ought implies can;" if we attempt to derive norms for particular activities from descriptive accounts, the norms must have the activities in question as their range of application. My general conclusions are that cognitive science of science will provide inadequate descriptive accounts of science if it retains cognitive individualism, and second, that a naturalistic philosophy of science with a normative dimension must have a descriptive basis that takes into account the distinctive social nature of science. I now turn to brief descriptions of cognitive science, cognitive science of science, and the common threads in the accounts of the four protagonists.

2. Cognitive Science.

Cognitive science is the cluster of disciplines that appeals to internal representations, generally represented by computer models, to explain phenomena such as language learning, visual perception, memory, problem solving and thinking. These phenomena are studied as they arise in both humans and machines/computers.

The disciplines that contribute most to cognitive science are cognitive psychology, artificial intelligence, linguistics and philosophy. The neurosciences are also strongly represented in the, albeit loosely organized, cognitive science community. Some authors claim that anthropology has a place in cognitive science (Gardner 1987) and some that sociology does (Mandler in Kintsch et

al. eds. 1984), and the debate regarding the proper delineation of the field to some extent echoes the debates over what cognition is (see Gardner 1987, Baars 1986). Questions of what disciplines ought or ought not to be included in cognitive science are to a large extent red herrings in characterizing cognitive science. Yet the issue of the difference between cognitive science and behaviorism is important as it helps to locate more clearly the concerns of cognitive scientists (Baars 1986).

Focusing on phenomena such as memory and learning would not be alien to a behaviorist psychologist, but to attempt to explain these phenomena in terms of internal representations, of some sort or other, is to go beyond behaviorist tenets. Cognitive scientists' claim that internal representations should be used to explain cognitive activity distinguishes them from behaviorists who did not believe that there were any internal representations. The appeal to internal representations has led cognitive scientists to investigate phenomena that behaviorists were not concerned with, such as thinking and problem solving (cf. Baars 1986).

The further claim of cognitive scientists is that computers can be used to model or simulate the internal representations of humans. Some adhere to the stronger claim, that computers and humans in some sense share the same internal representations (see e.g. Pylyshyn 1984). So that a computer simulation of the relevant internal representation is the same as the representation that occurs in humans. The bulk of work in cognitive science involves some combination of computer modelling and appeals to internal processes to explain cognitive phenomena. This leads some cognitive scientists to claim that they need not be involved with the explanation of human behavior, even if it is cognitive behavior. They are concerned entirely with machine behavior, and the designing of machines that perform particular cognitive tasks well.

Cognitive science aims to bring as many phenomena as possible under one broad explanatory theory and the current one is *computationalism.* For the purposes of the dissertation I will use "cognitive science" to refer to the cluster of disciplines that explain cognitive phenomena in terms of postulated internal representations and those that provide computational models of these representations. The advantages of relying on this characterization pertain on the one hand to my specific area of interest in the dissertation, and on the other to more general issues about discussions

of cognitive science. First, on this characterization cognitive science of science workers can be considered as cognitive scientists, in so far as they appeal to internal representations to explain cognitive phenomena. Further they can be consistently described as borrowing results from cognitive science when they appeal to work in other subdisciplines of cognitive science, such as cognitive psychology or machine learning. Second, the more general issue is that, on the above characterization, cognitive science is something over and above each of its contributing subdisciplines. Cognitive psychology, for example is concerned with the production of models of the internal representations of human subjects, whereas artificial intelligence is concerned with designing machines which produce, under some interpretation, an instance of a cognitive phenomenon.

3. Cognitive Science of Science.

Cognitive science of science researchers have one central concern: to use techniques from cognitive science to study science or scientists, both in the present and in history. For example they provide computer models that they claim explain incidents in the history of science, or they borrow techniques from cognitive psychology to draw conclusions about the psychological processes which are involved in a particular scientific discovery.

There is a sense in which cognitive science of science is a sub-field of cognitive science: Some cognitive scientists do cognitive science of science work. Simon provides an example of someone who is first and foremost a cognitive scientist, turning his attention to the particular cognitive activities of science. Many of his co-workers in his cognitive science of science projects are primarily cognitive scientists working variously on machine learning, human learning and the development of expert systems. It is more difficult to support the claim that cognitive science of science is a sub-field of cognitive science in any deeper sense. I give two reasons why here. First, workers in cognitive science are divided over the status of "higher cognition" such as scientific discovery, which some claim is not a proper site for cognitive science research. Second, cognitive

science is not as unified an endeavor as it would appear in some standard accounts (Gardner 1987) and, as evidenced by my account above, there is no clear consensus over what its sub-fields are.

Three of the four workers I focus on are philosophers of science and much of the agenda for coghitive science of science has been set by addressing concerns in the philosophy of science, so a useful way to introduce CSS is to sketch the relationship between cognitive science and philosophy of science.

Many philosophers have been concerned with the so called naturalization of philosophy.[2] Quine made the term "naturalized epistemology" one to rally around (Quine in Kornblith ed. 1985) and claimed a precursor in Hume. The general idea of a naturalized epistemology is to utilize the best resources of the science of the time to assist in reaching conclusions about knowledge. Quine's original suggestion was to turn to work in psychology; other suggestions have been to turn to work in sociology or history (Laudan 1977, Hacking ed. 1981, Fuller 1988). The specific claim made by philosophers of science involved in cognitive science of science is that the way to a naturalized philosophy of science is to utilize the resources provided by cognitive science (Giere 1988, Thagard 1988, Churchland 1990). At a first glance this seems to be a move in the Quinean spirit, but the four protagonists take the claim in different, and sometimes mutually incompatible, directions.

4. Cognitive Science of Science and Representation.

I focus on a restricted sample of work in cognitive science of science for several reasons. The first is simply that the field is very small. Well worked out and articulate positions are few and far between. The second is that the selected work is representative of what has been achieved so far in cognitive science of science. And it is representative in providing examples of most of the current approaches to cognitive modelling in cognitive science, especially in providing four different

---

[2] The project of naturalizing philosophy in general is a more complex one. For example, what it means to naturalize ethics has been a subject of debate for centuries. In this dissertation I concentrate only on naturalizing epistemology and philosophy of science.

accounts of the structure of representation. I defend this claim by providing an outline of Chapters 2-6 and of the positions of the four authors.

Simon was one of the first to propose that scientific discovery could be explained rationally by producing a computer program which could execute scientific discoveries (Simon 1966). In making this claim he addressed an issue that concerned philosophers of science: Whether discovery, as opposed to justification could be explained rationally (see Nickles 1980b). In addressing this issue Simon developed many of the crucial techniques and methods used in cognitive science of science, the use of computer models being the most obvious one, a topic I discuss at length. Protocol analysis is an important resource for cognitive science of science workers and was given an explicit formulation by Simon (Simon & Ericsson 1984). Protocols are transcribed versions of the verbal reports given by subjects in psychology experiments. Cognitive science of science workers claim that scientists' laboratory notebooks serve as protocols from which we can derive accounts of the psychological processes of scientists. I give an account of protocol analysis in Chapter 2 and assess its implications for cognitive science of science. Finally, Simon introduces normative proposals resulting from cognitive science of science and I use his work to illustrate some of the issues which arise from basing normative claims on such research. Simon's work provides a good opportunity to introduce in more detail some of the central claims of cognitive science of science and to introduce some recurring critical arguments.

Simon's account of the structure of representation relies on the representation of any example of cognitive behavior by means of a search through a problem space. The representational structures are the search trees produced by the combination of such a search, directed by various heuristics or rules-of-thumb. Simon provides no account of the way in which scientific theories are represented within individuals, but his account of representation includes enough building blocks to construct such an account. The remaining three researchers all present accounts of the nature of theories *qua* representational structures, and it is the differences among these accounts that distinguish their respective approaches.

Thagard relies on empirical results and theoretical techniques from cognitive science to inform his philosophy of science, but he also relies on his own cognitive science work. He uses computer

models, which he claims provide a descriptive account of scientific thinking and a normative account of science. His claims on behalf of his computer model are more far reaching than Simon's as he deals not only with scientific discovery, but with theory change and hypothesis evaluation in science. Examination of Thagard's work provides an opportunity to investigate the difference between normative and descriptive accounts of science and to begin to confront the issue of whether cognitive science of science provides accounts of an individual scientist, a particular instance in the history of science, or science as a whole.

Thagard's stock of representational structures is borrowed from artificial intelligence. He claims that scientific theories are schemata, consisting of a collection of rules and structured concepts. These representational structures are built up of proposition-like elements. Rules, for example, are sentences expressing conditionals: If x is a metal, then x conducts electricity. These representational structures can be used by a system in any order, there is no linear ordering to their application, as there might be for example in the heuristics-driven searches that Simon envisages.

Giere's work does not rely on computer models, but appeals to findings in cognitive psychology concerning "mental models." This is one of the approaches to modelling internal representations in cognitive science. Giere calls for a science of science which utilizes the results of findings in cognitive science. He also attempts to incorporate some work from social studies of science such as The Strong Programme in the Sociology of Knowledge, whilst attempting to draw a sharp distinction between their theoretical commitments and his own. Examining this aspect of Giere's work enables me to present a clearer account of the purported distinction between the social and cognitive components to science, a distinction which I challenge.

Giere represents theories as families of models. Models can be pictorial or symbolic and the exact nature of their embodiment is never made clear by Giere. His account is designed to rule out subsumption under the traditional logical empiricists' representation of theories as sets of sentences or axioms and deductive rules governing their application.

Finally, I turn to Churchland's work. This work relies on resources in cognitive science not appealed to by the other authors: Parallel distributive processing, which is a new form of computer

modelling in stark distinction to those used by both Simon and Thagard,[3] and the neurosciences. Churchland's turn to cognitive science has led him to make strikingly different proposals than Giere and Thagard for philosophy of science. For example he attempts to construct an epistemology that does not rely on any sentential representations at all, rendering any appeals to usual semantic criteria useless. Churchland proposes that the structure of the human brain sets limits on the types of cognitive skills that can be attributed to people at any particular juncture. The same limits would not be set if we appealed only to abstract computational considerations as Thagard and Simon do.

In the final Chapter I present a more detailed account of cognitive individualism and demonstrate some of the problems that result from adopting the position. Second, I develop some considerations that motivate the claim that scientific cognition can be a social phenomena, and need not be characterized in terms of the properties individual cognizers. Third, I argue that cognitive individualist work in the cognitive science of science leads to grave problems when attempting to present normative claims that are applicable to science and do justice to scientific practice. Finally, I present four theses implied by the arguments throughout the dissertation that serve as a basis for an alternative empirical study underlying a naturalized epistemology that avoids the pitfalls of cognitive individualism.

---

[3] Thagard claims that he is using a parallel distributed processing approach, but there are many reasons why he should not be considered as doing so. I go into these matters in Chapters 3 and 5.

CHAPTER 2

HERBERT SIMON'S COMPUTATIONAL MODELS OF SCIENTIFIC DISCOVERY

1. Introduction.

Herbert Simon's work on scientific discovery is important for several reasons. First Simon was an early advocate of rational scientific discovery, contra Popper and logical empiricist philosophers of science (Simon 1966). This proposal spurred on investigation of scientific discovery in philosophy of science, as philosophers used and developed Simon's notions of "problem solving" and "heuristics" in attempts to provide rational accounts of scientific discovery (see Nickles 1980a and Wimsatt 1980). Second, Simon promoted and developed many of the crucial techniques and methods used in cognitive science. One is the use of computers to model internal cognitive processes, a technique central to his account of scientific discovery. Another is protocol analysis, the use of the verbal reports of experimental subjects in psychology to construct accounts of their

cognitive processes. Protocol analysis is given a detailed formulation by Simon (Simon & Ericsson 1984), and is modified for use in the study of scientific cognition by Kulkarni and Simon (1988). Third, Simon introduces normative proposals for science based on his computational investigations of scientific discovery (see also Zytkow & Simon 1988). Simon's work can be viewed as a contribution to naturalized philosophy of science, which centrally features the derivation of normative proposals from descriptive accounts of science.

In this chapter I describe and critically evaluate Simon's recent work on scientific discovery.[4] I focus primarily on *Scientific Discovery* (Langley et al. 1987), which documents many computer programs that purportedly make scientific discoveries, and "The Process of Scientific Discovery" (Kulkarni & Simon 1988), which is a detailed investigation of Krebs discovery of the ornithine cycle. I present several distinct criticisms of Simon's work. First, I argue that Simon's account does not distinguish between an individual scientist's mind, a social process involving several scientists, or a historical process, and so he does not demonstrate which of these is the most important component of scientific discovery. Yet Simon argues that scientific discovery can be adequately accounted for by appealing to cognitive processes. I offer two lines of argument to establish that this latter step is unjustified. The first is that Simon's method of protocol analysis does not provide sufficient evidence for the existence of the distinct cognitive processes he claims are involved in scientific discoveries. The second is that a sufficient account of scientific discovery cannot ignore social and historical components of scientific discovery as Simon's account does. I conclude that as a result of these failures Simon's descriptive characterization of scientific discovery is inadequate, and further that this inadequacy is due to Simon's cognitive individualism. I conclude the chapter by considering the normative dimension of Simon's account, and arguing that his computer models of scientific discovery can be best understood as contributions to what Clark Glymour has called "android epistemology." This latter discussion provides the point of departure for Chapter 3, as Thagard claims that he is providing a normative theory of science in general, derived from his

---

[4] Throughout the text of this chapter and the dissertation as a whole I will refer to Simon's collaborative work by Simon's name alone. The citations credit his co-workers.

computer models of various phenomena of science. I will begin this chapter with a detailed description of Simon's approach.[5]

2. Background to Simon's Computational Approach.

Simon's work on scientific discovery is an extension of his work in information processing psychology, or more generally in cognitive science. His central claim is that scientific discovery is an extension of human problem solving behavior. He argues that a scientific discovery is a simple problem solving procedure adapted to a complex environment. To understand Simon's position on scientific discovery we need to unpack the above claims by briefly describing their theoretical background in his work in cognitive science in general.

In The Sciences of the Artificial Simon proposed the hypothesis that

> a man, viewed as a behaving system, is quite simple. The apparent complexity of his behavior over time is largely a reflection of the complexity of the environment in which he finds himself. (Simon 1969 p. 25)

Simon's notion of behavior is limited to cognition, he wants an account of "thinking man" not "whole man," a "psychology of thinking" (Simon 1969 p. 25). Simon holds that all thought is information processing. "Cognitive processes are a sequence of internal states or mental representations transformed by a series of information processes" (Ericsson & Oliver 1988 p. 396). On this view the human system executing these processes is simple and adaptive. The claim about simplicity is important for Simon, he argues that there are very few intrinsic characteristics limiting the adaptation of man's thought. These intrinsic limiting characteristics are the limited capacity of short-term memory size, long-term memory fixation time (the time taken to move chunks of information from short-term to long-term memory), and the serial nature of information processing. According to Simon human thought is adaptive as many new methods can be learned to solve the same problem, it is in this sense that most of human problem solving and thinking

---

[5] Many of the techniques Simon uses are taken up by other cognitive science of science researchers, and a description of them provides a background for discussions in later chapters.

behavior is "artificial," or learned and subject to improvement (see Simon 1969, Simon 1977 Ch. 5.1).

Simon's account of problem solving (see e.g. 1966, 1969, 1972) relies on the above considerations. Problem solving is a process of "selective search" through a problem space. The process takes place using resources retrieved from long-term memory and stored, for the duration of the problem solving task, in short-term memory. The strategies used to solve various problems are learned. For example Simon claims that cryptarithmetic problems (mathematical problems in which letters of the alphabet are substituted for numbers) are solved by humans using certain selective search strategies, and these problems can be solved more efficiently by a subject if they are taught better strategies. This is the sense in which problem solving, like all psychological processes, is adaptive. Simon claims that any problem solving strategy that humans use is constrained only by the limitations of short-term memory capacity, long-term memory fixation, and serial processing. These limiting parameters were established by experiments carried out on humans executing various problem solving or concept acquisition tasks (see eg. 1969 pp. 28-31 & Newell & Simon 1972).

Simon has accumulated a large body of information on problem solving activities, and uses his general theoretical criteria to guide his further investigations in human cognitive behavior. He claims that scientific discovery is part of human cognitive behavior (see e.g. Simon 1966 and Langley et al. 1987). He also claims that scientific discovery can be explained in the same way problem solving is explained. So for Simon *scientific discovery is a human psychological process that is related to problem solving* (and learning), *and can be explained within the general theoretical tenets of his information processing psychology*. Simon has claimed that his approach to psychology would provide an explanation of scientific discovery since 1966 (Simon 1966), and the work I concentrate on in this chapter is an extension of this earlier work. The chief instruments of research in Simon's information processing psychology are the computer models of the particular psychological process under investigation, and such models are prominent in his research on scientific discovery.

Simon claims that computers work in the same way that the human mind does, in the sense that they share the same important intrinsic restrictions on their cognitive activity (see eg. Simon 1969

pp. 31, 48, 54 and 1977 Ch. 5.1). These are the restrictions of short-term memory capacity and long-term memory fixation time, as well as the important restriction that they execute tasks serially. Simon claims that computers and humans use no parallel activity in executing a particular task (Simon 1969 p.53). He argues, for example, that a selective search will be executed step by step, one branch at a time in both humans and computers.[6] Simon further claims that what appears to be complexity in a computer program, is a result of the complexity of the environment the computer program is adapting to (Simon 1969 p. 22). So, for Simon, computers are "simple" systems in the same way that humans are "simple" systems (Simon 1969 p. 22). Given these similarities purported to exist between computers' and humans' psychological processes, it is a short step for Simon to propose that computer models should be used to investigate human cognition. Simon concludes that computer simulation of human thinking is one of the most important methods of investigating human thinking. He was one of the first cognitive scientists to claim that all psychological theories should be presented as computer programs (the relevant simulations) (cf. Baars 1986). Given Simon's claim that scientific discovery is a human psychological process, it is consistent with his information processing psychology to argue that an explanation of it will be generated by producing computer models.

The above provides a brief outline of the background to Simon's work on computer models of scientific discovery. Information processing is currently the dominant experimental and theoretical approach in cognitive science (see eg. Baars 1986, Sternberg & Smith eds. 1988). Rather than take on this whole tradition, I concentrate my criticisms on the application of the approach to scientific discovery. I now provide a brief description of some models Simon uses in this latter work.

---

[6] The serial nature of human psychological processes is challenged by champions of connectionism or PDP. Churchland's work, and to a certain extent Thagard's, gives a different account of cognitive activity based on the challenge to the serial criterion.

3. Computers that Make Scientific Discoveries.

In *Scientific Discovery* (Langley et al. 1987) Simon describes a set of computer systems that purportedly make scientific discoveries. Primarily he is concerned with systems that can make data driven discoveries, or discoveries of a particular relationship in a certain (usually numerical) data set. Simon claims that the systems are sufficiently general in their potential application to shed light on the notion of scientific discovery in general, rather than on the particular scientific discoveries investigated. The computer programs, the BACON programs, and other related programs, are progressively developed to deal with gradually more complex data, for example data requiring qualitative laws and attribute ascription. The book contains a large array of programs that Simon claims have discovered Kepler's Third Law, Boyle's Law, Snell's Law and many more.

Simon claims that the work

> seeks to investigate the psychology of the discovery process, and to provide an empirically tested theory of the information-processing mechanisms that are implicated in that process. (Langley et al. 1987, p. 4)

So the work is firmly embedded in his overall information processing approach in cognitive science outlined above. According to Simon, the research "is mainly limited to finding a set of mechanisms that is sufficient to account for discovery" (Langley et al. 1987, p. 4). So Simon's claims are that scientific discovery is a psychological process, and that a sufficient account of this process will be provided by the computer models. I expand and examine these claims below, after describing the BACON programs in more detail.

BACON.1 is the least sophisticated of Simon's programs. He claims that it discovered Kepler's Third Law of Planetary Motion (expressed as $D^3 / P^2 = c$),[7] as well as Boyle's Law and the Law of Uniform Acceleration. The program was able to detect the relevant regularities in the lists of numerical values it was provided with in each case. For example in the case of the Third Law of Planetary Motion BACON was provided with values for D and P.

---

[7] D = distance from the sun, P = period of orbit and c = constant.

BACON.1 is a production system, a computer program that uses a set of production rules (conditional statements) to operate on a set of data. An example of a rule for detecting regularities in data is: "If the values of two numerical terms increase together, then consider their ratio" (Langley et al. p. 66). Each of BACON.1's sixteen production rules purportedly represents a heuristic. It is argued that such heuristics are the same as those used by scientists when confronted with the same data sets that BACON.1 was confronted with. Summing up BACON.1's activity for Kepler's Third Law, it produces the relationship $D^3 / P^2 = c$ when provided with a list of values for each of D and P. It does so by operating on the data with the production rules. It is argued that these in some way represent the heuristics that Kepler himself used in coming up with the law.

Simon supports his claim that BACON.1 incorporates general heuristics of scientific discovery by demonstrating that the same program can "discover" Boyle's Law, the Law of Uniform Acceleration, and Ohm's Law from lists of data. The program uses the same simple heuristics to generate different laws from different data sets. Simon claims that his program remains simple and yet general, complying with his requirements for the explanation of information processing systems outlined above. Simon's aim is to develop a minimal pool of heuristics that will have a wide range of application to data driven discoveries. This pool of heuristics is then tested against a sufficiency criterion: "Can a program that contains only these selected heuristics to guide it actually discover significant scientific laws only with a modest amount of computing effort?" (Langley et al. p. 33). The further argument is that if this question is answered in the affirmative the "mystery" is driven out of scientific discovery (Langley et al. p. 33). Simon concludes that on achieving the above goal he

> will have shown how simple information processes, resembling those that have already been identified in other kinds of problem solving, can give an adequate account of the discovery process. (Langley et al. p. 33)

I challenge this conclusion below, but first I will outline the project of adding together programs that deal with different kinds of discovery processes.

Simon admits that scientific discovery does not merely consist of the comparison of lists of data, and the generation of laws involving two variables. He introduces more production system based programs that "discover" laws involving qualitative relations, such as those in chemistry, laws

involving intrinsic characteristics, and laws involving more than two variables (All documented in Langley et al. 1987). All the programs share the general characteristics of being "heuristic based searches" through specific "problem spaces" (see eg. Langley et al. pp. 281-2). They are all constructed in a similar way to computer systems previously developed by Simon for problem solving. The relevant problem space for these systems is a set of states, including an initial state and usually a goal state. Operators allow one to generate new states from the current state and these state changes are regulated by the heuristics, which are represented as production rules. The heuristics make the difference between these systems and a system of exhaustive search. An exhaustive search is one that attempts every possible combination of the data. An example would be an attempt to open a combination lock by methodically trying every single combination. Even in a seemingly simple data set of two columns of variables an exhaustive search would provide many irrelevant relationships between the variables. So a heuristic such as "If the values of two numerical terms increase together, then consider their ratio" greatly limits the search through two lists of numbers.

Simon's aim is to combine the various discovery systems in one general or integrated discovery system. This system would either embody all the smaller systems and their production rules linked together in some larger system (Langley et al. p. 281), or be a large heuristic search system consisting of a collection of the most successful heuristics from all the previously developed systems (Langley et al. p. 281). These two approaches are different, and Simon argues that the former approach, interconnecting whole systems, will be most fruitful in producing a general scientific discovery system (Langley et al. p. 290). He envisages that such a system will use the output from one system, for example his program GLAUBER, as input for another system, for example the data driven system BACON.1 (Langley et al. 1987 p. 290). Simon claims several such sub-systems could be linked together, either working serially or in parallel, and this would provide us with more insight into the discovery process in the following manner. The process of discovery will be explained for Simon, when each instance of discovery can be captured as a part of one general theory, embodied in a large combined computer program. For the moment it is important to note the structure of the individual systems, and Simon's perception of the overall goal of his program. The individual

systems are programs that carry out heuristics based searches through varying kinds of problem spaces, and the overall goal is to provide an account of scientific discovery in terms of these "simple" information processing systems.

The next system I describe is constructed with a view to adding a further dimension to the all purpose scientific discovery program envisaged; that of designing experiments. In presenting this latter computer system Simon also presents a detailed account of a further technique that can be applied in cognitive science of science: protocol analysis. So far we have considered techniques for representing discovery as heuristics based search. Protocol analysis is a technique designed to isolate the correct heuristics (the relevant psychological processes) used by a particular scientist in a particular case of discovery in order that they might be implemented in a computer system that simulates this discovery.

4. The Use of Protocol Analysis in the Study of Scientific Discovery.

In "The Processes of Scientific Discovery: The Strategy of Experimentation" (Kulkarni & Simon 1988) Simon discusses Krebs' experiments that led to the discovery of the ornithine cycle. Simon explains that the BACON programs did not broach the issue of where data came from in data driven discovery. He claims that the processes of designing experiments and observation were not investigated and that these latter are investigated in the work on Krebs. Simon's computational analysis of scientific discovery is governed by a guiding principle that scientific discovery is a collection of psychological processes and that these can be elicited by studying the work of scientists throughout history who have made important discoveries. Therefore the work on Krebs is intended to fit in with the multi-component discovery system approach outlined above. Simon looks to Kepler to derive an account of data driven theory discovery, and to Krebs to derive an account of experimental design and a "program of observation." I now describe this latter account.

Simon points out that "thinking aloud protocols have been used extensively as a tool for obtaining insights into psychological processes in problem solving" (Kulkarni & Simon 1988 p. 140)

and yet they are not available in the case of scientific discoveries. To understand this limitation we require a brief description of thinking aloud protocols.

Thinking aloud protocols are the text, cleaned up to some extent, from a person's verbal report on what she was doing whilst performing a particular task, usually a problem solving task, as a psychologist's subject. If a subject is presented with a problem, such as recognizing a pattern in a number of lists of three integers, her verbal report will be a process of "thinking out loud," and the protocol will be the transcription of this verbal report. There are several different types of verbal reports, for example those provided whilst the task is under way, and those provided immediately after the task is completed. I will attend to these differences in detail below in section 4. At the moment it is important to understand the aim of deriving such protocols. The protocols are used to derive an account of the psychological processes, mental representations, underlying the particular activity under examination (Simon & Ericsson 1984, Ch.1, and Ericsson & Oliver 1988). The account of underlying psychological processes derived in each case is constrained by a more general theory. In Simon's case his information processing psychology guides the research (Simon & Ericsson 1984, Ch. 1, and Ericsson & Oliver 1988). Without recourse to such a general theory Simon points out that protocol analysis would amount to little more than William James' notion of introspection. Recall that Simon's account of cognition is that psychological processing is information processing and memory is two layered, dividing into short-term and long-term memory. On this theory data from short term memory is the most readily available to the subject *during* the execution of a particular task, such as problem solving. So the protocols from problem solving tasks detail predominantly information from short-term memory, and the derived psychological processes are required by the general theory to be serial processes. I now turn to the use of protocol analysis in investigating the special case of scientific discovery.

There are no protocols of scientists' thinking processes, "since the research leading to [scientific] discoveries sometimes spans months or years [and] it is not practical to gather continuous protocols of the process" (Kulkarni & Simon 1988, p. 141). Also, since the discoveries investigated have already taken place, it is impossible to derive the relevant verbal reports during the execution of the discoveries. Simon resorts to a combination of accounts by recollection, accounts from published

papers, and accounts from diaries and laboratory notebooks. He argues that the recollections of the scientist under study, in this case Krebs, are unreliable and that published papers generally pay attention to explaining and justifying a discovery, rarely to describe how a scientist made it. According to Simon the clearest picture of the scientist's thought about a particular problem from day to day is provided by the investigator's laboratory notebooks, claiming that they provide an invaluable insight into the cognitive processes that lead to the production of published papers.

Simon argues that he can provide a reliable account of the *psychological processes* involved in Krebs' discovery of the ornithine cycle by relying on Holmes' (Holmes 1980) historical work on Krebs'. Holmes' account is constructed from a combination of Krebs' published work, his laboratory notebooks and interviews with Krebs made years after the discovery. Holmes' account provides the necessary protocols from which Simon derives heuristics, the proposed psychological processes underlying Krebs' discovery. These heuristics have been embodied in the production system KEKADA. As in the BACON system the heuristics are implemented as production rules. The production system is designed to follow the same discovery path that Holmes' account shows Krebs followed. The discovery is more complicated than ones reproduced by BACON, and is divided into three main stages: The discovery of the ornithine effect, the determination of the scope of the effect (is it only due to ornithine or also due to derivatives?), and the discovery of the reaction path, especially the important point that ornithine works as a catalyst.

KEKADA contains many components to assist it with the task. For example it is provided with a large store of background knowledge, which Simon claims corresponds to the background knowledge Krebs possessed. It is able to modify the background knowledge store, that is add to it or delete from it, as the process is carried out. KEKADA also contains an interesting component that rates results against a metric of surprise. If they are unexpected relative to prior knowledge, they are surprising and given a special status within the system. The end result of Simon's work on Krebs is the computer program KEKADA that he claims simulates Krebs' discovery of the ornithine cycle, and provides an account of the psychological processes involved in the discovery.

Overall Simon claims that Krebs' discovery of the ornithine cycle was due to a set of psychological processes that Krebs possessed, and that these are captured in the computer program

that simulates the discovery. These psychological processes have been isolated in the form of a heuristics based system that can be combined with the other systems developed by Simon in Scientific Discovery, to provide a more complete scientific discovery system in the manner outlined at the end of section 2 above. He proposes that the combined system will give a general account of scientific discovery. According to Simon KEKADA provides the component deriving data from experiments, which he argues is necessary in a general discovery system.

5. The Descriptive Adequacy of Simon's Account.

Philosophers of science have turned to studies of scientific discovery as a reaction to the claims of Popper and the logical empiricist philosophers that discovery was not amenable to rational explanation (see e.g. Nickles 1980b). Scientific discovery is also the subject of investigation of historians and sociologists of science (see e.g. Brannigan 1981, Pickering 1984, Galison 1987). The picture of discovery that arises from these investigations is by no means monolithic, rather one of a complex and varied activity. Discovery is part psychological activity, part sociology of group acceptance, and part historical accident and timeliness.[8] In contrast we see that Simon's account centers around the development of computer programs that arrive at the same results as great scientists in history. Simon argues that this approach provides a *sufficient* set of mechanisms to account for scientific discovery (Langley et al. p. 4). Further he argues that the success of these computer programs will "show how simple information processes ... can give an adequate account of the discovery process" (Langley et al. p. 33). Simon's descriptive account of scientific discovery shares none of the richness of the picture that arises from research in philosophy, sociology and history of science, and I argue that this makes it an insufficient account of scientific discovery.

---

[8] This list by no means exhausts the components of scientific discoveries, but certainly is a minimal list of necessary conditions for a scientific discovery. Scientific discoveries can also be crucially dependent on instrumentation and interpretations of data for example. As different scientific discoveries may have different crucial features there is no one list of necessary conditions that are sufficient to distinguish all scientific discoveries.

Let us begin with the use of protocols. Simon's own observations about the weaknesses of the method of protocol analysis can be extended into arguments against his use of this method to elicit the psychological processes involved in scientific discovery. I conclude that protocol analysis provides insufficient evidence to support the existence of the proposed psychological processes underlying in scientific discovery.

Two methods of direct verbalizations are used to generate verbal reports from psychology subjects: Thinking aloud and retrospective accounts (My account follows Simon & Ericsson 1984 and Ericsson & Oliver 1988). The former are recorded as the subject carries out the task under study, and the latter are recorded immediately after the activity under study has taken place to make sure that the subject still has the relevant information in short term memory. In the paper on Krebs Simon claims that scientists' laboratory notebooks are closer in nature to retrospective reports than thinking aloud reports (Kulkarni & Simon 1988). We have no thinking aloud reports in the Krebs case as the protocols are from a historical study not a psychology experiment. When retrospective reports are used in psychology experiments they are made according to specific guidelines for remembering what was thought about during task performance (Ericsson & Oliver 1988). So Simon's comparison between laboratory notebooks and retrospective reports is a weak one. Although laboratory notes are taken at the end of particular tasks (and even at the end of the day or the week), they are not taken at the end of a particular psychological process. The notion of coming to the end of a psychological process is not a relevant factor for a scientist determining when to make notes in a laboratory notebook.

Scientists do not primarily aim at recording their thought processes during experiments when making laboratory notebooks. For example Millikan's notebooks (see Holton 1978) contained columns of figures and comments such as "beautiful, publish this." This is good evidence that he was not always concerned with recording his psychological processes, rather with recording his results and commenting on their usefulness. Scientists in the laboratory record important results, or outline replicable procedures for themselves to use on a future occasion, or for graduate students

or technicians to use in their absence.[9] Laboratory notebooks do provide useful data about scientific practice, but it is not necessarily evidence for the existence of particular psychological processes of a particular scientist.[10] Such writings could also be used to generate an account of the social processes involved in a discovery (see Latour & Woolgar 1979). On the evidence of laboratory notebooks it is not only difficult to distinguish between different psychological processes, but it is also difficult to distinguish between the psychological processes of individual scientists and the more interactive processes of all the participants in the laboratory. The type of evidence provided by scientists' writings does not force one to the conclusion that particular psychological processes underlie scientific discovery.

In conclusion the analogy between verbal reports in psychology experiments, and laboratory notebooks as a resource for work on scientific discovery is strained for two reasons. First laboratory notebooks contain recollections, which may not have been made immediately after the putative psychological process they relate to occurred. Second, the laboratory notebooks do not contain information specifically about psychological processes. The laboratory notebooks can only be used as data from which an attempt to derive an account of psychological processes is made, and they could equally be used to derive an account of social processes.[11]

The use of scientists' more public writings to gain information about their psychological processes is even more problematic. One of the problems with retrospective reports in psychology is that subjects often "fill in" their reports with information not directly reproduced from memory (Simon & Ericsson 1984). They will perhaps give some *plausible reasons* for a particular activity instead of trying to remember the *actual processes* they went through (cf. Nisbett & Wilson 1977). Scientists' published work is almost exclusively concerned with giving a plausible account of the

---

[9] It is worth noting that graduate students and technicians often make entries in the laboratory notebooks. Certainly they cannot be recording the internal psychological processes of their supervisor. We see below that Henseleit, Krebs' assistant, took many of the notes the Holmes study was based on, yet Simon treats these as protocols.

[10] Tweney (1989) who derives an interesting account of what he calls "external memory" from a study of Faraday's notebooks (see also Gooding & James 1985.).

[11] There are several possible uses of laboratory notebooks once they are accepted as data from which various accounts can be derived, social accounts are not the only alternatives to Simon's type of account (see f.n. 7 above).

results (in the form of reasons for these results), the investigator is rarely if ever required to remember her/his psychological processes at the time of producing such results. Further, scientists do not attempt to distinguish between reasons for results, and the psychological processes that led to such results. No principled method is available to distinguish between the two types of post facto reports: Ones that present psychological processes, and ones that present plausible reasons for a particular act (cf. Nisbett & Wilson 1977).

Leaving the issues surrounding protocols as insufficient evidence for the existence of psychological processes, a second line of argument challenges Simon's claim that such processes provide a "sufficient" account of scientific discovery. Drawing evidence from work in the sociology of science, I conclude that his cognitive individualist account of scientific discovery is not sufficient as it cannot account for the social nature of scientific discovery.

Simon's approach to explaining scientific discovery is directed by his information processing psychology. Simon investigates scientific discovery as a process of "thinking man" (or thinking machine) (Simon 1969). His position is that the scientists under investigation use one or more of a common stock of psychological processes. These processes are heuristic driven search mechanisms. An important question is why one person's use of some shared psychological processes would produce a scientific discovery, whilst another person's use of it would not. Simon claims in earlier work (reviewed above in section 2) that the environment is the essential governing factor in producing different results with the same processes (Simon 1969, 1957), and yet he devotes no time in *Scientific Discovery* to describing how the environment is instrumental in producing the particular scientific discoveries he investigates. It is consistent with Simon's information processing psychology to argue that, given our shared psychological processes any human could come up with scientific discoveries, if he or she were put in the right environment. Putting it another way, it is consistent with Simon's account that factors other than the simple psychological processes he claims we possess are instrumental in producing scientific discoveries, and yet he leaves no room for such factors. Despite this deficiency he claims that he provides a sufficient account of scientific discovery. I now consider some factors that a sufficient account of scientific discovery must to account for.

Simon pays no attention to the issue of how one establishes that a scientific discovery has been made. A discovery's acceptance by the relevant scientific community is essential to its status as a discovery. And it is hard to separate this acceptance procedure from the process of discovery itself, a point argued by Brannigan in his *Social Basis of Scientific Discovery* (1981). Brannigan uses several examples from the history of science and exploring to illustrate his claim that the acceptance of a discovery by the relevant social group (the social context of the discovery), and the actual psychological process of discovery are indistinguishable (cf. Woolgar 1988 pp. 58-65). For example he assesses Columbus' "discovery" of America (Brannigan 1981 pp. 120-142) and Mendel's "neglected" discovery of the genetic basis of inheritance (Brannigan 1981 pp. 89-119). In both cases Brannigan argues that the special social contexts determined these discoveries. He argues that it was the preparations for Columbus' voyage and the recognition of his achievement by royal sponsors that distinguish his discovery of America. And for Mendel it was the emergence of a context within modern biology for his work that rendered it a significant discovery, and it was not until such a context arose that it became a significant discovery.

The distinction between acceptance and discovery could perhaps be cashed out in terms of a clear distinction between "cognitive" and "social" factors of the scientific discovery process. On this distinction the cognitive component of the discovery would be that part explained by Simon's psychological models, say chronologically the part of a discovery up to the submission of a paper reporting the findings. The social component of the discovery could be the particular peer review process that led to the acceptance of the paper by a distinguished journal. But this hypothetical picture is too limited and obscures the complexity of scientific discovery. Simon's own account of Krebs' discovery of the ornithine cycle gives us enough information to question any account based on such a straightforward distinction between "cognitive" and "social" factors.

Krebs' work on the ornithine cycle was carried out with an assistant. For much of the time Krebs' assistant Henseleit did all the experimental work and took all the laboratory notes, whilst Krebs was pursuing more theoretical work on this and other projects (Holmes 1980). Simon paraphrases Holmes' account of the discovery, which includes an account of Henseleit's

contributions, yet KEKADA models the putative psychological processes of an individual scientist. Whether Krebs could have carried out the work leading to his discovery by himself is irrelevant here, as Simon aims to explain the actual discovery of the ornithine cycle (Kulkarni & Simon 1988 pp.140- 143). But this discovery was produced by two cognitive agents whose interactions were instrumental in the discovery. KEKADA however is an idealized version of the possible psychological processes of an individual discoverer of the ornithine cycle. Thus KEKADA is not a model of the actual discovery of the ornithine cycle. Here we have a clear example of a scientific discovery, the relevant cognitive product, which was produced by more than one working scientist, or by social interaction.

If the interactive nature of scientific discovery were accepted, there would still be nothing in principle that prevents computer modelling of such activity.[12] For example Simon could claim that he was modelling the discovery of the ornithine cycle by producing a production system that characterized two heuristic based problem solvers, and embodied them in a system that combined and synthesized their results. But Simon is a cognitive individualist with regard to scientific discovery.[13] He holds that the cognitive process of scientific discovery can be accounted for by a model of an individual's psychological processes. The cognitive individualist approach prevents him from being able to provide a sufficient account of scientific discovery as it leaves important facets of scientific discovery unaccounted for, such as the interactions of a group of researchers essential to the eventual production of the scientific discovery, the relevant cognitive product.

For Simon scientific discoveries are produced by the psychological processes of an individual system, be it an individual scientist or computer program. Yet Simon's method of protocol analysis

---

[12] Rob Cummins' SOFT program is an example of a program that models the cognitive activity of groups of people (Cummins 1983). See also the discussion in Chapter 3 of modeling group activity.

[13] It is important to note the qualification "with regard to scientific discovery." In Simon's work on administrative behavior and his use of the notion of satisficing, the claim that the individual was the prime unit of analysis was not central. See for example his *Models of Man* (1957), which is interestingly subtitled "Mathematical Essays on Rational Human behavior in a Social Setting." Simon's work in information processing psychology has many affinities with his work in organizational behavior, which was neutral with regards its units of analysis. The work was applicable to individuals or groups, such as business organizations. I detect a tension between Simon's work on organizations and his work on scientific discovery, the former is neutral over its units of analysis and the latter is cognitive individualist. A possible resolution would be to view scientific discoveries as produced by organizations, and so the relevant heuristics would govern group behavior. Simon nowhere indicates that this is the way his work on scientific discovery should be understood.

does not provide sufficient evidence to establish the existence of these psychological processes. Even if an account of the relevant psychological processes could be provided it would not provide a sufficient account of scientific discovery, as it cannot account for scientific discoveries arising from social interaction. Simon's cognitive individualist account cannot encompass the richness of scientific discovery revealed by sociologists and historians of science.

6. Normative Accounts of Scientific Discovery and Android Epistemology.

Simon's computational models of discovery may not provide a sufficient account of scientific discovery, but perhaps they fulfill the role of refuting the claim, once held by the majority of philosophers of science, that scientific discovery is not amenable to rational analysis. Certainly Simon claims that his programs do this, but can this claim be sustained? Simon quotes Popper from *The Logic of Scientific Discovery*:

> ...The work of the scientist consists in putting forward and testing theories. The initial stage, the act of conceiving or inventing a theory, seems to me neither to call for logical analysis nor be susceptible to it. (Langley et al. 1987, p. 38)

In contrast Simon claims that there can be a normative theory of discovery. But if Simon has not provided a sufficient descriptive account of scientific discovery, what becomes of his normative account? I will suggest that Simon's normative account is best understood as a contribution to what Clark Glymour has called "android epistemology" (Glymour 1987).

Logical empiricist philosophers and Popper concentrated on the development of theories of confirmation or justification of scientific theories. They assumed a split between discovery and justification; the former was not amenable to logical analysis, whilst the latter was. Since Hanson's work in the late fifties and sixties (see e.g. Hanson 1958), the most sustained philosophical discussion of scientific discovery in the literature is collected in Nickles' two volumes on scientific discovery (Nickles 1980b). The philosophers represented in this volume are still, like Popper and the logical empiricists, concerned with the normative dimension of science, but the "friends of scientific discovery" (Nickles 1980a) hold that normative accounts of scientific discovery are

possible. For example some ways of going about scientific discovery are better than others. Most "friends of scientific discovery" have been concerned with elevating the status of scientific discovery from a mysterious process to a process amenable to rational analysis. The latter task involves providing a normative account of discovery. These philosophers also aim to provide accurate descriptive accounts of various scientific discoveries (see Nickles 1980b, Vol. II). They borrow historical or sociological methods to achieve this aim. An assumption driving this work is that if an accurate descriptive picture can be given of a great scientific discovery, it will inform the derivation of a normative account of scientific discovery in general (Nickles 1980a).

Simon has a similar project to the historically oriented philosophers of science (friends of scientific discovery), as he also aims to move from a descriptive to a normative account; the descriptive account constrains the norms derived. Simon focuses on both "regulative" and "evaluative" norms. Philosophers of science have traditionally been concerned with evaluative norms, for example norms for assessing a good theory. Regulative norms are those which, if followed, should produce effective procedures, including scientific discoveries. This distinction between regulative and evaluative norms is parallel to the one proposed by Nickles, who calls them "generative" and "consequentialist" norms (Nickles 1987).

Simon introduces his normative theory of discovery with the following claim: "The efficacy ("rationality," "logicality") [sic] of the discovery process is as susceptible to evaluation and criticism as is the process of verification" (Langley et al.1987 p. 39). Simon is explicitly addressing Popper and the logical empiricists and their separation of the context of confirmation, or verification, from that of discovery. He claims that "a normative theory of discovery would be a set of criteria for judging the efficacy and the efficiency of processes used to discover scientific theories" (Langley et al. p. 45). Simon claims that this theory "rests on contingent propositions such as 'If process X is to be efficacious for attaining goal Y, then it should have properties A,B, and C'" (ibid.) (the evaluative normative concern), and that "given such norms, we would be justified in saying that a person who adhered to them would be a better scientist" (ibid.) (the regulative normative concern).

Simon's work in *Scientific Discovery* is based on the assumption that there is no one scientific method, rather that there are several methods applicable over many domains of science. He calls

these "weak methods" to contrast them with more powerful specific methods within a particular domain of research. According to Simon weak methods are to be judged against the limiting case of random search. He claims that scientists are very seldom involved in random search, and rational activity is distinguished from random search by the fact that the *best* use of weak methods is employed. So in the formula for the normative theory of discovery quoted above, variables A,B, and C correspond to weak methods. Simon goes on to substitute "heuristics" for the notion of "weak methods." Hence the normative theory of discovery is restated as: "Rationality for a scientist consists in using the best means he has available - the best heuristics - for narrowing the search down to manageable proportions" (Langley et al. 1987 p. 47).

Simon's normative theory looks less like a normative theory of scientific discovery in general, than a theory of rationality construed as efficient search through a problem space. Of course Simon's descriptive theory treats discovery as a form of problem solving, and on his account problem solving is just an heuristics based search through a problem space. So, from the point of view of his information processing perspective, rationality and rational scientific discovery may amount to nothing more than efficient search. But this is not a general normative theory of scientific discovery. It is still an open question whether Simon's descriptive account captures scientists' psychological processes, and hence whether a scientist who adopted Simon's regulative norms, or more specifically used the best heuristics, would make better discoveries. It may well be true that if scientists were information processors whose work was best characterized by search through a problem space, they would become better discoverers if they used the best heuristics available. But as we have already seen Simon's descriptive account of discovery is far too limited, so a normative theory derived from this account can have only a limited application.

Simon's work is minimally consistent with that of naturalistic philosophers of science (for example historically oriented philosophers of science) who claim that a normative account of scientific discovery can only be developed on the basis of, and at the same time as a descriptive account (cf. Laudan 1977). The problem for Simon is that his descriptive account does not do justice to the complexity of the scientific discovery process, for example the social interactions involved in the process. Consequently the normative account he derives can provide no directives

for groups of scientists. Further it provides no directives for different instances of complexity in scientific practice, such as when it would be best to move to a different level of explanation to solve a scientific problem. In biology, for example, the solution to a particular problem might require shifting from the cellular to the biochemical level. What Simon's account does provide is a set of norms for guiding the simulation of further scientific discoveries, provided simulation can be achieved by representing the activity in terms of an heuristics based search through a problem space. Simon even claims his approach cannot "replicate the historical details of various scientific discoveries" (Langley et al. p. 62). Instead it can provide models of how such discoveries "*might* occur." If this is the overall claim of *Scientific Discovery*, then it is an entirely normative one, but we have seen that Simon also aimed to provide a descriptive acount. I will conclude by suggesting that Simon's normative claims may be best understood as guidelines to assist in building good scientific discovery machines, and so his work is best understood as a contribution to what Glymour has called "android epistemology" (Glymour 1987).

Glymour has proposed the name "android epistemology" for the production of the norms that regulate machines such as Buchanan and Mitchell's META-DENDRAL (Buchanan and Mitchell 1978), which have attained some level of success in scientific discovery (Glymour 1987). This project involves the development of machines to solve problems that humans have been accustomed to solving, especially problems that have traditionally interested philosophers. Scientific discovery is of central interest to many philosophers, so for Glymour the development of norms for machines that make scientific discoveries is work in "android epistemology."

Glymour has proposed a logic of scientific discovery implemented as a computer program (Glymour et al. 1988 and Glymour and Kelly 1989). The explicit difference between Glymour's approach and Simon's is that Glymour is not concerned with modelling actual human psychological processes (Glymour 1987, 1988). Glymour claims that "conventional artificial intelligence programs are little theories. The more theories look like theories of reasoning, the more the description of the program looks like a piece of philosophy" (Glymour 1988 p. 200). Glymour's view of philosophy is close to logical empiricism. He claims that the philosopher's concern is to give a "reconstruction of a domain of knowledge or form of reasoning" (Glymour 1988 p. 201. cf.

Glymour 1981). Glymour shares with the logical empiricists the assumption that reasoning processes can be abstracted from their context. In the case of scientific practice it is argued that the scientists' reasoning can be appraised independently of other scientific practices. Glymour diverges from the logical empiricists in arguing that one *can* provide a logic of scientific discovery, which for him is a theory of the reasoning that produces scientific discoveries. He claims that AI programs provide him with the formal capability of presenting such a theory. Finally, Glymour's account is an entirely normative one. His normative theory of discovery is proposed as a theory of how to make the best scientific discoveries (regulative norms). Or even stronger: How to go about discovering the truth. Recently he has claimed that he can give an account of the reasoning involved in discovering "the truth and nothing but the truth" (Glymour and Kelly 1989). This approach sets very high external standards by which his norms are to be judged.[14] Simon and Glymour's normative accounts share the same goals. Simon aims to provide an account of how the best scientific discoveries will be made. The way the account is implemented is in the construction of computer programs. Such programs bear little relation to the actual practice of scientific discovery, and so they fulfill one of Glymour's requirements for contributions to android epistemology, as they do not replicate human endeavor (Glymour 1987). The requirement is that the android epistemologist avoid what Glymour calls the "anthropocentric constraint": "[T]hat the algorithms executed by an android in performing a task must, at some appropriate level of description, be the *very same* algorithms that people execute in performing that task" (Glymour 1987 p. 74). While Glymour explicitly avoids the "anthropocentric constraint," Simon avoids it by default, due to the insufficiency of his descriptive account of scientific discovery.

Simon and Glymour's accounts of scientific discovery are in direct competition if they are both understood as android epistemology. The decision between the two accounts may be determined by generality of application, say by the scope of the regulative norms in each account. If we consider the applicability of regulative norms in terms of the maxim "ought implies can," then neither account is generally applicable to human scientists. If one is concerned with the production of good

---

[14] I return to Glymour's work later in the dissertation, and specifically to the issue of discovery of the truth. For example this is relevant in assessing Thagard's approach to scientific discovery.

science *per se*, then it is an empirical question which account is more generally applicable. The answer will depend on the quality of the new scientific work that the relevant norm-guided computers produce in the future.

Android epistemology is only one facet of the enterprise of cognitive science of science, others concern the description and explanation of human cognitive practices.[15] Glymour explicitly rejects this latter project. Simon, on the other hand, arrives at android epistemology by default, due to the inadequacy of his descriptive account. He provides guidelines for the design of efficient computer programs. Simon's account of scientific discovery does little to increase our understanding of scientific discoveries made by humans throughout history, and provides no useful regulative norms for *groups* of scientists practicing research currently. These are two of the goals that a naturalized philosophy of science might achieve, and Simon's work fails to reach them.

In the next chapter I turn to Thagard's computational philosophy of science, which tackles other facets of science than scientific discovery. The issue of developing a normative theory of scientific discovery is one that Thagard attends to, and acts as a point of comparison between his work and Simon's.

---

[15] Of course one cannot hold this sharp division between the goals of android epistemology and other more descriptive goals of cognitive science of science if one presupposes that humans and computers cognitive capacities are both computationally bounded. Much empirical work in cognitive science shows that although there are many deficiencies of human cognitive practices, the important bounds to human cognition are not purely computational (see eg. Faust 1984. and cf. Cherniak 1986).

CHAPTER 3

PAUL THAGARD'S COMPUTATIONAL PHILOSOPHY OF SCIENCE

1. Introduction.

Paul Thagard applies techniques from Artificial Intelligence (henceforth AI) to issues in the philosophy of science. He claims that AI provides a richer set of resources than logic for investigating science, rendering his approach more fruitful than logical empiricist analyses of science. Thagard wishes to provide an account of science that is psychologically realistic, pays attention to the history of science, and yet retains a normative dimension. If successful Thagard's approach would not only challenge traditional philosophy of science, but replace it with a new research agenda in which philosophical theories of science would be presented in the form of computer programs and tested on data from case studies in the history of science. Thagard, like Simon, offers an account of scientific discovery, but he also accounts for the nature of scientific theories, scientific

explanation, theory evaluation, the relationship between theory and experiment, group rationality, and more general epistemological issues such as justification and the status of realism.

I argue that Thagard's computer models atttempt to explain too much, and as a result he does not provide an adequate explanation of any of the issues he tackles. He runs together several explanatory goals and attempts to reach them all using one, somewhat limited, approach. First I argue that Thagard does not clearly establish the units of analysis of his study. The issue here is whether he is explaining individual psychology or successful scientific reasoning (or inference). Second I argue that in cases when he is clear about his units of analysis he applies the same computer models in each case, producing the unlikely conclusion that individual psychology and group decision making are explained by the same model. Third, I argue that as he applies the same computer models to psychological processes, social processes, and abstract reasoning, he attempts to explain too much, and as a result his account of science is inadequate. Finally, I return to Thagard's attempts to give his account a normative dimension arguing that the normative account he provides presupposes, rather than derives from, his "empirical" account. This argument turns on the awkward status of computer models interpreted as a source of empirical results. I conclude that Thagard, like Simon, contributes to what Glymour has called android epistemology, an enterprise that is highly normative and yet has little to do with the way humans actually practice science. This position contrasts with Thagard's avowed aim of providing norms for scientific inference that can be acted upon by practicing scientists.

2. A Brief Outline of Thagard's Project.

The most recent stage of development in Thagard's project of computational philosophy of science is a computer program ECHO that comparatively evaluates competing theories against a common stock of evidence (Thagard 1989b). ECHO embodies a theory of explanatory coherence, in which the most coherent theory is chosen from two competing theories. ECHO has recently been used to demonstrate the superiority of Darwin's theory of evolution over creationism and

Lavoisier's oxygen theory over the phlogiston theory (Thagard 1989b). ECHO "demonstrates" the superiority of Lavoisier's theory by assessing the ability of the theory to account for the relevant evidence from chemistry experiments. The phlogiston theory by contrast is shown to be capable of accounting for only a small part of this evidence. According to Thagard ECHO is able to make these comparative judgments because it is a connectionist machine, a notion that requires some elaboration.

Each hypothesis of Lavoisier's theory is represented in ECHO as a node in a network of nodes. A further set of nodes represent the evidence to be explained. The connections between the nodes are regulated by Thagard's theory of explanatory coherence. So if two nodes, say a piece of evidence and a hypothesis, cohere,[16] the connection between them will be strengthened. If a piece of evidence and a hypothesis contradict one another the link between them will be weakened. The starting position for a run of the ECHO program is to have one set of nodes representing the hypotheses from one theory (for example Lavoisier's) and another set of nodes representing the hypotheses from the other theory (the phlogiston theory) a further set of nodes represents the evidence to be explained by both theories. The nodes are proposition-like representations and the network is produced at the software level, Thagard does not have a hard- wired connection machine. The machine goes through cycles in which the activation level of each node is effected by those of the other nodes until it settles; no more changing of activation levels occurs. In the final state of ECHO in the Lavoisier case, Lavoisier wins out, because the hypotheses from his theory have formed stronger connecting links with the evidence and with one another than the phlogiston hypotheses. Thagard calls the final state "harmany;"[17] this is the state of system coherence. The whole system has reached coherence as a result of the pair-wise relationships of the various nodes. The best theory, for ECHO, is the theory that produces this state of "harmany."

---

[16] According to Thagard "propositions P and Q cohere if there is some explanatory relation between them" (1989b p. 436). He claims that this explanatory relation is exemplified in four ways: "1. P is part of the explanation of Q. 2. Q is part of the explanation of P. 3. P and Q are together part of the explanation of some R. 4. P and Q are analogous in the explanation they respectively give of some R and S" (1989b ibid.). Thagard claims that "explains" is a primitive in his theory.

[17] The name is intended as a tribute to Gilbert Harman.

Thagard claims that ECHO represents Lavoisier's reasoning process in settling on the oxygen theory rather than the phlogiston theory. When ECHO is used to assess Darwin's theory, he claims that it represents Darwin's reasoning process. Thagard argues that *ECHO attempts to reproduce the psychological reality of theory evaluation.* ECHO is to be treated as a model of the psychological processes of individual scientists. But this is not all, Thagard claims that the theory of explanatory coherence embodied in ECHO is a theory of how people should reason to the best theories or hypotheses.

Thagard's *Computational Philosophy of Science* (1988) provides the background for the ECHO project. Here Thagard argues that computational philosophy of science will involve a *descriptive* theory of the way scientists actually think and also a *normative* theory of how scientists should think. The two theories are to be developed together, the one informing the other. Thagard first concentrates on developing a theory of abductive inference, the inference from data to a new hypothesis concerning that data. This theory is embodied in another computer program PI (short for process of induction). Once an account of the generation of new hypotheses is developed Thagard turns his attention to the evaluation of hypotheses. The idea of assessing the best hypothesis adds normative considerations to his descriptive theory of reasoning.

The computer program PI is intended as a model of an individual scientist's explanation of phenomena such as the propagation of sound. According to Thagard PI represents the knowledge of a scientist and the process by which that knowledge is used. The representational task is achieved by utilizing "rules" organized by "concepts," these are technical concepts from AI and cognitive psychology (1988 p. 16). Thagard argues that the use of these concepts provides us with a better account of the nature of theories than has previously been provided in philosophy. Thagard's own example of the wave theory of sound illustrates this point.

"Rules" and "concepts" are specific data structures, which are the most fundamental structures in a symbolic computing system and are operated on by algorithms. Rules consist of relationships between condition and action conditions, for example: "condition: If x is sound" and "action: Then x propagates." PI contains many such rules and can add rules to its stock or remove them. This feature is allowed by providing the rules with activation levels, the more useful the rule during a run

of the system the higher its activation level, and vice versa. Each of the concepts in PI includes its own place within a hierarchy of concepts so the concept for sound contains superordinates of "physical phenomenon" and "sensation," and subordinates such as "voice" and "music" (1988 p. 17). Concepts are also attached to lists of rules that describe the general properties of the concept. Some rules attached to the concept of sound are: "If x is heard, then x is a sound" and "If x is a sound, then x is transmitted by air." Theories, such as the wave theory of sound, consist of rules and concepts and a history of past successes in problem solving tasks. The wave theory of sound consists of the concepts "sound" and "wave," the theoretical concept "sound-wave," the rules "If x is a sound, then x is a wave" and "If x is a sound, then x is a sound-wave," and the information that it has been used to explain why sound propagates and why sound reflects. For Thagard this "complex of interlinked structures" is what constitutes a theory (1988 p. 40). The theory is then used to explain phenomena, a process that PI also models.

Explanation for Thagard is parasitic on the notion of theory; theoretical explanation is problem solving using explanatory schemas. A schema is a "large complex unit of knowledge expressing what is typical of a group of instances" (1988 p. 198), and involves some sort of abstraction or generalization. Schemas are larger and more encompassing representational structures than rules or concepts, they produce a set of expectations that can be triggered when a relevant example is presented to the system. Thagard uses schemas to make sense of idealization in scientific theories. A theory must explain a large range of laws and instances of those laws, and to this extent it must be in some way abstract, the notion of schema is intended to provide the relevant abstraction. On Thagard's account explanation is the use of such explanatory schemas for problem solving, so explanation is a form of problem solving. PI "explains" the general law that sound propagates by creating the explanation problem: "Start: x is sound, Explanadum: x propagates." Then using various mechanisms of problem solving it provides an explanation. Thagard claims that his idea is not to reduce explanation to simple problem solving, he is merely working with a limited example. He envisages that PI could be extended to deal with contextual issues involved in explanation such as those brought up by Bromberger's flagpole example (Bromberger 1966).

Thus far Thagard claims that his account is a descriptive account of the nature of individual scientists' psychological processes. I remarked above that in *Computational Philosophy of Science* Thagard provided the background for the claim that ECHO was both a descriptive model of scientific thinking and a model of how best to think. PI is also intended to play this double role, but Thagard first defends his attempt to derive a normative account from a descriptive one.

The combination of the normative and descriptive accounts is central to Thagard's project, he holds that it is possible to develop normative and descriptive accounts in parallel, by making sure that the one is kept informed by the other. The resulting computer programs embody all the elements of the normative account. ECHO evaluates theories according to a theory of explanatory coherence, and PI produces the best hypotheses to explain a given range of data. This interdependence of descriptive and normative accounts is based in a position Thagard calls "weak psychologism," presented as a model entitled "From the Descriptive to the Normative" bearing the acronym FDM (1988 p. 133).

Weak psychologism is to be contrasted with strong psychologism and anti- psychologism. Thagard presents anti-psychologism as the view, defended by Frege and Popper among others, that epistemology and logic have no relation to psychology (see Haack 1978). Logic and epistemology are concerned with normative matters - the kinds of inferences we should make and the nature of justification - and psychology with descriptive matters - the kinds of inferences people actually make. According to Thagard the anti-psychologists further argue that to examine individual belief systems as revealed by psychology would lead to a hopeless subjectivism. Strong psychologism, what the anti-psychologists are interpreted as rejecting, is the view that logic is descriptive as well as prescriptive of human reasoning. Weak psychologism is the position that logic is prescriptive of mental processes, "it uses psychology as a starting point since it presupposes an empirical account of the mental processes about which to be prescriptive" (1988 p. 7). Thagard claims that weak psychologism escapes the charge of subjectivism because it separates the issue of the description of inferential practices from the question of what inferential practices are normatively correct. This approach requires a detailed argument to defend weak psychologism, especially as Thagard claims

that his overall project is best viewed as "a computationally oriented attempt to describe some possible results of a weak psychologistic research program" (1988 p. 8).

According to Thagard his attempt to derive normative accounts from descriptive accounts, embodied in the model FDN (from the descriptive to the normative), has had several precursors: the historical orientation in philosophy of science (HPS), wide reflective equilibrium in ethics (WRE), and the attempt to derive norms for logical inference from psychology (FPL). Thagard's own model uses the best features of each of these approaches, but has most similarities with HPS, as his focus is on scientific reasoning.

Thagard characterizes work in HPS as the combination of empirical historical work and reflective philosophical work. The historical work produces detailed case studies of various incidents in the history of science (descriptions of scientific practice) and philosophical work produces an account of the norms that guided the particular scientific research in question and derives a more generally applicable set of norms. A crucial assumption is involved here, this is that the norms (methods) applied by the scientists in question were the best ones. This assumption can be defended according to Thagard by considerations (given the benefit of hindsight) about the success of the theories produced by the particular scientific research under study (1988 p. 118).[18] On Thagard's account, the crucial difference between HPS and FPL, is that in the former case we have uncontested cases of exemplary behavior from which to derive our norms. He argues that successful work throughout the history of science can be clearly distinguished from unsuccessful work, and so it is clear what empirical cases should guide our derivation of norms.[19] In the case of logic, he claims, it is not clear where to look for cases of exemplary practice, and so there can be no straightforward derivation of norms from the norms used in an exemplary case of inferential practice. Thagard claims that several factors need to be taken into consideration when deriving norms of inference from psychology: first, the maxim "ought implies can" puts a restriction on what

---

[18] Thagard's model of HPS owes much to Laudan's discussion of the relationship between history and philosophy of science (Laudan 1977).

[19] This approach is decidedly whiggish, as there is no reason to suppose that the norms we project backward onto particular episodes of scientific theory change were instrumental in the success of the ascendant theory at the time the episode took place.

norms people can be expected to act in accordance with. Second he claims that an inferential system is a matrix of four elements that must cohere: normative principles, descriptions of inferential practices, inferential goals, and background psychological and philosophical theories. Coherence is assessed by three criteria:

> 1. Robustness: to what extent do the normative principles account for [inferential] practices? 2. Accomodation: to what extent do available background theories account for deviation of [inferential] practice from the normative principles? 3. Efficacy: given background theories, to what extent does following the normative principles promote the satisfaction of the inferential goals? (1988 p. 129)

The system that provides the desired norms of inference is the most coherent system according to these criteria.

The final stage of Thagard's account is the move to the general model FDN, with this model he hopes to cover all cases in which the normative is derived from the descriptive. He claims that the model is dynamic, that is the principles, background theories, and practices are considered against the optimizing criteria simultaneously. What he has in mind here is that given an initial set of practices, principles, and background theories the optimizing criteria can be used to generate a new set of principles that cover new developments in the practices to be accounted for. So this model enhances the HPS model by taking into account notions such as efficacy and accommodation, as well as the considerations about certain cases being considered as exemplary inferential practice.

Thagard connects this account with his computationalism by claiming that the way to test the theories of inference developed is to instantiate them as AI programs. The dynamic nature of his FDN model can then come into play, as the runs of the program will demonstrate the relationship between the principles, practices and background theories. So computational philosophy of science is a weak psychologistic program involving the testing of theories of scientific inference, by observing the behavior of computer programs that embody these theories. PI and ECHO are examples of such computationally embodied theories, PI performing abductive inference to rules (treated as hypotheses), and ECHO using a theory of explanatory coherence to chose between competing theories.

In the next section I turn to the first of my criticisms of Thagard's project. I examine the relationship between the description of individual psychological processes and the prescriptive task

of providing accounts of correct inference. I argue that Thagard's descriptive account fails due to his emphasis on providing an account of successful scientific inference.

3. Models of Psychological Processes or Models of Abstract Reasoning?

In this section I argue that Thagard fails to support his claims that he provides a model of individual scientists' psychological processes. I argue that he confuses the task of giving an account of psychological processes with the task of providing an adequate account of abstract scientific inference. I show that one account will not suffice to describe the psychological processes underlying scientific inference and to characterize the structure of correct inference. The collapsing of these two projects is a symptom of cognitive individualism.

Thagard aims to provide an account of the cognitive processes of scientists, this aim is linked to what he calls psychological realism.[20] He wants an account that is psychologically realistic, an account that explains data from psychology on scientific thinking. As his approach is to provide computer models of these processes, the best way to assess their psychological realism would be to compare their performance with data from human performance. Thagard's approach runs into problems in assessing which of these data are relevant.

If we grant for the moment that Thagard's aim is to simply provide a descriptive model of scientific thinking (I revise this assumption below when I consider Thagard's normative project), there are several ways he could go about this. These divide up along the following two lines. First, he could produce a model of thinking, or reasoning, in general and see how well this model applies to scientific thinking, and finally compare the model's performance to human performance. Second, he could produce a model of exemplary scientific thinking, and then compare its performance to certain exemplary human scientists. I will argue that Thagard fails to adequately distinguish these

---

[20] Thagard's sense of psychological realism is to be understood purely in terms of the match between the computer models and human psychological processes. He is not concerned with how it feels for each of us to carry out these particular processes, in fact he would not be unhappy with the claim that many of the proposed processes were "unconscious" (1989b p. 498).

two separable tasks. Each task is problematic in its own right, the first for reasons discussed in much of the cognitive science literature, and the second for reasons such as those provided in Chapter 2 of this dissertation. I will argue that Thagard's approach is predominantly the second, but is used to address issues that need to be confronted by the first.

Cognitive science in general attempts to provide accounts of human thinking and there are many well documented methodologies and approaches available for this task (see e.g. Sternberg & Smith eds. 1988). One approach is to provide a general set of rules and representations that would be sufficient to generate all human thought. The rules and representations can be implemented in a computer model which is then applied to various tasks and its performance assessed. The general pattern for such research has been to divide up the thinking tasks that are under investigation. So some researchers work on deductive logic, others on inductive logic, some on hypothesis formation and so on. The idea is that as humans can perform all of these tasks, any account of human thought must include accounts of each of these types of thought. Not all the results of cognitive science research retain this task specific modularity. Some more general types of representations are used to model several different thinking tasks. For example schemata, which Thagard uses, have been used in the modelling of various thought processes.[21]

Thagard takes the enterprise of cognitive science seriously and so presumably believes that a general theory of human thought will eventually be produced. What is unclear is whether he intends his work to be a contribution to such a theory. His model PI can make hypotheses about everyday objects, such as whether a certain creature is a bird, and apparently by the same mechanisms it can make scientific hypotheses (see Holland et al. 1987). The difference between representing everyday thought and scientific thought in PI is one of subject matter, as the same inference mechanisms that generate "If x flies, then is a bird," also generate "If x is sound, x propagates." If this is the case then Thagard's models of scientific thought will also be models of thought in general. The problem is that Thagard also wants models of the second kind, those that model exemplary scientific thinking.

---

[21] For example Chi, Feltovitch and Glaser (1981) use schemata to represent the thought processes of physicists solving various problems; and Schank and Abelson (1977) introduce schema-like representations to explain how we order food in a restaurant.

Thagard aims to model exemplary scientific thinking, such as Lavoisier's choice of the oxygen theory, by using Echo to model the psychological processes involved in this theory choice. In this case Thagard's model is not meant to be a model of everyday thinking, rather the model instantiates a theory of explanatory coherence, which, he argues, provides the principles underlying the choice of one theory over its rival. This theory of explanatory coherence is not merely a descriptive theory, it includes norms designed to guide in the choice of correct theories. The model of Lavoisier's theory choice is ECHO applied to evidence from chemistry, plus the hypotheses from both the oxygen and the phlogiston theory. The model of Darwin's choice of evolutionary theory is ECHO again, but this time the evidence is from biology and the hypotheses are from evolutionary theory and creationism. There are several questions that can be raised here. First, what is ECHO a model of? Is it a model of the thought processes of any successful scientist who has had to chose between two theories? Or is it a model of the abstract reasoning involved in the inference to the best theory on the basis of coherence criteria? Second, how much psychological realism is intended in the ECHO case, and of what kind? Is ECHO only realistic in specific cases of psychological processing, or is it realistic in general?

I conclude that ECHO is a model of abstract reasoning from the following argument. If ECHO is only applicable in special cases, then Thagard cannot claim that his programs are general psychological models. But we would expect such models to have some degree of universal applicability, unless each person has entirely different psychological processes. But if ECHO is an accurate representation of psychological processes in all humans, then it does not explain why only certain people make scientific discoveries. People who on this view have the same psychological mechanisms, for example Lavoisier and Priestly, ended up on two sides of a debate about theory choice. Understood as a model of scientist's actual psychological processes, Thagard's view prohibits this kind of occurence. At best Thagard's models represent abstract inference, and not actual psychological processes.

Thagard has placed too much emphasis on the traditionally philosophical tasks of providing a theory of the reasoning involved in theory choice to make his account a viable theory of psychological processing. Thagard's approach would look more plausible if he dropped all mention

of psychological realism. Then he would be left with an account of scientific reasoning that is different from the logical empiricist's due to the nature of its representational structures. Thagard believes that his models have more scope than this, but he may expect too much of them, in fact he argues that they can explain so much, that the claim that they can explain anything at all begins to look vacuous.

So far I have contrasted models of abstract reasoning and models of psychological processes. I now turn to the contrast between models of individual psychological processes, or individuals carrying out inferences, and models of social groups. Next I return to the notion of abstract reasoning and discuss the role of normative accounts, accounts of correct reasoning.

4. Psychological Models and Social Interaction.

In this section I argue that Thagard attempts to use a model of individual psychological processes to account for socially interactive processes. There are several stages to my argument. First, I show that one of the scientific case studies Thagard uses his model ECHO to "explain" is not appropriately accounted for in terms of psychological processes. Second, I turn to Thagard's attempt to model the inferences of a jury, which is a *prima facie* case of social inference. Third, I consider Thagard's own attempts to reconcile his approach with the social nature of scientific practice. I argue that he underestimates the extent to which scientific practice involves social interactions. Finally I conclude from the arguments in this section and the previous one that Thagard's models are used to explain too much, and as such do not adequately account for the phenomena.

Thagard's reconstruction of Lavoisier's choice of the oxygen theory does not do justice to the historical facts. Several of the respondents to Thagard's "Explanatory Coherence" (1989b) make this point and it is one worth developing. Thagard offers ECHO's choice, according to criteria of explanatory coherence, between the oxygen and the phlogiston theories in chemistry as a model of Lavoisier's psychological processes. Yet he draws his evidence for setting up the model from a paper

written late in the oxygen debate, the 1783 paper "Reflexions sur le Phlogistique." Thagard says that "the input given to ECHO represents Lavoisier's argument in his 1783 polemic against phlogiston" (1989b p. 444). And yet he claims that ECHO models the actual psychological processes of Lavoisier in chosing the oxygen theory over the phlogiston theory. By 1783 Lavoisier had established the oxygen theory in some detail, and both this theory and the phlogiston theory had gone through several modifications. The historical evidence indicates that Lavoisier himself was convinced of the falsity of the phlogiston theory in 1772, the so called "crucial year" (see Guerlac 1981, cf. Perrin 1989). The point of the 1783 "polemic" was to persuade other members of the scientific community of the superiority of the new oxygen theory. And so, as Giere correctly points out (Thagard 1989b replies), Thagard is modelling the arguments that Lavoisier presents, and yet it is not clear that he is modelling the psychological processes that led Lavoisier himself to choose the oxygen theory. Perhaps ECHO can be better characterized as a model of the dispute between Lavoisier and one of his opponents in the dispute, but this again undermines the claim that ECHO models Lavoisier's actual psychological processes.

Wetherick charges Thagard with having produced a sociological model, and not a psychological one (Thagard 1989b replies, p. 489), a similar point to the one I am making. The process of comparison between the arguments a scientist puts forward in defense of a theory, and the arguments their opponents put forward is more of a social process than a psychological one, but Thagard denies, without explanation, that his models are sociological. But Thagard also applies ECHO to a case, that looks to be a *prima facie* social process; jury decision making.

Thagard uses ECHO to model the decisions of juries in two court cases, he presents the case for the prosecution as one theory, and the case for the defense as the competing theory. The evidence is the same for both cases, although on occasion defense and prosecution will introduce contradictory evidence, and both bits of evidence are included. ECHO produces a decision as to which "theory" coheres better, and thus pronounces the defendant guilty or innocent. The question that is pertinent in this case is what is ECHO modelling? Is it modelling the psychological processes of any one particular jury member, or all of the jury members? Given Thagard's concern for the psychological realism of his account (see above pp. 8-9 and p. 12) perhaps his primary aim is that

ECHO model the psychological processes of individual jurors. I contend that it is modelling the inference of the whole jury, and that this is a *prima facie* case of a social process.

If Thagard's intention were merely to model the abstract reasoning that is involved in jury decisions, then the question of what he was modelling would not be as problematic. The instantiation question would not be appropriate. We would not need to ask what processes actually produced such reasoning. As I have noted Thagard is concerned with the descriptive issue, although he has been shown to be ambiguous in his presentation of it. In what sense can he say that ECHO models the jury's inference? Perhaps Thagard does intend to model the psychological processes of individual jury members, but if this is the case we would need eight runs of ECHO. On the other hand if all the jury members psychological processes were the same, then just one run of ECHO would suffice. There is a problem with this approach. If all the jury members psychological processes were the same, and this was a sufficient condition for all their inferences to be the same, a minimal conclusion would be that we would always get unanimous verdicts. The fact is that not only can hung verdicts be brought, but in some cases one or more jury members will be in disagreement with the rest, despite the fact that after prolonged discussion a unanimous verdict is eventually returned. Given that on Thagard's model anyone with the same psychological processes would reason to the same conclusions it becomes a serious question as to why we need more than one person in a jury at all. His model misses important phenomena of the jury decision process.

What in fact happens in jury decisions is a process of negotiation. Often one member will obstinately stick to her belief about the defendant's guilt, despite all the other members being convinced of the defendant's innocence. This kind of case can effect the outcome if the obstinate juror can win some support, pursuading the others, one by one perhaps, that the defendant is guilty. It has been shown in the social psychology literature that this kind of situation can lead to the majority leaning to the one obstinate individual's view (see e.g. Paulus ed. 1989, Ch. 9).

There are other cases exemplifying the social nature of jury decisions. For example some jury members may not understand the significance of a particular piece of evidence, and once the others have explained this to them their opinion may change thus producing a verdict. Jury decisions are not reducible to the psychological processing of their individual members, and certainly not in the

way Thagard has in mind. Thagard produces models that produce the same output as the modelled situations, but without any indication of the interrelationship of the contributing components. To produce a model that replicates a jury's verdict does not shed light on the process that led to that verdict. The parallel point is that *to model the choice of a "correct" scientific theory does not shed light on the processes leading to that choice.*

Thagard does claim that ECHO is potentially applicable to legal reasoning, and as I remarked above this is a reasonable claim, so long as he is concerned with *abstract legal reasoning.* Thagard is claiming too much if he claims that ECHO can be a descriptively accurate model of the social process of jury decision making, and by parallel reasoning that it can be a descriptive model of Lavoisier's inference to the choice of the oxygen theory. I now leave the argument from the case of legal reasoning on one side and return to the scientific case to consider Thagard's own comments on the role his models have in accounting for the social nature of science. I argue that Thagard is working with an impoverished notion of the social nature of science.

Thagard introduces two conflicting accounts of the relation between his computational approach and the social nature of science: in *Computational Philosophy of Science* he introduces the social nature of science as an example of the further potential application of his computational models, and in "Scientific Cognition: Hot or Cold?" (1989a), he presents an interpretation of the interest theorists' view of scientific belief change as socially determined in contrast to his own "cognitive," or computational, account of belief change. Ironically, Thagard claims in both cases that he can provide computational models of the social phenomena he refers to, arguing that this helps to demonstrate the superiority of his computational approach over sociological approaches to understanding science. But in explaining so called social phenomena with his psychological models, he presents the models as explaining too many phenomena and hence the "explanations" they provide become vacuous.

Thagard first introduces the social nature of science while defending his use of parallel computation in his models (1988 p. 186). He argues that parallel computation has an advantage because scientific communities can be viewed as parallel processing systems, and so they can be better modelled by parallel computational systems. These claims do little to defend the use of

parallel computation, as Thagard's computational model of the scientific community is simply one of a set of scientists all feeding information into a central review system. Such a simple model hardly forces us to adopt parallel computation to model it. But my main point is not to criticize Thagard's defense of parallelism, rather to bring out his account of the social nature of science.

On Thagard's view science involves groups of individuals with different characteristics, for example "audacious but reckless thinkers" and "careful but less original critics" (1988 p. 187) He introduces this view in the context of the issue of group rationality, that is addressing the question: what group characteristics would produce the best results in science? He argues that we can answer this question empirically (or we will be able to soon) by testing the performance of computer models of different kinds of groups of individuals, for example a group of conservative Kuhnian scientists compared to a group of more reckless Popperians (ibid.). Such computer models would be modelling group operations in contrast to PI, which models the problem-solving processes of individuals (1988 p. 188).

Thagard gives no details of how such group interactions would be represented. He presents scientific activity as a group of relatively autonomous reasoners feeding their results into some form of central processor, and so misses some of the more subtle nuances of the social nature of science. First, he fails to notice as I argued above, that if ECHO is a descriptive model, then it is being used simultaneously as a model of social and psychological processes. Thagard explicitly denies that ECHO is a sociological model (1989b replies, p. 491) thus confirming this point. Second, he gives no account of what the significant differences between a computational model of social processes and one of psychological processes would be. He does claim implicitly that PI is different than a social model, but gives no indication of what the difference is. Appealing to the capacity of parallel systems for modelling social processes indicates that he would prefer parallel models, but the very feature of parallelism itself does not distinguish characteristically social processes from psychological ones. Third, he introduces his notion of the social in what he considers to be a normative context, and yet despite his claim to abide by the maxim "ought implies can," he claims that empirical constraints do not come into play. Let us examine this third point further.

Thagard claims that the hypothetical division of labor between experimentalists and theoreticians "may reflect only individual differences in talents and inclinations" (1988 p. 187) and hence have no normative significance. But this claim conflicts with the maxim "ought implies can," which Thagard espouses, in the following way. Differences in inclinations and talents are the very kinds of considerations that force the adoption of certain norms to induce good scientific work from a group of variously talented individuals. If Thagard proposes divisions of labor that contribute to the production of good science, he needs to consider the empirical constraints on imposing those divisions. Among these empirical constraints are the very inclinations and talents of the scientists whose activities the norms are designed to guide. So here Thagard's notion of the social is hypothetical, and has normative implications, and yet makes no attempt to match up with empirical studies of the social nature of science. In contrast he does attempt to confront such empirical studies in his alternative presentation of the social nature of science, which I now consider.

In "Scientific Cognition: Hot or Cold?" (1989a) Thagard introduces the debate between philosophers and sociologists of science over the explanation of belief change. Philosophers, such as Laudan, claim that belief change should only be explained by appealing to models of rationality. Sociologists, such as Bloor, claim that belief change can be explained purely sociologically, and the type of explanation will be the same irrespective of the truth or falsity of the beliefs adopted. Thagard claims that neither approach is appropriate in the explanation of scientific change, from one theory to another, and that his computational approach is superior. The manner in which Thagard characterizes the social is important here, so I leave discussion of explanation of belief change according to criteria of rationality on one side.

Thagard claims that sociologists of science have attempted to explain the adoption of scientific beliefs in terms of the interests of those adopting the beliefs. His model ECHO "says nothing about the interests of [the scientists] who adopt particular theories" (1989a pp. 76-8) and thus ECHO is a model of "cold cognition." Cold cognition is "reasoning immune from motivational factors," and hot cognition is reasoning involving "motivations (interests) driving belief change" (1989a p. 72). Thagard argues that "we can rephrase the dispute between philosophers and sociologists in terms of this distinction: the former think that scientific cognition is primarily cold, while the latter think

that it is primarily hot ..." (ibid.) He then claims that "to have a chance of settling this dispute, we need well worked out theories of hot and cold cognition" (ibid.). His proposal is that we present computational models of each of the types of cognition and then test them on cases of scientific belief change. ECHO and PI are models of cold cognition and Motiv-ECHO and Motiv-PI are models of hot cognition, and for Thagard "the key empirical question for understanding science is whether the best account of the thought processes of scientists is given by cold models ... or by hot models..." (1989a p.). Thagard started out with an attempt to represent the aims of sociologists of science, by presenting their explananda, and ends up proposing that a computational model of the motivated thought processes of individuals will serve as an adequate presentation of the sociological account of science. There are several things wrong with this proposal.

First, Thagard has presented an account of the social nature of science, described above, which is entirely different to the one presently under consideration. The former account presents science as a process involving many individuals interacting with some sort of central reviewer to produce scientific results. The second account *reduces* the sociological account of science to an account of motivated reasoning in individual scientists. Second, Thagard's proposal simply misrepresents the sociologists of science's approach. He only attempts to present one type of approach in sociology of science, the interest theory, and uses this as a place holder for all sociological theories of science, which radically misrepresents the diversity of sociological views (see Woolgar 1988). Further the sociologists propose a theory such as the interest theory specifically to address issues in belief change that cannot be explained by appealing to the psychology of individual scientists (see e.g. Barnes & Bloor 1981). To attempt to reduce this type of explanation to one about motivated individual thought is to miss the point of the sociologists work. If Thagard had explicitly claimed that he could better deal with the explananda by using a individualist psychological model, then his approach would be more adequately motivated. Instead he simply claims that we can represent the sociological approach adequately by providing a model of motivated inference. Perhaps it would not be surprising if the computer model of sociology of science were rejected in favor of the models of cold cognition, given that the computer model provides an entirely inadequate model of scientific belief change.

Let us recap Thagard's attempts to deal with social processes. First, he claims that ECHO can model jury decision making, and yet he provides a model of either individual psychological processes, or of abstract reasoning, and not one of social processes, of which jury decisions are a *prima facie* case. When given the opportunity to admit that he is modelling a social process he explicitly denies this. Second he claims that science is a social process, and our best accounts of scientific rationality will come from models of groups of scientists, and yet he claims that because this is a normative process, empirical constraints derived from the observation of the actual social structure of science are irrelevant. This contradicts his claim that computational philosophy of science ought to present a normative account which is first and foremost guided by the principle of "ought implies can." Finally, he claims that sociology of science is best represented, for the purpose of testing its viability, as computer models of motivated reasoning in individuals. We have here three notions of the social. The first ignores the social, or its existence is denied. The second is more promising, as it represents an attempt to come to terms with the social nature of science, but the model provided ends up being more based on normative concerns than with a concern to give a descriptive representation of the social nature of science. The third is a further attempt to model the social nature of science, this time in different terms to the former approach, and in a manner inconsistent with that approach. One method allows for the possibility of significant social cognitive products, the other attempts to reduce social cognition to motivated individual cognition.[22]

In every case that a decision between two alternative viewpoints arises Thagard suggests that we build a computer model of each of the views and compare their performance. In some cases, for example in the treatment of Lavoisier, we have seen that this approach involves crucially misrepresenting the views to be compared. In many cases Thagard's computer models simply explain too much: for any given explananda he claims that his computer model provides an explanation. They simultaneously give accounts of real psychological processes, social processes,

---

[22] Once the idea that there can be social cognitive products is recognised, explaining how they arise becomes an important goal of naturalized study of scientific practice. Philosophers who are concerned with this explanatory goal include Fuller (1989), Hull (1989), Kitcher (1990) and Sarkar (1983). In the final chapter I return to this general issue as it relates to the nature of scientific theories.

and best cases of scientific reasoning. Without an adequate account of the relationship of all these explananda, and further an account of the exact explanatory role of the computer models, the claims that such models are explanatory becomes vacuous. Thagard's models may provide interesting versions of how particular cases of scientific change could have taken place, given certain empirically unrealistic constraints, but they do not explain scientific change. In the final section I turn to the issue of normative models made without empirically realistic constraints.

5. Thagard the Android Epistemologist.

In this section I return to Thagard's attempt to produce a theory of scientific inference and argue that despite Thagard's claim to follow the maxim "ought implies can" when introducing norms, he only considers a limited case of the application of "can": the capabilities of computer models. But if Thagard can only provide a theory of inference for computer models, he has not lived up to the challenge of providing an account of the norms that govern actual scientists' practice. Thagard's work is more appropriately understood as a contribution to android epistemology, than to naturalized philosophy of science.

I have argued above that Thagard actually explains little about the nature of science because he attempts to explain too much using the same model. Perhaps his computational models, like Simon's, are best understood as models of abstract scientific inference, and as such are predominantly normative.[23] The relevant explananda of such models are phenomena such as correct inference, convergence on the truth, and what contributes to any theory being the best scientific theory. These are the traditional explananda of philosophy of science. I use the term "cognitive individualism" to refer to the practice of giving an account of these phenomena in terms of the psychological processes of individual scientists. We have seen in this chapter that Thagard is a

---

[23] Thagard's notion of "normative" does not have as much force as traditional notions. Normative accounts are traditionally marked by what kinds of practices the norms would rule out, Thagard's normative "account" rests on the assumption that what has been acheived so far in science is the best result for science, and so his normative account does not produce any *projectable* norms for future scientific practice. See also f.n. 2 and 3 above and p. 20 above.

cognitive individualist in the relevant sense, but some of his projects can be assessed in the way one would assess traditional approaches in philosophy of science. Here I argue that Thagard pays little attention to the maxim "ought implies can" and is more concerned with providing answers to the traditional questions of philosophy of science. Such answers are provided without applying restrictions from the relevant descriptive theory of science, and are provided in the form of computer programs. This renders Thagard's project android epistemology, as it fits Glymour's requirements (Glymour 1987).

We saw in Chapter 2 that Glymour claims AI programs are like little theories, and the more they are like theories of reasoning the more AI is like philosophy. Thagard has produced theories of reasoning in the appropriate way, as AI programs. Before continuing with the main argument it is worth clearing up a confusion between Thagard and a commentator on his work on ECHO. Dietrich (Thagard 1989b, replies p. 474) claims that Thagard's work could be described as "computational positivism." Thagard replies that his work bears no relation to logical positivism as it is neither logical nor positivistic. Dietrich's point is more subtle than this reply allows for. One issue here is a historical one, Thagard is not so much following the logical positivist tradition as the later tradition of logical empiricism, including the work of Hempel, Carnap and Reichenbach. But I intend to continue neither the historical nor the terminological dispute here as both obscure Dietrich's main point.

Thagard shares some of the important features of the logical empiricist program, those that are shared by android epistemology. First, theories are characterized formally, not by logic, but still in an abstract manner. Second, Thagard attempts to account for the relevant phenomena of science in terms of the features of his abstract representations of theories and inferences about such theories. Third, given the ambiguity of the object of Thagard's models, he is in the same position as the logical empiricists in that the question of how the theories and reasoning processes are instantiated is still an open one.

Thagard falls between two equally problematic positions here. The one he wants to deny is the logical empiricist position of accounting for science merely in terms of abstract theories and reasoning processes. Yet the cognitive individualist position of accounting for phenomena such as

the truth of scientific theories in terms of the psychological processes of individuals is equally problematic. Thagard believes that his "weak psychologism" holds a middle ground between these two positions, but he tends, as Dietrich pointed out, more toward the former position as I will now argue. In my argument I substitute "android epistemology" for "computational positivism."

In "Explanatory Coherence" (1989b) Thagard claims to provide a theory of explanatory coherence. Such a theory addresses the philosophical concern of providing an abstract, generalizable account of how one theory turns out to be better than another. In *Computational Philosophy of Science* Thagard claims he can provide a defense of scientific realism, the view that "science in general leads to the truth" (1988 p. 139). Again this is a philosophical concern that requires him to provide an abstract and generalizable account. Due to their level of abstraction and generality both of these issues can be discussed without reference to either scientists' psychological processes, or computer models of scientists' psychological processes. Thagard considers his account of the psychological processes of scientists to be an essential component of his approaches to both the above issues. If the separation between these issues is clear in his work, we can conclude that his account is not a naturalistic one, as a naturalistic approach requires taking into account empirical evidence derived from a descriptive theory. To attend to the normative concerns of philosophy of science without recourse to empirical results is to reject naturalism. Further if his account produces computer programs that perform activities humans cannot perform, then it is android epistemology.

Concentrating on the notion of theory focuses the case against Thagard. Recall that he characterizes theories as representational structures, schemata consisting of rules and concepts. This characterization implies that theories are psychological entities, in the heads of individual scientists (see section 1 above). Yet when ECHO compares such theories, it is according to abstract characteristics such as their explanatory coherence. ECHO compares theories in their complete state, according to various formal criteria, for example coherence relations between hypotheses and evidence posits. But this process does not take place in scientists heads. Whole theories as structures only come to light once they have been accepted by the scientific community. So individual scientists do not compare whole theories, formal representational structures, to one another. Thagard only creates this appearance in the Lavoisier case by selecting historical evidence from very

late in the oxygen debate. Philosophers of science present theory choice in this manner, as they rationally reconstruct the relevant theories and then compare them to each other according to "rational" criteria. Thagard's approach is the same, he rationally reconstructs theories, not using formal logic, but using the representational structures of AI.

The issue of the nature of theories recurs throughout the rest of this dissertation. Both Giere and Churchland propose ways of accounting for the traditional philosophical notion of theory in cognitive science terms, which in both cases involves an attempt to place theories in the heads of scientists. Thagard makes the same cognitive individualist move producing problematic results move. The range of phenomena philosophers try to account for with their abstract notions of theory are more general than those that can be captured by treating theories as psychological representations. For example whether a theory is a true representation of the world is not a function of its role as a psychological representation.[24]

Finally, Thagard's approach culminates in the production, and comparative studies of, various computer models. ECHO considers all the evidence that supports a given theory simultaneously, and all the relevant hypotheses of that theory. It does so, because it acts with the benefit of Thagard's hindsight, he knows the outcome of the particular theory choice being modelled, and knows the evidence that it turns out was the most relevant. There are two problems here concerning psychological realism. First, we do not, and perhaps cannot consider all the relevant evidence for a particular theoretical claim, let alone a whole theory. Second, scientists working on the frontiers of their field have no access to historian's hindsight. So Thagard's claims that ECHO models the actual psychological processes of scientists are mistaken. What ECHO does provide is an attempt to address issues in the philosophy of science using a new, and more versatile range of formalisms, those borrowed from AI. But this makes Thagard's task a contribution to android epistemology, and given the android epistemologists' lack of concern for descriptive accuracy, this is a position far removed from naturalism.

---

[24] I develop this issue further in Chapter 4 below.

Thagard, like Simon, ends up contributing to android epistemology, which is only part of the task cognitive scientists of science present themselves with. A major goal is to explain how scientists actually do science. Cognitive science of science researchers seek an account of the psychological processes of actual practicing individual scientists. In the next chapter I turn to Giere's attempt to provide such an account.

CHAPTER 4

RONALD GIERE'S "COGNITIVE" THEORY OF SCIENCE

1. Introduction.

Inspired by recent empirical successes in the cognitive sciences Ronald Giere proposes to provide a cognitive theory of science, replacing philosophical explanations of science, and sociological accounts of science. According to Giere the theory would be more descriptively accurate than philosophical explanations of science, and superior to sociological accounts of science because of its focus on the capacities of individual scientists. If successful, such a theory would be a valuable contribution to the interdisciplinary study of science. Philosophers of science who have already been forced to re-assess their explanations of science, because they fail to be descriptively adequate, would welcome the theory Giere promises. Unfortunately, Giere's theory does not live up to his expectations. Some of the reasons for the theory's failure are idiosyncratic. Others point to more

general difficulties that the theory shares with other work in cognitive science of science we have already discussed. A successful cognitive theory of science would be one that shares the goals of Giere's theory, yet avoids both these sets of difficulties.

I argue in this chapter that Giere fails to provide an adequate cognitive theory of science. First, I argue that his notion of a "model" is not robust, or specific enough to ground his account of representation, or to support his defense of realism. Giere's inappropriate use of empirical evidence in support of his philosophical arguments is one source of these problems. Second, I argue that Giere's "satisficing" model of scientific judgment fails to meet the standards he sets for it. Although he argues that his "satisficing" model is a cognitive model of the judgments of individual scientists, I show that this argument is inconclusive. Third I show that Giere's account of science is relativist, in the sense that the interest theory in the sociology of science is relativist, despite Giere's argument to the contrary. Finally I relate the shortcomings in his theory to the adoption of a cognitive individualist approach.

2. An Outline of "Explaining Science."

Giere argues that his *Explaining Science* provides the basis for a cognitive theory of science. He argues that the study of science must be scientific itself to be successful, and its closest allies in current science are cognitive science and evolutionary biology. Giere draws on empirical results from both of these fields to support his philosophical arguments, but he draws on cognitive science predominantly. His work differs from the work of Simon and Thagard, examined in the previous two chapters, in not using computer models. Giere uses results from empirical work done on human subjects.

Giere wants to answer two questions in detail: What do scientists do? And, why do scientists do what they do? He claims he requires high standards of descriptive adequacy in addressing the former question, to enable him to provide adequate explanations in addressing the latter question. Giere's philosophical position can be clarified by comparing his strategy for answering the latter

question with logical empiricist and historically oriented philosophers of science. Giere emphasizes the accurate description of actual scientific activity as a basis for the explanation of science. The logical empiricists used rational reconstructions of scientific reasoning to achieve the same goal. Giere has a strategy similar to historically oriented philosophers of science, who use historical accounts of incidents in science as a basis for explanations of science. Giere's overall argument in the book is that his cognitive account of representation and judgment provides an adequate basis for his explanatory account of science.

Giere argues that his new naturalized philosophy of science, a cognitive theory of science, will take the place of both logical empiricism and the historical orientation in philosophy of science. He also argues that his cognitive theory of science will form a necessary part of any account of science, even a sociological or historical one.

The book can be roughly divided into three sections, plus an introduction that provides an overview of positions to date in philosophy, sociology and history of science. The first section offers a development and defense of an account of representation. The second section addresses scientific judgment, and the third section is an application of Giere's whole cognitive theory in a case study.

In the first section Giere argues that the representation issue in science centers around two questions: By what means do scientists represent the world? And, do these representational structures stand in some relation to a real world or not? The issues are familiar ones in recent philosophy of mind and philosophy of science (see e.g. Cummins 1989, Fodor 1987, and Leplin 1984). Giere's answers to the questions are: First, that representational structures are models of some form or other. Specifically that scientific theories, the most important form of representational structures in science, are families of models. Second, that models have a "similarity relation" with the real world. There are respects and degrees of agreement between the models and the world. Giere derives his account of models partly from cognitive science, and partly from his own study of scientific textbooks.

Giere contrasts his account of the structure of representations with the accounts of empiricist philosophers of science, and constructivist sociologists of science. He uses his account to defend realism against the arguments of these anti-realist theorists. The final part of the section on

representation (Chapter 5) involves an account of Giere's own experiences as an observer in a particle physics laboratory, and is designed as a spirited rhetorical defense of realism.[25] Giere's venture in to "laboratory life" is intended to support his arguments against specific types of anti-realism by demonstrating the necessity of realism for an adequate account of scientific activity. I present Giere's account of representation in more detail in section 3.

The second main section of *Explaining Science* is an account of scientific judgment. Giere holds that judgment is "a natural activity of human beings" (p. 6, Giere 1988. All future references to Giere's book will be by page number only.). This is a view he claims to share with cognitive scientists, and he dubs it a "cognitive" account of judgment. Further he argues that this natural human activity is centrally important in science. Giere's view, that judgment in science is a part of the natural human activity of judgment, contrasts with that of Bayesian theorists in philosophy of science. Bayesians hold that judgment in science is rational decision making guided by criteria of rational choice. These criteria are developed independently from evidence in cognitive psychology about human decision making.

Giere shares with Bayesians the focus on decision strategies rather than other forms of judgment, arguing that these are the most important facets of scientific judgment. He builds his case against these theorists on evidence from cognitive psychology (e.g. Nisbett and Ross 1980, Kahneman, Slovic and Tversky eds. 1982 and Faust 1984). He argues that the psychological evidence discredits the Bayesian account, and that a "satisficing model" of decision making most adequately accounts for scientific judgment. Giere's satisficing model is derived from one proposed by Herbert Simon in his work on administrative behavior (see Simon 1945), and accounts for the decision of an individual scientist who is presented with a choice between theoretical models through the "satisfaction" of various criteria. The satisfaction level is set by the model's fit to the world, its correctness (an epistemic value), and the scientist's interests, such as social, metaphysical and professional interests.

---

[25] Giere claimed in conversation that Chapter 5 is not intended as an argument to the best explanation for realism, though it does appear to be one to both this and other readers.

Giere goes on to apply his satisficing model of judgment to explain a number of cases in the recent history of physics. Finally, he attempts to incorporate the important insights of the interest theory, held by some prominent sociologists of science (e.g. Bloor 1976 and Shapin & Schaffer 1985), into his view of scientific judgment. He argues that his notion of satisficing is compatible with interest theory, and complementary to it. I present Giere's account of satisficing in more detail in section 4.

In the book's final section Giere applies his cognitive theory of science to the adoption of continental drift theory in geology. Specifically he uses it to explain why continental drift was not adopted until the nineteen sixties, when it appeared that sufficient data was available to support the view in the nineteen twenties.

In addition to his "cognitive" accounts of representation and judgment Giere presents parts of an evolutionary account of science. This involves two claims; first, that cognitive abilities are the mechanism for the evolution of science, and second that the social context of science and the pressure of the real world provide the environment in which these mechanisms work (see e.g. p. 222 and p. 277) I do not present any arguments against Giere's evolutionary account. He uses it at various points in his book to bolster his argument. He claims: "The power of evolutionary models is considerable, and I shall not hesitate to employ them whenever they seem fruitful." (p. 15). But later he claims:

> I will not be concerned in this work to develop the evolutionary analogy in detail. My primary concern is with the more specific cognitive processes of representation and judgment. (p. 18)

Giere does not put as much weight on the evolutionary aspect of his theory as he does on the cognitive aspect. The evolutionary strand of Giere's view is not argued for at length, and so I do not argue against it in any detail. I focus my attention on his "cognitive" theories of representational structure and of judgment.

3. The Structure of Scientific Representations and Scientific Representation.

In this section I argue that Giere's model based account of the structure of scientific representations is not a cognitive one when judged by his own standards. I show that Giere does not give an account of how scientists represent. Second, he is unable to defend the realism so crucial to his account. To argue this point I show that Giere does not keep distinct the problem of the structure of representations, from the problem of representation *per se*.[26] He concentrates on bringing evidence from cognitive science to bear on the broader philosophical question of representation, rather than concentrating on the specifics of a theory of the structure of scientific representations, which can be supported by the same evidence. He thus fails to achieve his goal of providing a cognitive theory of scientific representations, and finally cannot defend realism.

At the beginning of Chapter 3 of *Explaining Science* Giere claims that "one of the primary means by which scientists represent the world is through the use of theories" (p. 62). He goes on to claim that "any account of science must face questions like 'What are theories?'" (p. 62) He argues that philosophers' accounts of theories have been too general and that historians and sociologists have been content to provide accounts in terms of "beliefs" or "concepts," which are left unexplained. Giere argues that the kind of explanation required to adequately account for theories, as well as "beliefs" and "concepts," must be one "that employs resources beyond those of folk psychology" (p. 62). Such an explanation would be a "cognitive theory of representations," and Giere provides an account of the standards such a theory must meet.

Giere argues that

> **a theory of explanation is not to be judged by philosophical standards, but by the standards of the cognitive sciences. That is, an empirical theory of explaining would be judged by the sorts of evidence relevant to theories of other higher level cognitive activities such as language comprehension and problem solving. (p. 105)**

I focus on the way these standards apply specifically to a theory of the structure of scientific representations. Giere argues that the cognitive sciences recognize that people have various

---

[26] Throughout this chapter I use "structure of representations" or "representational structure" in place of "representations." Much of my argument hangs on the distinction between "representations" and "representation," and the above paraphrase retains the sense while being much easier on the reader.

cognitive capacities which they employ "in everyday interactions with the world" (p. 5). Further the central notion in the cognitive sciences is "that humans (and animals) create internal representations of their environment" (p. 6). The claim when applied to science is that "within the framework of the cognitive sciences, theories would be some sort of representation" (p. 5) Or in more detail, "scientific theories should be regarded as similar to the more ordinary sorts of representations studied by the cognitive sciences" (p. 6). Given these considerations we expect an account of the structure of scientific representations that is cognitive, and informed by evidence from cognitive science. I will argue that Giere does not present such an account.

Recall that Giere's objective is to answer the question "What are theories?" within the context of a cognitive theory of representational structure. Giere develops the concepts of "model," or "theoretical model," and "theoretical hypothesis" to answer this question. Theories consist of two elements: "(1) a population of models, and (2) various [theoretical] hypotheses linking those models with systems in the real world" (p. 85). I initially focus on the notion of models, and the role they play in theories. I will return to the notion of "theoretical hypotheses" later in this section.

Giere argues that

> theoretical models are the means by which scientists represent the world - both to themselves and for others. They are used to represent the diverse systems found in the real world: springs and pendulums, projectiles and planets, violin strings and drum heads. (p. 80)

He adds that models are "abstract entities" (p. 78), and non-linguistic entities (pp.79-80). Further he argues that "models," or "theoretical models...function as 'representations' in one of the more general senses now current in cognitive psychology" (p. 80) This is the crucial point that Giere does not establish.

The problem of representational structure in cognitive psychology, or more generally in cognitive science, is characterized as follows:

> Although we know that states of and processes in the nervous system play the role of representations in biological systems, it is an open question just which states and processes are involved in which activities, and how. (Cummins 1989, p.1)

Cognitive scientists provide competing responses to the problem of:

> ...discovering a way of characterizing representations that will allow us to understand both their physical instantiations and their systematic roles in mental processes.(Cummins 1989 p.1)

Giere does not address the problem of characterizing representational structures. His account of models does not comply with any of the constraints that would apply to a cognitive theory of representational structure. Giere develops an account of theories that can rise and fall independent of any considerations of its adequacy as a theory of representational structure in cognitive science. I now go on to examine this point in more detail.

For Giere, "representations scientists construct cannot be too radically different in nature from those employed by human beings in general" (p. 62). He points out that much work has been done on this subject in cognitive science, yet he does not connect his argument with this work. Instead he tells his reader "I shall not approach the question of what scientific theories might be by looking first at what cognitive scientists have to say about how humans represent their world" (p. 62). He turns to "scientific representations themselves" (p. 62), which he argues are found in scientific textbooks.

Giere describes the salient features of several textbooks in classical mechanics. He describes how these books are generally organized. He then describes their presentations of the linear oscillator, the simple pendulum, and the damped linear oscillator. He notes that the texts do not refer to these systems as real, rather they treat them as ideal systems that satisfy equations (p. 70). He uncovers such widespread use of idealization and approximation in these textbooks (and in his observations of the practices of working scientists), that he is led to claim that idealization and approximation are of the essence of science (p. 78). So he argues that an "adequate theory of science must reflect" such a central feature of scientific practice (p. 78).

"Models" are the feature of Giere's theory designed to account for the phenomena of approximation and idealization. He argues that the idealized systems discussed in science texts are "theoretical models." A model such as the "simple harmonic oscillator" is "a constructed entity," it has "no reality beyond that given to it by the community of physicists" (p. 78). It is in this way that models are "abstract entities," they have "all and only the properties ascribed to them in standard texts" (p. 78).

The next step in Giere's argument is to account for the way models are combined to form theories. This step is also derived from the study of science texts. I noted above that Giere presented

theories as a "population of models." This notion is expanded by his argument that any model in a theory must bear some family resemblance to other models already in the theory, and further that

> nothing in the structure of the models themselves could determine that the resemblance is sufficient for membership in the family. That question is solely a matter to be decided by the judgments of members of the scientific community at the time. This is not to say that there is an objective resemblance to be judged correctly or not. It is to say that the collective judgments of scientists determine whether the resemblance is sufficient. (p. 86)

Theories, on this account, are families of models that are combined together according to the decisions of the relevant scientific community. This is not an account of theories that sheds any light on their role as representational structures for an *individual* scientist. Giere's focus shifts from presenting an account of theories as scientists' representational structures, to an account of theories as a collection of abstract entities. To connect the two accounts Giere needs to show how the theories (or models) are instantiated in the human cognitive system. This is a connection that is required by Giere's own standards for a cognitive theory of representation. I now turn to Giere's suggestion concerning this connection.

Giere defends his study of textbooks by arguing that "the modern physics textbook evolved to its present form because this form is well adapted to the actual operations of human cognitive capacities" (p. 89). His evidence is from Larkin's (Larkin et al. 1980) study of expert physicists' problem solving activities. In the study, thinking aloud protocols were collected from physicists solving problems. The results indicate that physicists "first select from memory a representation of the problem - a model - and work from there. What is retrieved from the physicists' long-term memory, is not the axioms of mechanics, but an appropriate model" (p. 89). Giere argues that as the physicists in the study

> learned their mechanics from similar texts. It therefore would not be surprising to find that their knowledge of mechanics is organized along the same lines as the textbooks. (p. 89)

Giere takes the further step of claiming that the textbooks are adapted to human cognitive capacities. There is a large gap in Giere's reasoning here. Let me recap the steps. First Giere claimed he would provide a cognitive theory of scientists' representational structures, to be judged by the standards of cognitive science. He then proceeded to give an account of theories as collections of abstract entities referred to as models. This account was derived from the study of science textbooks. Finally, he argued that textbooks were structured in the way they are because they reflect

the structure of human cognition. But even if correct, which is doubtful, this latter claim does not bridge the gap between the results of the textbook studies, and a cognitive theory of representational structure.

Giere argues that his study of science textbooks will result in a cognitive theory of scientists' representational structures, because textbooks are "well adapted to the structure of human cognition" (p. 89). By parallel reasoning we can argue that cookbooks, knitting patterns, and car-engine manuals are adapted to human cognitive capacities. All of these are didactically successful, and all contain models (in Giere's sense) that we use to represent our world. We cannot assume that every didactically successful instructive text gains its success from its structural similarity with human cognition, because didactic success fails to pick out any particular representational structure. The structures (models) in textbooks, cookbooks, and so on, are represented in several manners. For example, they are represented as data structures, say as lists or flow charts, or graphically as diagrams, or propositionally in descriptive prose. All of these representations have considerable didactic success. Hence the presence of particular representational structures in didactically successful physics texts is not by itself evidence for the existence of the structures in human cognitive make-up. Further, Giere's account fails to pick out any of the many possible representational structures competing for prominence in cognitive science. Any one structure in a textbook can be represented by any of the possible representational structures proposed in cognitive science.[27] Thus Giere fails to characterize any one type of cognitive representational structure used by scientists. To this extent he fails to reach his own goal of making an "explicit linking of scientific models with the "schemata" of the cognitive sciences" (p. 20). I now turn to Giere's treatment of the problem of representation.

Giere does not focus exclusively on the problem of representational structure, he is also concerned with the problem of representation. This is the problem of explaining the relation between representational structures and world in terms of whatever empirical theory of

---

[27] Some examples of such representational structures are: Sentences in the head that correspond to sentences in the subject's language (Fodor 1975); computational data structures, such as lists of data operated on by production rules (see e.g. Langley et al. 1987); norms for comparing concepts with the subject's previously acquired concepts (Kahneman & Miller 1986).

representational structure one adopts (see Cummins 1989). Materialist philosophers of mind are concerned with the relation between representational structures in peoples' brains, and structures in the world. I call this relation the head/world relation. A connected, but distinct relation is the focus of work in the philosophy of science. This is the relation between theories and the world. In philosophy of science theories are discussed independently of their instantiation in particular scientists' brains. I call this relation the theory/world relation. The terms of debates about these two relations are quite distinct. Giere conflates both forms of the representation issue without paying adequate attention to distinctions in the terms of debate, as I will show.

I have argued that Giere does not give an adequate cognitive theory of representational structure, and without this a theory of representation is hard to get off the ground. Throughout my argument above I focussed on Giere's account of scientists' representational structures. The representation relation applicable to this account is the head/world relation, which has most often been discussed in the philosophy of mind. The theory/world relation is the focus of work in philosophy of science. Giere shifts in his discussion of representation from the former to the latter.

Giere captures the relation between a model and the system it represents by appealing to the notion of "theoretical hypotheses." Theoretical hypotheses are linguistic entities, "namely a statement asserting some sort of relationship between a model and a designated real system" (p. 80). According to Giere, theoretical hypotheses can be true or false, depending on whether the stated relation holds. The relation between the model and the real system it represents is not one of truth or falsity, as neither relata are linguistic entities (p. 81). Giere suggests a notion of "similarity" for the relation. So theoretical "hypotheses claim a *similarity* between models and real systems" (p. 81). The claims of similarity are made in terms of "respects" and "degrees." An example of a theoretical hypothesis is the following:

> The positions and velocities of the earth and moon in the earth- moon system are very close to those of a two-particle Newtonian model with an inverse square central force. (p. 81)

That is a Newtonian two part system resembles the earth-moon system with respect to "position" and "velocity," to a "very close" degree. Giere goes on to argue:

> That theoretical hypotheses can be true or false turns out to be of little consequence. To claim a hypothesis is true is to claim no more or less than that an indicated degree of similarity exists between

> a model and a real system. We can therefore forget about truth and focus on the details of the similarity. (p. 81)

He continues:

> ...the relationship that does the heavy representational work is not one of truth between a linguistic entity and a real object, but of similarity between two objects, one abstract and one real.(p. 82)

These arguments are made in terms of "abstract entities." So Giere has not established the relationships between representational structures in scientists' heads and the world, the head/world relation. Instead he has accounted for the relation between models, or scientific theories, and the world, the theory/world relation. Further he has not established the connection between these two relations in his own account, which is essential to making the account of representation dependent on that of representational structures.

Next Giere enters into the realism debate in philosophy of science. There are several realism debates in philosophy of science, and each debate hangs on different sets of evidence and arguments (see e.g. Leplin ed. 1984). Giere addresses realism with regard to scientific theories as opposed to the debate over scientific entities. He is a realist, and the realism he adopts is as follows:

> Scientific realism is the view that when a scientific theory is accepted, most elements of the theory are taken as representing (in some respects and to some degree) aspects of the world. (p. 7)

Similarity is the representation relation that does the work in his realist view. He argues that his view avoids problems that several other realist theories in philosophy of science entail. For example it is argued that realist theories cannot cope with the notion of approximation to the truth. Giere claims his theory copes well with this challenge, as all similarity judgments are made in terms of respects and degrees, which embody a notion of approximation. The problem for Giere is not so much his adoption of realism, but his attempts to defend it using evidence from cognitive science. I now examine one such defense of his realist theory.

Giere makes the connection between his realist account of the relation between theories and the world, and the cognitive theory of scientific representational structures in the following way:

> Accumulating evidence from the cognitive sciences, including even the neurosciences (P.S.Churchland 1986), suggests that human cognition and perception operate on the basis of some sort of similarity metric (p. 81).

But Giere is mistaken to appeal to cognitive science to support his argument for the adoption of similarity as the relevant representation relation between theories and the world. He can gain no

support from such evidence for his account of this relation. Whether models are similar to real systems does not hang on facts about human neurological makeup, as, for Giere, the similarity relation stands between abstract objects and physical objects. If it exists, the similarity relation between any abstract object and the real object is not a relation mediated by the human cognizer. The relation is a metaphysical or semantic one, and cognitive science tells us nothing about relations of this sort. Cognitive science might shed light on how we as cognizers recognize such relations, but it does not give us a means of investigating the relation itself. Giere tries to move from evidence in cognitive science, to a claim about structural relationships between theories and the world, and it does not work.

Giere cannot defend his account of the *representation relation*, by an appeal to cognitive science. Further, as similarity grounds his account of realism, his appeal to cognitive science does not contribute to his defense of realism. Recall that, for Giere, realism is the view that a scientific theory is accepted if most elements of the theory are taken as representing aspects of the world. As the theories turn out to be abstract objects, their relation to the world has nothing to do with human cognizers, and so evidence about cognizers does not help establish the case for realism. Evidence from cognitive science may simply be irrelevant to Giere's concerns at this stage. Even if it were the case that he had provided a cognitive account of *representational structures*, evidence from cognitive science about such structures would not be relevant to establishing the nature of the relation that existed between these structures and the world.[28]

Throughout his account Giere confuses the problem of representational structure with the problem of representation. These problems require two different kinds of evidence, and approach. I have argued that Giere's account fails to address either problem. In the next section I turn to Giere's cognitive theory of scientific judgment.

---

[28] Certainly if Giere had established a cognitive theory of representational structure it would put constraints on what kind of account of representation was open to him (cf. Cummins 1989). Giere has not established such a theory, so this approach is not available to him. Notice that the arguments I use to undermine Giere's theory are independent from the serious difficulties with similarity as the representation relation, which also present grave problems for Giere's theory (see e.g. Cummins 1989 pp. 27-34).

4. The Satisficing Model of Scientific Judgment.

In this section I criticize Giere's satisficing model of scientific decision making. I argue that Giere does not establish this model as either a cognitive account of scientific judgment, or an account of individual scientists' decision making. I then argue that Giere is mistaken over the role of relativism in the interest theory in sociology of science, and further that his own satisficing model collapses into a similar kind of relativism.

In Chapter 6 of *Explaining Science* Giere argues that "most students of the scientific enterprise speak naturally of scientists as 'choosing' or 'deciding' that some model is correct" (p. 141-2). Earlier he argued that "in the framework of the cognitive sciences ... the selection of a particular theory as the best available would be a matter of individual judgment"(p. 5). Giere aims to provide an account of these decisions, or judgments, that is "consonant with current thinking in the cognitive sciences" (p. 141). His topic is not scientific judgment in general, but scientific decision making. One reason he gives for narrowing his range of inquiry is that "there exists a highly developed set of concepts and principles for talking about decision making" (p. 141), and no corresponding set to talk about judgment in general. A second reason is that the decision to accept a theory or a model is very important in science, and has been the focus of attention for all "students of scientific life" (p. 142). Giere proposes a satisficing model to explain scientific decision making.

Giere's model is based on Simon's model of the decisions made by administrative agents (see Simon 1945 and Simon 1957 esp. Chs. 14 & 15). These agents have limitations imposed upon their abilities to "gather, store, and process information about their immediate decision-making context" (Giere p. 158), and yet they still make a decision. Simon proposed that the decision made was the one that was satisfactory, given a "satisfaction level" previously agreed on by the agent. Administrative agents may be unable to construct a coherent preference structure for their options, and they cannot generally calculate the expected utility of the options, but they are, according to Simon, able to distinguish satisfactory decisions from non-satisfactory ones. If they find no satisfactory options agents can lower their satisfaction levels until one of the options becomes

satisfactory. Giere's satisficing model is a special case of Simon's, designed to apply to decision making by individual scientists. Giere does not follow Simon's account closely, but he does note the parallels between their respective projects (p. 161). I will first consider the details of Giere's model.

Giere interprets the agent making the decision as an individual scientist. The scientist is faced with a choice between two models, S and D. There are two possible states of the world. In the first case model S is similar to the real system under study, and in the second case model D is similar to the real system. There are also two possible outcomes: In outcome A model S is most like the world, and in outcome B model D is most like the world (p. 161). The scientist is presented as valuing correct outcomes over mistakes (p. 163) (a point I will return to below). Giere adds that "applying a satisficing strategy requires the additional restriction that the outcomes representing correct decisions be regarded as *satisfactory*, while those representing incorrect decisions are regarded as *unsatisfactory*" (p. 163). Giere argues that this is not a trivial restriction as "one can imagine a scientist whose career and skills are so bound up with a particular model that to reject the model would make it impossible to go on doing science" (p. 163). Giere labels scientists in this predicament "closed-minded," and offers Priestley as an example of such a scientist (p. 163). The "(minimally) open-minded" scientist regards both correct outcomes as satisfactory (p. 163). The choice between the two satisfactory options depends on evidence from experiments. Giere postulates that if model S is correct, the range of experimental results is RS; and if model D is correct, then the range of experimental results is RD. The choice for the scientist is now: "If RS is the actual result of the experiment, then choose ... S. If RD is the actual result, choose ... D" (p. 166). If the experimental results fall between RS and RD, Giere argues that the experiment will be repeated or another one designed. Having set up his model, Giere requires a way of testing its applicability.

Giere admits he can provide no evidence from cognitive psychology, or any other field, that humans in general are satisficers (p. 159). Fortunately there is also no evidence that establishes that humans are not satisficers. It is simply the case that very little experimental work has been done on the subject. As he cannot establish his case that scientists are satisficers from the generalization

that humans are satisficers, due to lack of evidence, Giere attempts to establish "the case for scientists ... on its own terms" (p. 160). He rephrases the task as a question: "How could one best go about determining whether a satisficing model fits actual instances of scientific decision making?" (p. 179) He answers that he could devise and carry out experiments on scientists, but he does not have the resources (p. 179). Instead he uses

> ... the traditional method of the historian or philosopher of science, the analysis of scientists' writings. In addition, having had access over an extended period to scientists engaged in ongoing research, I can supplement the examination of both published and unpublished writings with more ethnographic sorts of data, particularly observations and interviews. (p. 179)

Recall that Giere says he will present a cognitive account of scientists' decision making. For example he says "my claim will be that scientific judgment is a natural cognitive process" (p. 94). Yet all he can establish using these research methods is "that sometimes real scientists are satisficers" (p. 179). This is far short of the goal of presenting a cognitive theory adequate to the task of explaining scientific judgment.

If all Giere can establish is that "sometimes real scientists are satisficers" (p. 179), he is in a very weak position to argue that his *cognitive* theory of scientific judgment has any general application. He certainly has a model that gives plausible interpretations of some scientific judgments, but this is not enough. One would expect a cognitive theory of judgment to isolate one or more processes that are involved in all, or at least most, cases of scientific judgment. Giere merely presents some examples of decisions in physics (the choice of relativistic Dirac models (pp.185-198)), and in geology (the choice of continental drift (Ch.8)) that can be interpreted as involving satisficing strategies. Giere states that the case of continental drift in geology was "not chosen at random" (p. 227), but chosen because it was especially well suited to his approach. This is not a strategy of articulating a model then testing it, he is using his model as an interpretive heuristic. Giere does not provide a generalizable cognitive theory of judgment, he merely provides an interpretive framework that can be plausibly applied to a few cases of scientific judgment.

Giere does provide two criteria for distinguishing a cognitive theory. First, it employs "the resources of the cognitive sciences" (p. 2). Second, it takes "the individual scientists as the basic units of analysis" (Giere 1989 p.8). He cannot rely on the first criterion to support his theory of judgment, because, as he even notes himself, there has been no work on satisficing in cognitive science (p.

159-60). Certainly there has been no work testing the validity of the claim that satisficing is a decision strategy in individual human subjects. Therefore his theory is not cognitive according to the first criterion. Does it satisfy the second criterion? No, as there are also problems establishing that the theory is required to take individual scientists as its units of analysis.

Giere merely postulates that the satisficing model be applied to individual scientists' decisions. He says "let us redeploy Simon's model, interpreting his agents as scientists" (p. 161).[29] He provides no argument or evidence for this postulate, nor does he indicate how the model is instantiated in individual scientists. Giere consistently interprets individual scientists' decisions in terms of the satisficing model, but no additional evidence is brought forward that the satisficing model forms part of scientists' cognitive makeup. The crucial link to the individual that would legitimate Giere's postulate is simply never provided.

Let us return to the question of whether the satisficing model is a cognitive theory of scientific judgment, or decision making (see e.g. pp. 6,9,21,141). Giere admits he cannot defend his satisficing model by drawing on evidence from cognitive science, and he has not indicated what kind of cognitive activity satisficing is. Second, Giere merely postulated at the outset that satisficing is an activity of individual scientists, a postulate that we are not forced to accept. Even if we accept this postulate, an account in terms of individuals is not necessarily cognitive.

Both representation and judgment are cognitive processes of individual scientists according to Giere. I argued in section 2 that he does not establish the case for representation as he construes it. Now we see that he does not establish the case for the satisficing model of judgment. Giere provides an interpretive framework that focuses on individual scientists, but he does not provide a cognitive theory of scientific judgment. Let us now examine his claim that the satisficing model is compatible with interest theory in the sociology of science, returning, as promised, to the role of values in the satisficing model of scientific judgment.

Giere points out that in using decision theoretic models the "necessity for dealing with values or interests is explicit from the start" (p. 161). He allows that several interests or values are at play

---

[29] Simon's model of decision making, on which Giere's model is based, applied equally to groups, such as businesses.

in any decision between theories. Giere's minimally open-minded scientist will always value *correct* outcomes as satisfactory. But professional, social or metaphysical interests may also come into play (p. 163). Giere's recognition that the values involved in scientific decision making include a wider range of interests forms the connection between his "cognitive" account of scientific judgment and the Edinburgh School's interest theory in the sociology of science. I focus on two points here. First, I will argue that Giere is mistaken to claim that relativism is not essential to the interest theorists view. Second, I will argue that Giere's own cognitive theory collapses into relativism, because of his social constructivism.

Rather surprisingly Giere argues that his satisficing model of scientific judgment is consistent with interest theories in the sociology of science.[30] In such theories interests causally determine scientists' belief adoption. Interest theorists are naturalists in this regard; they are aware of the distinction between reasons and causes for belief adoption, but give priority to a causal account of belief adoption. On their view the social and political milieu can be responsible for the generation of a particular scientist's beliefs. More generally, social and political interests can be the most important factors contributing to the choice of one scientific theory over another. They see interest theory as a challenge to philosophers' accounts that scientists' beliefs are explained by appealing to the truth, or to the rationality of the scientist.

Giere argues that the apparent conflict between the interest theory and his cognitive theory of science

> arises because of the doctrines that have been *associated* with interest theories but are not essential to an interest theory.... These doctrines are anti-realism and relativism. (p. 165)

So Giere, a self-proclaimed realist and non-relativist, argues that he can consistently adopt a version of their view without either of these doctrines. But relativism is essential to the interest theorists' view, as a detailed examination will show.

Barnes and Bloor[31] argue that

---

[30] A good programmatic overview of interest theory is Barnes and Bloor in Hollis and Lukes eds. 1982, for more detail see Bloor 1976, and for an interest based case study in history of science see Shapin and Schaffer (1985).

[31] Barnes and Bloor are founder members of the Edinburgh School in the sociology of science. They are the most prominent defenders of the interest theory.

> far from being a threat to the scientific understanding of forms of knowledge, relativism is *required* by it. Our claim is that relativism is *essential* to all those disciplines such as anthropology, sociology, the history of institutions and ideas, and even cognitive psychology, which account for the diversity of systems of knowledge, their distribution and the manner of their change. (In Hollis and Lukes eds. 1982 p. 22, my emphasis.)

Barnes and Bloor go on to say that the version of relativism that they hold is based on three postulates:

> (i) the observation that beliefs on a certain topic vary, (ii) the conviction that which of these beliefs is found in a given context depends on, or is relative to the circumstances of the users. (Hollis and Lukes p. 22)

And finally a third "equivalence postulate" requires that "all beliefs are on a par with one another with respect to the causes of their credibility" (Hollis and Lukes p. 23). This third postulate embodies the claim that, regardless of a belief's truth or falsity, its *credibility* is equally problematic. Here credibility is to be understood as the fact that a belief is adopted. For the interest theorists, one should investigate the contingent factors that causally contribute to the adoption of any belief, regardless of whether it is true or false.[32]

The interest theorists argue that true and false beliefs do not fall into different classes. The words "true" and "false" are evaluative, but only relative to a specific social context. They argue that there are no context independent ways of distinguishing between true and false beliefs. The claim that truth and falsity are context dependent is essential to their view. The interest theorists argue that, since there is no context independent distinction between true and false beliefs, the adoption of all beliefs can be given the same causal explanation. In the case of a "true" belief (relative to the context) the causal explanation will account for the evaluation "true" (in that context). The fact that the belief is evaluated as "true" by those who adopt it is not causally relevant in its adoption. I am not concerned here to appraise the interest theory in the sociology of knowledge. Its notions of "belief" and "cause" have been the subject of philosophical scrutiny already (see e.g. Brown ed. 1984, Roth 1987, Fuller 1988). I simply want to demonstrate the essential role that relativism plays in the theory.

For the interest theorists scientific claims have no special status. This is an important point of difference with Giere's position. For interest theorists there is no set of epistemic values that

---

[32] Barnes and Bloor argue directly that cognitive psychological accounts are incompatible with their account, so not *all* contingent factors are to be investigated. (see e.g. Barnes and Bloor 1982 p.32).

explains the adoption of scientific theories. Scientists merely provide one, albeit very coherent, account of the world, among a plethora of alternatives. The interest theorists are relativists because they do not acknowledge any context independent way of distinguishing between true and false beliefs (cf. Woolgar 1988). According to Giere, our most recent scientific theories are in some sense objectively better than earlier ones, they have a better fit to the world (p. 56). And further, for Giere scientists value the theories with better fit, more correct theories, over previous theories. Giere therefore wishes to be anti-relativist in the crucial sense. Turning now to Giere's arguments for this position, I will show that Giere has no way of providing a privileged status for the value he calls "correctness" in his account over any alternative values he invokes. I conclude that, despite Giere's argument to the contrary, his own account of scientific representation cannot be used to defend his position against relativism.

Let me first make it clear that Giere wants to avoid the kind of relativism that the interest theorists' embrace. There are several places where he makes this position clear. He argues that his cognitive theory "allows scientists to be real people with a full range of human interests while also being engaged in something like 'the pursuit of truth'" (p. xvii). Later he argues that sociologists of science, including interest theorists "utterly fail to explain the obvious success of science" (p. 4). He argues that there is "something important missing from the sociological account" and that is a "causal interaction between scientists and the world" (p. 4). Finally, Giere argues:

> Scientists, it is said, value truth over error. I prefer to speak of the similarity between a model and the world rather than the truth of statements, but the intent here is the same. (p. 162)

Giere does require more than the relativist sociologists offer. He claims that his account is grounded in the similarity relation between models and the world. Recall (from section 2) that models, for Giere, are cognitive representations of real systems in the world, and stand in a similarity relation to these systems. Giere argues that the value of correctness in his satisficing model of judgment depends on his account of representation, but I will show that his account of representation cannot be successfully used in this manner.

Simply stating that scientists value correctness over mistakes says no more than the interest theorists, who we saw argued that scientists use evaluative terms to distinguish their beliefs from one another. For a theory to avoid relativism, Giere demands that it demonstrate a causal

interaction between the scientists and the world. The point I will establish is that Giere's theory fails to meet this demand.

I argued in section 2 that Giere fails to establish that models are instantiated in the human cognitive system. He merely establishes that they are abstract entities. Now, the relation of similarity between an abstract entity and the world is not a causal interaction between a scientist and the world. Still worse for Giere's case is his claim that:

> ... one could say that the systems described by the various equations of motion [models] are socially constructed entities. They have no reality beyond that given them by the community of physicists. (p. 78)

And again: "... models are deliberately created, 'socially constructed' if one wishes, by scientists" (p. 93). This notion of model provides no opportunity for establishing a direct causal link between scientists (as opposed to their models) and the world. Such a direct causal link would be provided by certain of the scientists' cognitive mechanisms that operated in the same way for all scientists regardless of their differing social contexts. But Giere has explicitly invoked the scientists social context in his account, and this places him in the relativist camp, as I now show.

Giere says that models are socially constructed, and the admission of social constructivism into a theory has until now committed the theorist to relativism (see e.g. Woolgar 1988). Social constructivists, along with interest theorists, emphasize the context dependent nature of scientists' beliefs. Social constructivists make the further claim that scientists' representations originate from social interaction, and not any special interaction the scientists have with the real world. Giere could argue that his adoption of social constructivism does not render him a relativist, but a restricted realist. His realism might be established case by case, for each socially constructed model, by establishing its causal connections to the world. But the outcome in each particular case will be the product of the social group who negotiated the social construction in the first place. Further, in each case, as Giere admits and as can be seen from much of the literature in philosophy of science, there are usually at least two alternative models presented for any range of phenomena. Different groups of scientists socially construct different models to represent the same phenomena, and thus construct separate ways of representing reality. Therefore, if we combine Giere's account of socially

constructed representations, with any plausible account of diversity of opinion in science, even restricted realism ends in relativism.

Giere set out to defend his satisficing model as an explanation of how scientists value correct outcomes. He argued that the notion of correctness was grounded in his account of representation. This was supposed to provide the necessary causal link between scientists and the world, to give support to their valuing correctness over mistakes. I have shown in section 2 that Giere's account of models as representations did not establish them as part of scientists' cognitive makeup. And now I have noted that Giere claims models are socially constructed entities. The relation between socially constructed entities and the world is not the causally direct one required to ground the value of correctness. His model provides no basis for distinguishing the epistemic value of correctness as more important than any other value involved in theory choice. Finally Giere's position itself collapses into relativism, and so he cannot justifiably claim to have presented an alternative to relativism.

5. Giere's Cognitive Individualism.

Giere confronts many familiar problems in the philosophy of science, for example realism about theories and theory choice, and attempts to tackle them from a new perspective, a naturalistic perspective, which he calls a cognitive theory of science. I have argued that this theory fails in many crucial respects to live up to Giere's own standards for a cognitive theory. An example is its failure to characterize scientists' representational structures, and another is its failure to establish any cognitive mechanisms of scientific judgment. The one feature that is omnipresent in the theory is the approach of treating individual scientists as the units of its analysis. Giere argues that he can account for more scientific phenomena in terms of individual scientists than philosophers can in terms of relations between sentences, or sociologists can in terms of social relations. I argued in section 2 that he could not defend realism in terms of his cognitive theory, as the cognitive theory was not applicable to debates about semantic, or metaphysical relationships. I argued in section 3

that the theory was committed to relativism, leaving it in the same position with regard the role of truth in scientific decisions as the interest theory in sociology of science. Both these arguments cast doubt on the potential of this particular cognitive theory of science to successfully replace other theories explaining science.

I introduced the expression "cognitive individualist" (in Chapter 1) to describe philosophers of science who argue that a naturalized philosophy of science will best explain phenomena in terms of the cognitive capacities of individual scientists.[33] Giere fits this description at the level of his goals, but he does not successfully execute his program. His work illustrates the potential pitfalls of a cognitive individualist approach. The cognitive individualist approach places a great burden on the cognitive capacities of individual scientists. There are two separable issues here. The first involves the range of phenomena cognitive capacities are invoked to explain. The second involves the demonstrable limitations of our cognitive capacities. In conclusion I will briefly examine these issues.

The range of phenomena to be explained in science includes the force of epistemic values, say in the success of science and the process of scientific change. These highly complex phenomena are over- simplified when characterized entirely in terms of individuals' cognitive capacities. For example the phenomena of scientific change are not exhausted by an account of individual scientists' theory choice. The adoption of a new theory by the full range of scientific institutions previously endorsing its predecessor does not reduce to a case by case examination of each individual scientists' choice to adopt that theory.

The second issue here is that human cognitive capacities are actually limited. For example, human subjects consistently violate many established canons of right reasoning (see Kahneman, Slovic & Tversky eds. 1982, and Faust 1984). Much experimental work in cognitive psychology indicates that human cognitive capacities fall short of our previous expectations for them. Placing the burden of grounding epistemic valuation, for example, on a system (the human mind) that demonstrably does not reason well may not be a productive strategy.

---

[33] Analogously a naturalistic epistemologist who argues that important issues in epistemology will be explained by appealing to the cognitive capacities of the individual knower is a cognitive individualist.

Giere's cognitive individualist approach would be on stronger ground if it had some independent motivation. Giere does appeal to "empirical successes" in the cognitive sciences, but gives no details. He gives no examples of successful applications of a cognitive individualist approach, although he begins several of his arguments with the postulate that they apply to individual scientists. In his discussion of satisficing we saw that he postulated without argument that satisficing was the judgment strategy of individual scientists. The prior stipulation that individuals should be the units of analysis in a cognitive theory of science is not a sufficient motivation for the adoption of this approach. It also conflicts with a requirement Giere himself imposes on the study of science: "One should not put *a priori* restrictions on what might prove useful in explaining the phenomena of modern science" (p. 2). The undefended postulation of cognitive individualism is just such an *a priori* restriction.

Giere's book is ambitious, and this is one reason for its problems. He does recognize the need for a new approach to explaining science that will overthrow descriptively inadequate accounts in the philosophy of science. He further recognizes the need to integrate the work of several disciplines in achieving his overall aim. Giere may not provide an adequate version of the cognitive theory of science he calls for, but he does go further than Simon or Thagard in acknowledging the complexity required in such a theory.

CHAPTER 5

PAUL CHURCHLAND'S NEUROCOMPUTATIONAL PERSPECTIVE ON SCIENCE

1. Introduction.

Paul Churchland observes (1979) that a recent lesson in philosophy was that philosophy of mind could not be successful without paying attention to the philosophy of science. As philosophers came to realize that their theories of the nature of mind needed to keep pace with developments in psychology and neuroscience, they needed the resources of philosophy of science to assess these scientific theories of mind. In *A Neurocomputational Perspective* (1990) Churchland argues that the reverse is now true, philosophers of science must now pay attention to developments in philosophy of mind and the sciences of the mind. He claims that the important questions about science can

best be addressed by applying a neurocomputational theory of the mind. Like the other philosophers considered in this dissertation Churchland proposes a program to replace traditional philosophy of science and yet answer all the questions it posed, and more. Churchland seeks answers to philosophy of science questions in terms of a theory of the structure of the human brain, and models of the brain produced by neuroscientists and workers in parallel distributed processing in artificial intelligence. Churchland derives an account of the nature of theories, explanation, and conceptual change from computational neuroscience.

I argue that Churchland is mistaken in assuming that his approach rules out the need for sociological approaches to the study of science. Although Churchland may have provided a satisfactory account of the underlying representational structures possessed by individual scientists, this does not amount to an adequate account of the nature of scientific theories, explanation, and conceptual change. Churchland's accounts of the nature of theories and explanation is too all encompassing to have anything specific to say about science in particular, without appending sociological and historical accounts. As he acknowledges that to establish a naturalized epistemology one's empirical account of science must be satisfactory, this threatens the range of application of his proposed naturalized epistemology. Finally he has no clear way out of a relativism that renders his position consistent with that of social constructivists, which is a position he is likely to reject given his commitment to the idea of representation as progressively capturing the nature of the "real world."

2. An Outline of Churchland's Neurocomputational Approach.

In the last chapter of *Scientific Realism and the Plasticity of Mind* Churchland argues that epistemology needs to reject its advocacy of sentential representation, that is that "the current state of an epistemic engine is relevantly and adequately represented by a set of sentences or propositions" (1979 p. 127), in order to secure itself a viable future. Churchland's argument here is largely a negative one, and his positive proposal is merely the speculative one that the form of

non-sentential representation that will replace the sentential approach will be found by studying the nature of the human brain. In *A Neurocomputational Perspective* he introduces work in computational neuroscience that purportedly starts to fulfill the promise of providing an alternative representational structure for human knowledge.

The first half of *A Neurocomputational Perspective* is dedicated to matters in the philosophy of mind and the second section to matters in the philosophy of science. Although the first section provides useful background to the second it is not relevant to the central issues of this dissertation, and so I do not refer to it. In the second section Churchland proposes an account of representation, which he claims adequately accounts for the nature of scientific theories, the nature of scientific explanation, and the nature of conceptual change. The account of representation derives from work in computational neuroscience and parallel distributed processing research in computer science. In this section I provide a sketch of enough details of Churchland's account of representation to provide background to the more critical sections below.

Churchland's account of representation derives from work in Parallel Distributed Processing (PDP) in Artificial Intelligence (AI) and Psychology. We saw in Chapter 3 that Thagard also derived his account from work in PDP, but there is a crucial difference between Churchland and Thagard's account of representation. The computer networks that Thagard works with have nodes that correspond to proposition-like objects, hypotheses, evidence posits, and so on, but Churchland's aim is to propose a representational system that does not rely on sentential representation. Churchland's requires that nodes be understood as neurons, so the similarity between the networks and the human brain is highlighted. Representation is then distributed across the whole network, rather than being contained within individual units. Churchland describes an informative example of such an artificial network, which detects the difference between rocks and mines on the sea bed given an input of sonar echoes returning from the respective objects.

All networks are constructed from three layers of nodes, an input layer, an output layer, and a "hidden" layer. The representation of the input in the rock/mine network is as a vector of the activation levels (between 1 and 0) of the thirteen input nodes, and the output is represented as either of two vectors at or near $< 1,0 >$ for a mine echo and at or near $< 0,1 >$ for a rock echo.

All of the input nodes are connected to all of the hidden unit nodes whose activation strengths are determined by the number of connections, their size or *weight*, their polarity (stimulatory or inhibitory), and the strength of the incoming signals. The network initially has these activation levels randomly assigned and is trained to produce the correct output vectors by feeding it known input vectors and checking the output. When the input is a mine the weights of the hidden units are adjusted after each successive presentation of a mine echo until the output converges on $< 1,0 >$. Similarly the output for rock echoes is moved towards $< 0,1 >$. The network in question (See Gorman and Sejnowski 1988a, 1988b) was trained to a point where it could recognize the difference between ambiguous examples of input vectors, and also produced results for input vectors whose source was unknown. The important point for Churchland is not so much the success of this particular network, but how it represents.

Churchland claims that the "'knowledge' the network has acquired, concerning the distinctive character of mine echoes, consists of nothing more than a carefully orchestrated set of connection weights" (1990 p.167). The input layer represents thirteen aspects or dimensions of the stimulus to the network. The layer of hidden units, seven in this case, is important in mediating these combined activation levels to the output layer. Churchland has us consider

> the abstract space whose seven axes represent the possible activation levels of each of the seven units, then what the system is searching for during the training period is a set of weights that *partitions* this space so that any mine input produces an activation vector across the hidden units that falls somewhere within one large subvolume of this abstract space, while any rock input produces a vector that falls somewhere in the complement of that subvolume. (1990 p.168)

This then leaves the top half of the network with the job of distinguishing between these two subvolumes. The "knowledge" in the system is represented by these two subvolumes in hyperspace. But the system's knowledge can be represented in more detail on the same account. There are regions at the center of the mine-vector subvolume that represent *"prototypical* mine echoes," which will produce output vectors at or near $< 1,0 >$, and in the regions nearer the surface of the subvolume represent atypical or problematic echoes with output vectors such as $< .6,.4 >$ (ibid.). So far we have a way of representing prototypical and less prototypical cases of rocks or mines, but what Churchland requires is a more general account of representation.

Churchland introduces several other examples of networks, including one that turns printed words in English into speech, and another that detects the boundaries of objects from smooth gray scale pictures of the objects. Churchland concludes that the success of these networks indicates that

> it is plain that [they] have contrived a system of internal representations that truly corresponds to important distinctions and structures in the outside world, structures that are not explicitly represented in the corpus of their sensory inputs. (1990 p.177)

He continues as follows:

> The value of those representations is that they and only they allow the networks to 'make sense' of their variegated and often noisy input corpus, in the sense that they and only they allow the network to respond to those inputs in a fashion that systematically reduces the error messages to a trickle. These, I need hardly remind, are the functions typically ascribed to *theories*. (ibid.)

Finally he concludes against the sentential account of representation that "an individual's overall theory-of-the-world ... is not a large collection or a long list of stored symbolic items. Rather it is a specific point in that individual's synaptic weight space" (ibid.). This, then, is Churchland's account of the nature of theories, but he also promises to give an account of the nature of explanation and the nature of conceptual change.

These latter two accounts are parasitic on the account Churchland has provided of the nature of theories. He claims that

> to explain the phenomena of *conceptual change*, we need to unearth a level of subconceptual combinatorial elements within which different concepts can be articulated, evaluated, and then modified according to their performance. The connection weights provide a level that meets all of these conditions. (1990 p.178)

Although the partitions in a particular weight space correspond more to the common sense notion of concepts, Churchland argues that to analyze the "cognitive dynamics" of conceptual change in a revealing manner we must resort to the lower level of connection weights. But for the purposes of his own exposition of conceptual change Churchland sticks to the level of partitions of weight space. He argues that perfect identity of weight configurations would produce an identical partition of weight space, but "almost identical partitions" can be achieved with different weight configurations. In other words "synaptic contrasts in one place may compensate for further synaptic contrasts in another place, so that the functional profile of two brains may end up practically the same" (1990 p.234). For these reasons Churchland adopts the partitions in weight space as the "closest available neural analogue of what the philosophical tradition conceives as our 'conceptual framework'" (ibid.).

Churchland's account of explanation is the following:

> Explanatory understanding consists in the activation of a specific prototype vector in a well-trained network. It consists in the apprehension of the problematic case as an instance of a general type, a type for which the creature has a detailed and well-informed representation. (1990 p.210)

Churchland urges, against the possible objection that his account reduces to one of mere classification, that "prototype vectors embody an enormous amount of information" (1990 p.212). Prototypes in neural networks can have as many as 108 elements each constituting "one dimension of a highly intricate portrait of the prototypical situation" (ibid.). This points to a further perceived advantage of Churchland's account, "different people may have different levels or degrees of explanatory understanding, even though they classify a given situation in what is extensionally the same way" (ibid.). And this is explained by the richness and accuracy of that individual's prototype for a given situation. Finally Churchland claims that his account of explanation has the advantage of unifying the different types of explanation, for example causal, functional, and moral explanations. What unifies these disparate types of explanation is that they are all explained by prototype activation, but in each case a prototype of a different "character" is activated. Churchland offers a list of such prototype characters, for example property-cluster, etiological, practical, and superordinate. I return to these in section 4 below.

Underlying this whole perspective is the important assumption that the neural networks Churchland describes provide good models of the behavior of actual human brains. The difference between artificial networks and our brains on this account is the difference in the number of hidden units, the number of layers of units, and the number of connections on each unit. In each case the human brain has many more. Another important general theme in Churchland's perspective is that human brains are continuous with animal brains. Again this is on the basis of the assumption that each is being described accurately by the model, and so human brains are simply more complex, they have more individual points in weight space than animal brains. Churchland believes that if we properly understand the brain, we can understand the nature of all human cognitive endeavor, and even discover hitherto unknown, or inconceived of cognitive endeavor. Further he believes that the best way of understanding the structure of the brain is provided by neurocomputational models.

3. The Nature of Scientific Theories.

In this section I isolate two characterizations of theory from Churchland's account. The first is that theories are a point in weight space, and the second is that theories are any sort of guiding conceptual scheme. It is the second characterization of theories that is more pervasive and drives Churchland to his account of theories as points in weight space.

I argue that rather than giving an account of the nature of scientific theories, Churchland gives an account of the nature of individual representational capacity in terms of the conceptual schemes individuals can possess. If his account were one of scientific theories it would be a cognitive individualist one, but as it turns out there is nothing in Churchland's characterization that distinguishes scientific theories from any other conceptual schemes, for example a particular religion or common sense. Therefore Churchland has not provided an account of scientific theories, as distinct from any other kind of conceptual scheme. At most he has provided an account of the representational structure that might underlie any individual's displayed competence in the application of (part of) any particular theory. I return to the compatibility of this version of Churchland and relativistic sociology of science in section 5.

One of Churchland's central goals is to overthrow sententially based theories of representation, for example the logical empiricist's characterization of theories as lists of axioms from which predictions can be deduced. In place of this characterization of scientific theories Churchland proposes an account of theories as points in weight space. An individual's theory of the world is represented by the configurations of weights in its neural net. I have outlined the technicalities of this account in the previous section, now I highlight the overall view of theories that leads Churchland to propose this particular characterization.

Churchland variously proposes that theories are conceptual schemes, world views, ways of dealing with the buzzing confusion of sensory input. Theories can be anything from folk psychology to unified field theory, and from the representations an infant has to help itself avoid bumping into furniture, to the tutored common sense we might use to guide our choice of an election candidate.

Churchland concludes that "no cognitive activity whatever takes place in the absence of some theory or other" (1990 p.188). This characterization of theory is too all encompassing to generate an account of the nature of scientific theories. This is partly because Churchland characterizes theories purely in terms of any one individual's representational capacity, as opposed to focussing on the various uses of scientific theories, and their communicability. I generate this argument by first reviewing Churchland's reasons for rejecting the logical empiricist's account of theories, and noting that his objections are cognitive individualist ones.

First I must respond to a possible objection, which is that Churchland may accept that scientific theories are no different than any other conceptual schemes. I deal briefly with this objection here, and return to it when assessing the compatibility of Churchland's view with those of sociologists of science in section 5.

Churchland provides an adequate rebuttal to the charge that all theories should be treated as equally important in the following claim:

> ... our best and most penetrating grasp of the real is still held to reside in the representations provided by our best theories. Global excellence of theory remains the fundamental measure of rational ontology. And that has always been the central claim of scientific realism. (1990 p.151)

So he does want to maintain a distinction between scientific theories as representations of reality and other kinds of theories. Perhaps that distinction is merely one of degree:

> [My] perspective bids us to see even the simplest of animals and the youngest of infants as possessing theories...The difference between us and them is not that they lack theories. Rather, their theories are just a good deal simpler than ours in the case of animals, and less informed than ours in the case of human infants. (1990 p.188)

To rebut the objection, it is sufficient to establish that Churchland does acknowledge a difference between theories. Therefore it is reasonable to ask for an account that distinguishes scientific theories. As Churchland does have a normative agenda, perhaps the relevant distinctions are to be made in the methods of assessing theories, rather than in their intrinsic nature as representational structures, I return to this issue in section 6.

I turn now to Churchland's cognitive individualism and the role it plays in his rejection of the logical empiricist characterization of theories. Churchland claims to challenge the logical empiricist's characterization of theories, in particular he mounts an attack on an account of learning, which he attributes to the logical empiricists. This is "...the depiction of learning as the rule-governed

updating of a system of sentences or propositional attitudes..." (1990 p. 154). Yet he acknowledges that the problems beseting the logical empiricists are those such as the confirmation paradoxes. These are not problems with a theory of learning, but with an abstract presentation of the nature of theories, produced with a view to understanding their justification. The construal of the logical empiricist's characterization of theories as a theory of learning, allows Churchland to go beyond the usual list of acknowledged failures of the program by adding a new list of failures regarding its inability to account for the empirical constraints on a human learning subject. This leads Churchland to the hasty conclusion that we should consider a different form of representation for theories than the sentential. The problem with Churchland's argument is that it only works if indeed the logical empiricists', and for that matter Popper's, accounts of theories were intended as accounts of the nature of learning by individual humans. This is an unfounded assumption.

The logical empiricists produced accounts of the nature of theories that were designed to display the logical relations between the various components of the theories. Logical relations are characteristics of groups of sentences, and hence the theories were characterized as groups of sentences. Although it is fair for Churchland to point out that this enterprise is a failure, it is not because it embodied an inadequate account of human learning. Churchland's conclusion that we need to look for a level of representation beneath that of sentences is not supported by his attack on logical empiricist accounts of the nature of theories alone. It may be the case that an adequate account of how we represent and learn theories requires a level of representation beneath that of sentences, but this conclusion is not fairly established by the question-begging postulation of logical empiricist accounts of the nature of theories as accounts of learning. Churchland subverts the logical empiricists' program into a cognitive individualist program, primarily directed at individual learning. This may be what his own project is ultimately concerned with, but it was not the central concern of the logical empiricists.

By treating theories as a kind of conceptual scheme that guides all kinds of perception, Churchland renders theories as representational structures of individuals.[34] He claims several

---

[34] His view here overlaps with Giere's. Giere proposes models as the internal representations that stand for theories. Giere's view is more liberal as models need not be mental models, they can be abstract entities of some sort or other, but they do not rely on sentences to represent.

advantages for his account, first, that his account supports the fact that the human system can sustain a vast number of theories about the world around it. Human perceptual knowledge is both "highly plastic" and "theory laden." Second, that his account renders Kuhn's notion of a paradigm comprehensible. Third, it sheds new light on Kuhn's notion of resistance to a new theory. Fourth, it gives support to the claims of Kuhn and Kitcher, among others, about the importance of theoretical unity. Before going further let us compare this list of advantages to those cited by Churchland as being advantages of the logical empiricists' characterization of theories: the view "... made systematic sense of how theories could perform the primary business of theories, namely prediction, explanation, and inter-theoretic reduction" (1990 p.154). Churchland accounts for a theory's explanatory power, and I deal with this issue in the following section. He subsumes the other two logical empiricist goals under Kuhn's notion of extending and articulating a paradigm into novel domains, so before I turn to this issue I will assess whether Churchland's account really has the advantages he claims for it.

I argue that although Churchland's account may display some advantages, it is limited by cognitive individualism. Attempting to account for the notion of a theory exclusively in terms of individual humans' purported representational capacity cannot do justice to the role that theories play in scientific practice. First let us consider the claim about the potential multiplicity of theories that we can entertain.

Churchland assumes for the following argument that we accept his account of theories as points in weight space, and further that neural networks accurately represent the structure of our brains, to some close degree of approximation. He then claims that given these assumptions "in principle the human cognitive system should be capable of supporting any one of an enormous variety of decidedly global theories concerning the character of its commonsense Lebenswelt as a whole" (1990 p.190). The number of such theories is dictated by the number of distinct possible positions in weight space and adds to $10^{100,000,000,000,000}$ where "[t]his is the total number of (just barely) distinguishable theories embraceable by a human given the cognitive resources we currently command" (1990 p.190).

So Churchland claims that an advantage of his view is that it accounts for the plasticity and theory ladenness of human perceptual cognition. But this has little to do with the nature of scientific theories. We could grant Churchland all the above and still ask the question "What is a scientific theory?" Only if one concedes that a scientific theory is a way of informing our individual perceptions of the world, does this become an advantage for Churchland's account. The fact that we have an enormous capacity for varying conceptual schemes does not help us decide either way on the issue of whether such conceptual schemes provide an adequate representation of a scientific theory. If the claim is merely that scientific theories are such conceptual schemes then, in Feyerabend's phrase, "anything goes." Churchland's demonstration of the plasticity of the human perceptual system is orthogonal to providing an account of the nature of scientific theories, unless one assumes at the outset that scientific theories are conceptual schemes that guide an individual organism's perception.

Churchland claims that the second advantage for the view is that it helps make sense of Kuhn's notion of "paradigm." Kuhn is vindicated by considering the learner of a particular paradigm to be a brain taught (as neural nets are) by exposure to prototypical examples. For the brain to command a paradigm is for it to have settled into a particular weight configuration centered around the prototypes. Churchland argues that even the most reflective individual would find it hard to articulate explicitly the knowledge involved in the paradigm, and yet at the level of representation offered by considering neural nets this knowledge is in some sense accessible (even if only by a microscopic investigation of the brain's structure). According to Churchland his view demonstrates why learning by example is the most effective technique of learning science.

This may provide us with a way of cashing out Kuhn's insight at the level of individual learners, but it is much weaker than an account of the nature of scientific theories. Churchland merely demonstrates that he can provide an account of representation that is arguably consistent with Kuhn's account of paradigm. There are other accounts of representation that would fit this bill. The one Kuhn himself prefers is the prototype account of concept representation developed by Rosch, which may be ultimately compatible with Churchland's account, but need not be spelled out in terms of brain structure (See Kuhn 1977a). Also Giere's model-based approach could comply

with the requirement that paradigms be inarticulate, as a particular model could be learned by the presentation of examples, and need not be capable of linguistic articulation by its user. This supposed advantage of Churchland's account is therefore one that is shared by other accounts, and not by virtue of being an account of the nature of theories, but in virtue of being an account of human representational repertoires.

Churchland observes that "Kuhn makes much of the resistance typically shown by scientific *communities* to change or displacement of the current paradigm" (1990 p.192 my emphasis). He goes on to explain how neural nets can develop problems in learning new theories from examples, and he extends this to the problems individual humans have in similar cases, alluding to the case of sublunary and superlunary physics. What Churchland establishes here is that computational neural nets can have difficulty learning a new perspective even in the face of many new examples, but how does this relate to the obstinacy of a scientific community?

One could speculate that perhaps certain members of the community settle into the new theory more easily than others, but then we still have the problem of explaining the ultimate adoption of the new theory by the community. If it were simply a case of counting heads this problem might be soluble.[35] But if Churchland is seriously concerned with explaining the resistance of a scientific *community*, then he needs to appeal to factors beyond the learning characteristics of individual members of that community. Revolutions in the Kuhnian sense have often been revolutions of small numbers of people, who carry the day for a new theory against the background of a majority of detractors. The kind of explanation that might be required here is one that demonstrated how the power structure of a particular scientific community was transformed to allow the minority view to gain a foothold. The fact that we have in hand a theory of individual learning capabilities and deficiencies does not help us with this task. Certainly Churchland's account may help us to explain why particular individuals failed to adopt a particular theory, but in the historical cases it is more appropriate to turn to the kinds of documented evidence we have available, which is not likely to be about the brain states of the scientists concerned. So although Churchland may be able to point

---

[35] Kitcher gives an account of the chemical revolution that implies that it occurred once a sufficient percentage of the community were converted to the oxygen theory, which would fit this pattern of explanation (1990).

to the kinds of mechanisms underlying the general case of individuals' resistance to theory, the most revealing story in any particular case may still be told by historians of science given the evidence base.

I observed above that Churchland subsumed prediction and inter-theoretic reduction, two goals of logical empiricists' accounts of theories, under one new goal: unification. He puts this as follows:

> Tradition speaks of developing a single "theory" to everything. Kuhn speaks of extending and articulating a "paradigm" into new domains. Kitcher speaks of expanding the range of application of a particular "pattern of argument." (1990 p.193-4)

Churchland claims that his view can unify these two aims by thinking in terms of the evolution of the neural net to subsume all input vectors under one "similarity space." So the particular neural net under consideration would have something like a global theory of the world. Churchland claims that this kind of convergence is unlikely, and what is more likely to happen is the emergence of several global theories, and the one an individual learner settles on "may be a function of the idiosyncratic details of its learning history" (1990 p.194). But this does not amount to the claim that an advantage of the account is that it lends support to claims about theoretical unity. Again Churchland is making a claim about individual representational capacity, this time the claim is about the potential a particular neural net has to unify input under one conceptual scheme. The idea behind highlighting theoretical unity in accounts of the nature of theories is to point to a virtue one particular theory has over another one, this is Kitcher's aim (in his 1981). To merely assert that a property of a representational system is that it can (or has the potential to) represent a conceptual scheme that unifies input, does not shed light on the property of theoretical unity as a virtue in an account of the nature of theories.

My insistence that an account of an individual's representational capacity does not amount to an account of the nature of theories requires some independent motivation. A scientific theory can fulfill several roles, for example Laudan (1977) points to two: it can guide the solution of problems, or can provide a more general world view, such as that provided by the cluster of theories gathered under the head of "evolutionary theory." Less current scientific theories are collected together in textbooks, and students can learn to solve problems according to the dictates of these theories, although they may later learn new and more sophisticated versions of the theories, or even learn to

forget everything they had learned about a particular theory. Theories may be applied piecemeal, for example one can use simply the laws of motion, or one law of motion, in solving a whole host of mechanics problems. This activity may be an application of Newtonian theory, but the student need never know large facets of the theory's battery of techniques, such as the calculus and the universal law of gravitation. Some empirical research on the solving of physics problems indicates that, not only do physicists at different levels of training apply different parts of a body of theory, but they appear to represent problems differently according to their level of expertise (See Chi, Feltovitch and Glaser 1981).

In concentrating on individual representational capacities as the chief way of articulating the nature of scientific theories many of their salient features are left unaccounted for. Churchland's insistence that we must adopt a sub-sentential level of representation leaves him unable to account for facets of theories that arise from their very method of public representation, which is in sentential (and pace Giere pictorial) form. The fact that certain laws can be hived off from the main body and learned and applied separately is one such feature. Theories are public property in science, and this claim need not be cashed out by appealing to theories dwelling in some special realm (such as Popper's World 3, Popper 1972), but simply by pointing to their material embodiment in textbooks, computer memories and so on (cf. Hacking 1983 p.123).

My claim is not that Churchland's account is inconsistent with the above account of theories, it might be usefully appended to such an account with certain modifications. First, Churchland would have to drop the claim that he has accounted for *the* nature of scientific theories. Rather he has provided an account of a representational structure that can account for our ability to represent and apply scientific theories at all their levels of complexity. Second, he needs to acknowledge the communal nature of the enterprise of science. Science does not progress by the work of one sole brain in possession of a huge global theory. Partly due to the way theories are developed over time by many individuals, the work of science under their guidance is carried out by groups of researchers often using entirely different portions of what is later to become a more encompassing theory. Churchland's aim in providing an account of the nature of theories is not to establish that we can give an account of an ideal science carried out by one super-brain, so an acknowledgement of the

division of theoretical labor in science should not be an anathema to him. The fecundity of Curchland's proposals about the nature of the human representational system may lie more in its potential compatibility with more encompassing accounts of science than the cognitive individualist one he currently adopts.

4. Scientific Explanation or Explanatory Understanding.

Many of the weaknesses in Churchland's account of the nature of explanation parallel those in his account of the nature of theories, and result from his cognitive individualism. Instead of giving an account of the nature of explanation he gives one of the nature of explanatory understanding for any one particular individual. Further he claims that his theory "portrays explanatory understanding and perceptual recognition as being different instances of the same more general sort of cognitive achievement: prototype activation" (1990 p.197). I have outlined the way this claim is spelled out in more detail in terms of his neurocomputational account above in section 2. He intends his account to apply equally to explanatory understanding in pre-linguistic creatures and to sophisticated practicing scientists, thus falling afoul of the criticism that his account has nothing special to say about scientific explanation, a parallel problem to that effecting his account of theories, which had nothing special to say about scientific theories.

In this section I concentrate on his notions of prototypes, arguing that appeal to prototypes to account for explanation does not require an account of how the brain of any one individual works. I conclude that in concentrating on individual explanatory understanding Churchland either avoids the task of giving an account of scientific explanation entirely or shifts the burden of his account of scientific explanation to the evaluation of explanations, a project I evaluate in section 6.

Let us for the moment grant Churchland the contentious point that explanation can be subsumed under explanatory understanding, and assess his account of explanation as prototype activation on its own grounds. His point is that for any situation we want to explain we can activate a particular prototype vector that will account for the situation. My argument against Churchland

is that if prototype activation is the most important aspect of his account of explanation, then the account need not be couched in terms of the structure of individual brains. What is important about the account is what distinguishes the prototypes from one another and how effective they are at carving up various situations into manageable chunks. I develop this point by considering Churchland's exemplary prototypes.

Churchland lists six types of prototypes, which correspond to six types of explanation: property cluster, etiological, practical, superordinate, social interaction, motivational. He admits that the list is not exhaustive, and further that perhaps at least one of the members could be removed by subsuming it under another (motivational under etiological). The issue is whether the list of prototypes in any way relies on the previously outlined theory of cognitive dynamics. If Churchland generated the list by considering all the types of explanations he could envisage, then the list could have been generated by simply reflecting on explanations in both scientific and every day cases. But then the list is potentially infinite (or at least extremely large) and/or entirely arbitrary, and certainly does not require prior commitment to a neurocomputational account of representation.

A distinction will help to clarify this point, that is the distinction between the structure our cognitive architecture forces on our concepts or prototypes, and the manifest distinctions between different concepts or prototypes. Work on concepts and the graded structure of concepts in psychology is often ambiguous over this distinction (see e.g. Rosch 1978), but when the distinction is attended to the tasks for the psychologist become more complex. Barsalou (1987) demonstrates some of the difficulties confronting an investigator who is aware of this distinction. He distinguishes results obtained from experiments on the structure of concepts from the implications of those experiments for accounts of cognitive architecture.

Churchland's observations on the manifest distinctions between prototypes are not presented as the result of empirical observation. He simply postulates types of explanation, and hence prototypes that such an explanation would require to activate. This seems flagrantly question begging. First, Churchland proposes an account of cognitive structure that highlights the role of prototypes, then he arbitrarily generates a list of prototypes invoked in explanations, finally he claims that such prototypes are instrumental in explanation. Some issues left unexplained are what

distinguishes one set of prototypes from another, why simply subsuming an event under a prototypes counts as an explanation, and what distinguishes scientific explanation from any other kind of explanation. These issues are not addressed by Churchland's account of representation.

Let us consider the first issue of what distinguishes prototypes from one another. Churchland's open-ended list contains motivational prototypes and social-interaction prototypes, among others. But some accounts of the social aspects of activities, for example Thagard's account of social factors in terms of motivated inference (see Chapter 3 above), subsume social prototypes under motivational prototypes. Churchland himself points out that motivational prototypes could well be subsumed under etiological prototypes (those concerned with causal explanations). Churchland's division of prototypes simply suits his goals of accounting for several different, antecedently defined, types of explanation. The list could be expanded to include social science prototypes as opposed to natural science prototypes, and observational prototypes as opposed to theoretical prototypes. But to allow such expansions is to demonstrate the lack of specificity to scientific explanations in Churchland's account, because there is a prototype for every kind of explanation. Or certainly there are no distinctions between different types of explanation, except that for any new type of explanation there is a new prototype.

Notice, finally, that generating lists of prototypes does not rely on an antecedent theory of representational structure. Even if the generation of such a list were more controlled, for example by developing it experimentally following the Roschian program, it is a separate issue to establish what representational structure captures the prototypes in individual humans. Investigators who take prototypes seriously have suggested various sorts of representational structures, including the sentential ones that Churchland explicitly rejects (see e.g. Holland et al. 1986, Thagard 1989, Barsalou 1987, Kahneman & Miller 1986).

If it is correct that Churchland's account of explanation disintegrates into a process of proposing prototypes *ad hoc* for new situations that require explanation, then he has no distinctive account of scientific explanation. Recall that I observed earlier that a weakness of Churchland's account was that it subsumed explanation under explanatory understanding for any one individual. If we now add together these two factors, it appears that Churchland is claiming that explanation is any

application of an individual's prototypes, however idiosyncratic, to any particular new situation. This story is plainly inconsistent with the practice of science, which involves the use of shared explanatory strategies by particular communities.

It may well be the case that these explanations involve the application of prototypes, but why particular prototypes are favored over others, and why particular prototypes are shared by particular communities is not addressed by Churchland. The picture might not be so bleak for Churchland if he placed the burden of discriminating between types of explanation on their evaluation. This would allow him to retain his wide and non-discriminating account of explanation and add a normative story about the appraisal of explanations. Such an appraisal might contain a component that distinguishes scientific explanations from, say, religious or common sense explanations or those of the insane. Notice that if there is such a normative account it has to take up an awful lot of slack for Churchland, as he now requires it to distinguish explanations from one another and different types of theories from one another.

5. Conceptual Change.

In this section I argue that Churchland's account of conceptual change merely provides an account of what might underlie given historical cases of conceptual change in any individual scientist, but to grant him this is not to concede that he has provided an adequate account of conceptual change. First, he argues that conceptual change is conceptual redeployment, but often the interesting point of a conceptual change in science is that a previously unknown concept is used to explain a phenomena, more or less well known. Second, he argues that his view is consistent with those of sociologists of science, despite holding that it is the "world" that determines what our concepts are, and not society. The sociologists of science he refers to explicitly reject this point. Also there is good evidence that it is only through the mediation of other humans (say parents) that infants learn things about the world. Churchland's view of infants as little physicists or epistemologists is mistaken. Empirical evidence indicates that they are more likely little "social

interactors" (Gellatny & Rodgers 1989). Churchland's account of conceptual change, even if it were correct, leaves much of the work in an account of scientific conceptual change still to be done, by philosophers, psychologists, sociologists and historians of science.

Recall that Churchland claims that conceptual frameworks are partitions of the weight space (see section 2). He claims on the basis of this that he can account for sudden and discontinuous conceptual shifts in the following manner:

> The crucial idea is *conceptual redeployment*, a process in which a conceptual framework that is already fully developed, and in regular use in some domain of experience or comprehension, comes to be used for the first time in a new domain. (1990 p.237)

The examples of such events are Huygens' realization that optical phenomena could be better understood in terms of waves, Newton's unification of terrestrial mechanics and superlunar mechanics, and Maxwell's reconception of optical phenomena as electromagnetic phenomena. Churchland claims that in each case no new concepts were produced, rather previously established conceptual resources were applied to new domains. He goes on to argue that "...so many of the historical examples fit this redeployment mold that one may begin to wonder if history contains *any* examples of real conceptual novelty" (1990 p.239). He counters that there are, and offers Faraday's "fields of force" as an example, but claims that such cases are rare.

Churchland is guilty of at least exaggeration on this point. Cases of new concepts are not extremely rare at all, they seem to be rather commonplace especially in science and social science: gene, commodity, molecule, virus, black hole, computer, gravity, photosynthesis, phlogiston, temperature. There seems to be no end to this list, not that this ought to persuade us that *no* conceptual change takes place due to conceptual redeployment, rather that it will not account for all conceptual change by any means.

A deeper challenge to Churchland's account of conceptual change, especially as it occurs in science, can be mounted if we allow him his account at the level of individuals acquiring new conceptions of phenomena, and instead we focus on his account of how this kind of change is triggered or why it should come about. On this issue I maintain, Churchland has nothing useful to say, as he ignores what may be the most important determinants of conceptual change within scientific communities: the social interactions within those communities. To be able to account for

one individual's adoption of a new conceptual framework, is not to provide an account of a scientific community adopting a new conceptual framework. I now expand this argument.

Churchland reviews several candidates as mechanisms for triggering conceptual change. The first is "blind *luck*." Churchland's claim is that the fact that "Maxwell's EM theory should have yielded a velocity for EM waves exactly equal to the known velocity of light was the sheerest serendipity" (1990 p.241). But Churchland does not want to build his entire account around luck. He turns his attention to Kuhn's notion of gathering anomalies as responsible for triggering conceptual change. Phrasing this entirely in terms of an individual's search through conceptual space, Churchland analyses this process as the repeated attempts to reconfigure old input in the hope that it trigger an already established prototype in a subvolume of activation space previously developed to cope with entirely different phenomena.

This interpretation of Kuhn's insight internalizes a process that can be explained in much less mysterious terms. If we consider the process of coping with anomalies at the level of a scientific community, it is easy to envisage researchers looking through materials from related fields, to find new ways of coping with a particular set of phenomena, or perhaps turning to experimental techniques developed in a different area of research to their own area. An example such as the use of x-ray crystalography in biochemistry is an example of such a procedure. This is better accounted for by the interrelations between scientific fields, and the division of labor within scientific fields, than by attempting to analyze the procedure in terms of the reconfiguring of the brain of any one individual. My point here is not that individual's involved in a "paradigm shift" do not reconfigure their activation spaces, rather that in the case of scientific conceptual change more needs to be said.

Despite his attempt to analyze Kuhn in terms of activation spaces, Churchland is convinced that anomalies are not the real driving force that triggers conceptual change, he claims that they might not even be necessary. He argues that it is the "desire" for theoretical unity that triggers conceptual change (1990 p.241). Churchland gives these examples of conceptual change driven by the "desire" for conceptual unity: the move to apprehending heat as mechanical energy at the molecular level, the development of special relativity unifying mechanics and electrodynamics, and general relativity attempting to unify accelerated and unaccelerated reference frames. What underlies these unifying

moves is the "impulse toward conceptual unity [which] is vitally important in any cognitive creature" (ibid.). Any one individual's "impulse toward conceptual unity" is hardly significant if he or she is ignored by the rest of the community, again especially in the scientific case. The work involved in getting a case of conceptual change established and accepted within the community involves experimental work and the writing and circulating of papers persuading ones colleagues of the viability of the new view.

Churchland conjures up the image of scientists as having brains that are already filled with every imaginable conceptual possibility, simply tinkering with new data until they trigger off one of the available conceptual frameworks. Perhaps the reader is given this impression because of Churchland's use of the metaphor of "conceptual space." He describes the massive capability of human brains for holding different conceptual schemes simultaneously (see sections 2 and 3 above), which may well be the case, but this is only a space that is *a priori* capable of holding a vast number of conceptual frameworks. If any particular human, or artificial neural net, has only been trained to perform a specific task, its conceptual capability may still be quite vast (if Churchland is correct it is a function of the number of points in weight space), yet its actual stock of concepts is minimal given their limited training (*pace* Churchland, activation space is only minimally partitioned). The conceptual space in science that is already filled is held in libraries and computer memories as well as in large collections of actual brains (cf. Hacking 1983). What practicing scientists do is bring this store of conceptual schemes to bear on a current problem set. This can be achieved in various ways. For example, we need not rule out the possibility of one individual having mastered all the literature in two scientific fields and applying the theories of one to the other's recalcitrant phenomena. But we also need to explain the division of labor among various scientists within a field working on the consolidation of a new conceptual scheme. This cannot be done by describing the affinities of individual scientists' brains for conceptual redeployment.

Churchland could object to this line of argument by admitting that there is a far deeper sense of conceptual change, that which is involved in the learning process that produces the original partitioning of the activation space, the learning of new conceptual schemes. Here again Churchland gives an account in terms of the training of individual neural nets. He discusses various methods

used to train up networks, and then extends the discussion by analogy to the human case. One point he is keen to make is that the "world" plays an important part in training up the networks, and this fact undermines social constructivist challenges that scientists' conceptual development is socially determined. Churchland spends little time developing this argument, but it is worth spelling out his case in detail as there are several reasons for concluding that he does not have a convincing case against the social constructivists.

Churchland argues against the social constructivists as follows: We need to decide between the social determination of conceptual change or the robust influence of the "world itself" on the process. In the many networks developed by neurocomputationalist researchers it is clear that the world itself is driving the learning process. But such networks are not operating in the complex social world of real science, nor do they have the pressure on them to instantiate socially acceptable functions. On the other hand, given that in non-social cases of learning (artificial networks, simple animals) the world is the instructor, why would the case of science represent a total "corruption" of an otherwise systematically successful process. Science has outperformed the purer but simpler creatures. Therefore there is no reason to hold the "skeptical" position that conceptual change is socially determined.

I will concentrate on two weaknesses in the argument. The first is the premise that simple networks are trained by the world and that this implies anything about human conceptual acquisition, and the second is that science would be *corrupt* if it were to involve socially mediated conceptual change. I finally comment on Churchland's assertion that rather than a "skeptical" account of knowledge he requires a better one, where he equates skeptical with socially informed.

Recall Churchland's example of the network that learned to distinguish between rocks and mines. The network was fed a series of sonar echoes reconfigured into thirteen components, one to each input node. After each trail the network was "tweaked," its activation weights were adjusted, until it started to converge on an output that corresponded to the distinction it was required to make. It is hard to envisage this as a case of the network interacting with the world, if the world is supposed to be something separate from the network's human operators. First, the input was recoded and second, the network was adjusted by its instructor who had a clear idea of the desired

output. It is hard to see what hangs on the distinction between the "world" and "society," or even how it can be made from the perspective of a neural network. The network simply gradually partitions its vector space, and then applies the "concepts" it has learned, but this has not necessarily linked it up with the world. Given Churchland's adoption of a notion of theory-ladenness across the board (recall his claim that there is no perception without a theory), it is not consistent to attempt to invoke such a notion as the "world," which is in some way brute and unmediated. Certainly the appeal to the world does not naturally flow from Churchland's account, as he admits his account is consistent with the strong program in the sociology of knowledge (1990, p. 248-9). His account is as constructivist as that of the strong program, as he can potentially account for our acquisition of all kinds of concepts, whether they relate the structure of the world or not. Perhaps Churchland postulates the notion of a "world" to distance himself from the "corrupting " influence of sociological accounts of concept acquisition. But the claim that the sociological account corrupts what really goes on is itself an unsupported one.

Churchland's premise is that in nonsocial cases of learning the "world" is the instructor (a premise whose integrity we have questioned), and he concludes that unless "institutionalized science somehow represents a total corruption of a process that shows systematic integrity elsewhere, there is no reason to adopt [an] extremely skeptical, antirealist social determinism" (1990, p. 248). Churchland appends the following to his conclusion: "On the contrary, science has outperformed those purer but simpler creatures" (ibid.). One could take this supporting observation to imply an entirely different conclusion to Churchland's, that is, that the reason for the success of science is its very social and institutionalized nature, a nature that Churchland often explicitly acknowledges. Churchland presupposes a metric of scientific success something like "the tendency to expose the real nature of the world," and yet his account seems entirely inconsistent with such a presupposition. In fact his account is consistent with much social constructivism. Further, there is recent evidence from social psychology that indicates that from very early on in infancy, our world is being ordered by social concepts (Trevarthen & Logotheti, 1989).

Trevarthen and Logotheti's work indicates that from early infancy humans are inclined to communicate and attempt to interact with other humans. They also point to results showing that

babies first object fixations are with objects that are much used in the household, such as the telephone. Their work evidences the child's need to set up communication strategies as soon as possible, and no reciprocal need is found to explore the physical world of inanimate objects. Of course all this could be entirely consistent with Churchland's account of concept acquisition, and his account of the nature of theories and explanatory understanding, if we remove the claim that the "world" is our primary instructor. Recall also that Churchland makes much of the continuity between human infants and adults, as does the view he sets himself in opposition to. Not only do infants rely on social interaction to master their world, so do adult scientists.

If Churchland's view is consistent with a sociological account of concept acquisition, does this mean that his view can replace the sociological one? In the case of the study of scientific conceptual change the answer must be no, for Churchland has provided nothing in his account that distinguishes conceptual change in science from any other sort of conceptual change.

I argued above that Churchland has one avenue left open to him to distinguish scientific theories and explanations from other varieties, to wit the claim that they are evaluated by distinct methods. In the last section I turn to this normative component to Churchland's neurocomputationalist account.

6. The Possibility of a Normative Neurocomputational Philosophy of Science.

In *Scientific Realism and the Plasticity of Mind* Churchland argued that it would only be after a new account of representation were established that an account of knowledge could be provided. His goal was a naturalistic epistemology based in a theory of how the brain works. We have seen that he now has a theory of representation based on how the brain works, or at least on how models of the brain work. The tasks for his naturalistic epistemology are also apparent: the assessment of the best theories and the best explanations. Churchland's primary candidate for an indicator of good theories is "explanatory unification." Those theories that unify the most phenomena, and those explanations that unify broad ranges of explanada are the best. Churchland's normative

account arises directly from his account of representation, and I argue that it cannot perform the descriminatory role Churchland hopes for. One way the view might be bolstered is by appealing to the direct connection between the representations in our brains and the real world, but this has already been shown to be a problematic connection to establish.

Having adopted a new account of representation means that Churchland's stock of evaluative criteria is changed. He no longer has access to "the usual semantic vocabulary of reference, truth, consistency, entailment, and so forth" (1990, p. 220). As the "cognitive kinematics here being explored does not have sentences or propositions as its basic elements" (ibid.), Churchland claims that the "various dimensions of epistemic virtue" have to be reconceived.[36]

Churchland first proposes a notion of misrepresentation, which is that the wrong prototype is activated for the situation at hand. The correct prototype is the one that is reliably activated when one of the class of situations, of which the one at hand is a member, is presented to the organism. He then claims that, on his account, "correctness" can be distinguished from "warrant." "High warrant is a matter of low ambiguity in the input" (1990 p.221). In other words if the input vector is similar to one that would activate another prototpe, then the warrant for the current prototype is low. These evaluative criteria are limited as Chuchland concedes, because they are only applicable in individual cases. What he requires is a way of evaluating whole systems of prototypes, or theories.

The criterion Churchland proposes for evaluating whole theories is "conceptual unification." Churchland characterizes this in two different ways and I will concentrate on the characterization in terms of prototypes.[37] Churchland claims that a network given the luxury of too many hidden units may correctly recognize samples of a given F from the training set, but when presented with a hitherto unexperienced F, slightly different from those in the training set, it may not recognize it. (F's could be rocks or mines for example.) This is because it is able to develop a new protoype for

---

[36] Notice the similarity between Churchland and Giere's approaches on this point. Giere points out that in rejecting sentences as the main vehicle of representation he no longer has access to truth, and so he adopts similarity.

[37] I noted above in section 4 that Churchland resorted to the use of prototypes in his account of explanation. In his account of the nature of theories he gives a slightly different characterization of conceptual unity in terms of activation spaces and similarity subspaces, it appears that these two characterizations are consistent, the one he adopts in the account of explanation being merely at a coarser level of description.

each new, and slightly different, example presented to it. If the network is given fewer hidden units (implying less capability to subdivide its activation space) and "forced to continue learning until it finds a single prototype region" (1990, p. 222), it will be prevented from "*ad hoc* and unprojectible learning." Churchland argues that conceptual unification is an important epistemological virtue, because "cognitive configurations having that virtue do much better at generalizing their past experience to new cases" (ibid.).

Rather than offering a general account of conceptual unification as an explanatory and theoretical virtue, Churchland has so far only given evidence that artificial networks that can configure many similar examples into one prototype are better learners. He needs more to make his overall accounts of scientific theories and explanations distinct from accounts of any individual's conceptual scheme and efforts at understanding. Worse still, Churchland's account does not seem to have any obvious extensions to the case of human scientists and their theories. He may be attempting to make this connection when he evokes Kitcher's notion of expanding the range of application of a "pattern of argument."

Recall from section 3 that Churchland argued that his account of theory had the advantage of encompassing Kitcher's aims in developing the notion of expanding the range of application of a pattern of argument. I argued that since Kitcher's aim is to produce criteria for deciding between scientific theories and Churchland's account is about individual representation in general, there is no necessary overlap. But Churchland argues in defense of his account of conceptual unification that Kitcher's view is closely allied to his prototype activation view of explanation, because Kitcher himself appeals to prototypes. Churchland goes further claiming that the only drawback to Kitcher's view is its reliance on a linguistic conception of knowledge representation. Now if Kitcher is appealing to the notion of prototypes with his "argument patterns" he need not be appealing to prototypes in Churchland's sense of the term. As I argued earlier, the notion of prototype is separable from the issue of their representation within a given individual. What is common to a group scientific theories as they appear in various texts, and the mark by which a prototypical theory may be recognized, is likely to be a configuration of symbols organised sententially. Kitcher abstracts from such cases to uncover the argument patterns of various theories (Newtonian

Mechanics is one example he provides), at no stage need he resort to a level of representation below the sentential (Kitcher 1981). One clear advantage of Kitcher's view is that it relates specifically to the evaluation of scientific theories. One can assess two theories by abstracting their argument pattern and assessing the extent to which each unifies a given range of explananda. Churchland can claim that his account of representation underlies Kitcher's account, but he cannot help himself to any of the normative bite that Kitcher's account may provide without first providing some way of distinguishing prototypes that stand for scientific theories from any other prototypes a network might possess.

I argued above that Churchland provides no account of scientific theories or explanations that intrinsically distinguished them from any other conceptual schemes or explanations of phenomena an individual might have. I proposed that although Churchland might not be able to distinguish scientific theories and explanations in this manner, he might be able to by giving an account of the unique way in which scientific theories and explanations are evaluated. Churchland's account of epistemic virtue is derived from his prototype activation model of explanation, and includes no elements that distinguish the evaluation of scientific theories from one another, or from the evaluation of any prototypes on the grounds that they unify more phenomena. Churchland's attempt to allign his view with Kitcher's treatment of explanatory unification as a criterion of theory choice fails because Churchland's view is couched at too low a level of representation to discern between different scientific theories. So Churchland has no distinctive account of the nature of scientific theories or scientific explanation.

There is one last avenue that Churchland might consider open to salvage his case, and it is one that is hinted at throughout his work: that we have some form of direct connection to the real world. If it is the case that we are in some way directly connected to the world, then we may have some way of assessing the difference between those theories that relate the structure of the real world, and those that are concerned with other issues such as social etiquette. In the previous section I argued that the idea that we have some direct connection with the real world can be nothing more than an *a priori* postulate on Churchland's part. I also argued that any connection between our representations and the real world is greatly mediated by social interaction from early

infancy onwards. If these arguments are convincing Churchland seems to have no access to an appeal to our connectedness to the real world to ground his account of representations that are special to science, and hence his account of the structure of individual's representations does not amount to an account of the nature of scientific theories, nor the nature of scientific explanation.[38]

One set of concluding remarks are in order to dispel any impression that I am advocating some sort of idealism, they also act as a connecting passage to the final chapter. The point I make against Churchland is that there is no direct connection between our representational structures and the real world, no unmediated connection. This is not to say that there is no mind-independent world as an idealist would say. Rather there are no socially unmediated representations of that mind independent world. Two final considerations make my point clearer.

First, consider the vast range of representations available in current science, even within one currently practiced science. Each is intended to represent various parts of the otherwise undifferentiated whole we call the "real world." Models of the structure of large molecules ignore the structure of the individual atoms that combine together to form those molecules, and yet they are treated as satisfactory representations. The double-helix model of the structure of DNA was a representational and scientific breakthrough, and yet it is not reasonable to presume that it represents the actual structure of real DNA molecules in ultimate detail. And it is not considered a weakness of the double-helix model that it did not represent the ultimate structure of DNA molecules. Churchland's underlying concern that we are in some sense able to represent the real structure of the world shows an allegiance to a mistaken view of science inherited from the logical empiricists, that ultimately there will be one grand unifying and true theory of all reality. To separate Churchland's account from these aspirations is advantageous, not detrimental, and he admits this himself at some points (see e.g. 1990 pp.149-51, 193-4).

Second, Hacking has argued that it is only those objects that we create in order to produce an effect that we can call real (see his 1983). He is at pains to point out that such cases of intervening can be assessed separately from cases of representing. My argument that Churchland provides an

---

[38] The kinds of considerations raised against Churchland here and in the preceding section bear a resemblance to the classic "no independent access" argument that arises for empiricists. The argument is succinctly put in Russell's treatment of Berkeley (Russell 1959).

account of individual representation, and yet no account that is relevant to the special case of science is consistent with Hacking's argument.[39] Churchland's type of account could perhaps be usefully combined with a sociological and historical account of any particular incident in science, but will not provide an adequate substitute for such accounts as it offers no insight into the special nature of scientific practice. And as Churchland points out, there is no hope for an adequate naturalized epistemology without an adequate empirical account of the nature of science. My claim is that an adequate naturalized epistemology will only result from taking into account factors other than the nature of individual's representational structure.

[39] Hacking has recently argued that his view is consistent with radical social constructivists such as Latour and Woolgar (Hacking 1988, Latour and Woolgar 1979). He argues that their account of the social construction of knowledge applies at the level of representing, but not at the level of intervening.

CHAPTER 6

COGNITIVE SCIENCE OF SCIENCE AND NATURALIZED PHILOSOPHY OF SCIENCE

1. Introduction

In this chapter I recap and develop in more detail the arguments presented in the above four chapters. First I develop the notion of cognitive individualism in some detail. Second, I introduce some conceptual considerations about the nature of social relations in general and use these to organize and reconsider the idea that cognition may be social. Third, I discuss the place of normative theories of scientific inference in cognitive science of science. I conclude with several theses that are supported by the arguments in the dissertation as a whole, and which could act as guiding principles for a naturalized epistemology that is not cognitive individualist.

2. Cognitive Individualism.

Cognitive individualism is the thesis that a sufficient explanation for all cognitive activity will be provided by an account of autonomous individual cognitive agents. On this view what is cognitive will be circumscribed by what can be described in terms of the internal processes of such autonomous cognitive agents. The cognitive individualist collapses issues of the generation of beliefs and the process of providing warrant for those beliefs by locating both processes within the autonomous cognitive agents. For example in the case of scientific reasoning cognitive individualists claim that the process of hypothesis generation is an internal psychological process of individual scientists, as is the process of hypothesis evaluation by argument to the best explanation. Thagard's work provides the best example of this kind of approach, as he explicitly addresses hypothesis evaluation.

Some cognitive individualists allow that science is an activity involving many practitioners, but claim that the important cognitive output of science will be explained by a theory of the individual scientist's psychological processes. The social nature of science, for these researchers, is simply the gathering together of these autonomous cognitive agents into groups. The fact that scientists are collected together in particular groups (sub-fields, laboratory teams, paper collaborators etc.) is not to be taken into consideration when an account of the cognitive output of science is desired. Thagard, Giere and Churchland all acknowledge the social nature of scientific practice, and yet all three attempt to construct accounts of the important cognitive output of science in terms of autonomous individual cognizers. Thagard's conception of the social nature of science being captured by a model of various autonomous individual scientists sending their theories, or hypotheses to a central reviewer is a good example of this approach.

There is a parallel between the work of traditional philosophers of science (in the logical empiricist tradition) and internalist historians of science and the protagonists in cognitive science of science. Traditional philosophers of science hold that the success of scientific theories can only be explained on rational grounds, for example by appealing to the features of the theory, its

closeness of fit to the phenomena or its explanatory coherence. These philosophers lack an account of how scientific theories are embodied. Theories were treated as abstract entities that could be assessed independently of embodiment. Traditional philosophers agree that social or psychological factors have nothing to do with the justified acceptance of a theory or its eventual success.

Cognitive individualists have taken up part of the traditional philosophers' task with the added feature that they have found a new embodiment for theories: the psychological make-up of scientists, or in three of the cases considered, computer models of the scientists' psychological make-up. This approach is consistent with most research in the cognitive sciences where a cognitive phenomenon is accounted for in terms of an internal psychological mechanism. For example mechanisms have been proposed that underlie deductive logic, abductive reasoning, inductive reasoning, pattern recognition, past tense verb learning and so on. This dissertation aims primarily to challenge such a reduction of scientific cognition. Many of the cognitive products that cognitive individualists want to account for in terms of psychological mechanisms are the results of social interactions, for example the discovery of the ornithine cycle, the discovery of the Mendelian Laws of inheritance, the discovery of oxygen. These interactions are not analyzable into their individual components, and hence the cognitive output cannot be traced to the psychological makeup of particular individuals.

Consider the case of evaluating actual hypotheses in science. If it is the case, as the protagonists studied in this dissertation claim, that individual scientists have mechanisms that evaluate the best hypotheses, why is it the case that scientists work in teams, ranging in size from two or three member laboratory groups to fifty strong research teams to do just this job? Perhaps the cognitive individualist would reply that individual scientists have the capacity to evaluate hypotheses by themselves, but they need not, or cannot do all the work by themselves if others have this capacity, and more important there may be time constraints on producing results, so therefore they have to work in teams. There are several problems with this reply. First consider the practical and empirical constraints on any individual attempting to do all the work evaluating a particular hypothesis that leads to a scientific discovery. If it is both practically and empirically possible for an individual to carry out the tasks, then a sufficient empirical account of such a scientific discovery may be one that

gave an account of the psychological processes of an individual scientist. For given the time such an individual could do the work. But if it is practically and empirically impossible for any individual to produce the scientific discovery (due to considerations about varying individual cognitive competence or available memory space) the empirical status of the cognitive individualist approach is threatened. Now the cognitive individualist may avail herself of the following route: turn the claim into an "in principle claim," or claim that in "ideal circumstances" the individual could complete the task.

The claim that in principle an individual could carry out the required cognitive task is not an empirical claim. It may explain the right range of phenomena, but not under the right empirical constraints. It is irrelevant to the concerns of the empirical scientist that something could in principle be the case, given certain background assumptions, especially if the background assumptions are not empirically tenable. Such in principle claims could be turned into normative claims about how a particular task might best be done, given say an infinitely capable individual processing device, but such normative talk does not address the empirical issue of how a particular task was carried out, or how a particular cognitive output was produced.

A special case of this kind of in principle claim arises in cognitive science of science as computers are used to model human performance. The tendency is to run together the limits on the actual performance of certain computer programs and the limits on performance of human scientists. Certainly this tendency is to the fore in Simon and Thagard's work. It may be the case that the computer program can arrive at a result in a more efficient way than an individual scientist, but this does not imply that human performance is best explained as a poorly working instantiation of the successful computer program. But it turns out that in some cases when routine human cognitive tasks are reconstructed in straightforwardly computational terms quite the opposite is true. That is computers (even in principle possible computers) cannot complete the tasks in anything like the time that we can. Cherniak (1986) provides a striking example of this by having us imagine a computer checking a belief set for consistency using truth tables:

> Suppose that each line of the truth table for the conjunction of all those beliefs could be checked in the time a light ray takes to traverse the diameter of a proton ... and suppose that the computer was permitted to run for twenty billion years, the estimated time from the "big bang" dawn of the universe

> to the present. A belief system containing only 138 logically independent propositions would overwhelm the time resources of this supermachine. (1986 p.93)

As Cherniak points out, even given the problems in individuating beliefs, 138 seems a very low number to count as the total human belief set. In conclusion proposing that computers functioning in such a way should be treated as representing idealized cases of human performance is somewhat misguided. This illustrates a problem that must be addressed by any computer model of human cognition.

When developing a naturalized epistemology it is empirical constraints that must be at the forefront. The naturalized epistemologist faces the task of providing norms for the best cognitive performance given the empirical constraints on the agents involved. When it comes to evaluating hypotheses, one way that scientists deal with such constraints is by working in teams, and therefore dividing the cognitive labor. In ignoring this fact cognitive scientists of science overlook a crucial facet in the production of scientific knowledge. A contrasting approach to epistemology, android epistemology, proposes that we search for norms that guide the behavior of computational devices that are not constrained in the way human scientists are. Suggestions have been made for combining these two approaches, which in Chapter 2 I treated as inconsistent, by using computers as "prosthetic reasoning devices." In this type of scenario computers carry out tasks that are beyond the computational capacity of the human scientists, given the practical constraints (see Faust 1984, Fuller 1989, Churchland 1990).[40]

3. The Social Nature of Science.

There are many ways of presenting case for the social nature of science. In the previous chapters I pointed to cases in which social relations affected the cognitive output. Here I present a more general case for the social nature of scientific cognition. I first propose that the cognitive force of

---

[40] Faust calls the type of reasoning that computers could perform better, and in a shorter time frame, "actuarial reasoning." He is thinking of the reasoning invovled in making large statistical generalizations based on large data sets that arise in social science.

scientific theories cannot be adequately accounted for without considering their social embodiment. Second, I rebut the possible objection that it is only current science that produces social cognition, since the arrival of the research team approach after World War II, and that previous science could be explained by a cognitive individualist account. Finally, I present some thought experiments that give conceptual motivation to the claim that the cognitive output of science is largely socially produced.

One motivation for turning to social rather than cognitive individualist analyses comes from considering the complex ways in which scientific theories are articulated out of a great number of varied components. Recall that in Chapter 5 I challenged Churchland's account of the nature of theories by presenting many features of scientific theories that were not captured by treating them as individual's representational structures. Theories can be embodied in textbooks, they can be presented piecemeal, students can use certain parts of a theory without ever learning other parts, they can be general world views such as "evolutionary theory," which change their content over time (consider the differences in evolutionary theory between 1890 and 1990) and so on.

Throughout the dissertation we have encountered several different accounts of the nature of theories: schemata, models, points in weight space (or partitions in weight space). What all these accounts have in common is the idea that the notion of a theory is captured adequately by a representational structure within an individual's head. What we need to capture in any account of scientific theories is each particular theory's constantly changing nature across time, and the fact that scientists using the same theory need not have the same parts of that theory available to them. An account of the individual scientists' representational structures does not do justice to these features of scientific theories if the representational structure aims to capture the whole theory. Even the rare cases in the history of science in which a scientist conceived of and applied a whole scientific theory by himself, for example Einstein's development of special relativity, can be adequately accounted for without requiring that the theory exist only as an internal representation of an individual.

Cognitive science of science researchers' accounts of the representation of whole theories may in part be the legacy of traditional philosophy of science. Philosophers of science have traditionally

been concerned with the *post facto* justification of whole theories. This task traditionally required no concessions to the historical development of such theories. So in cases of evaluating the best of two theories both are formulated in their entirety and then assessed by their ability to cope with the relevant observations. This "static" view of theories (to use Giere's term) is the one that Thagard, Giere and Churchland all explicitly strive to overcome. But in their new accounts of theory evaluation the static picture is recreated, this time in the guise of internal mental representations. Thagard reconstructs Darwin's psychological process of comparing evolutionary theory with creationism, and Giere reconstructs the process of comparing continental drift theory with the preceding static theory.

One way of accounting for the permanent presence of a theory throughout generations of research is to acknowledge its embodiment in textbooks, research papers, computer memories and so on, all of which are the shared property of the scientific community. Once this fact is acknowledged, one is not faced with the problems of how one transfers one whole theory from the head of one researcher to the next. One can also attribute beliefs about a theory's truth to scientists, without having to assume that they have a representation of that entire theory in their heads, which they have assessed as best describing the world. If science is conceived of as a communal project involving the communication of knowledge and the accumulation of knowledge for use by other, perhaps future, generations of scientists, it becomes obvious that theories need some form of embodiment other than as internal representational structures of individual scientists.

It may be objected against this account that the embodiment for theories that I propose only accounts for tokens of scientific theories, as opposed to types, which are representational structures. Certainly Churchland, who wishes to reject the sentential mode of representation, could claim that the public embodiment of scientific theories does not reveal their ultimate nature. I have two responses to such an objection, both to some extent pragmatic. The first is that much progress can be made in the empirical study of science without solving the problem of the ultimate nature of human representational structure. We can answer questions about the transmission of knowledge and the application of particular theories to new domains without talking in terms of individuals' representational structures. The second is that even if we accept an account such as Churchland's

we still need to explain why theories are presented in their public form in the way they are. Perhaps it is more problematic to account for how theories are learned and internalized in a different manner to their public presentation. Churchland's own account of representation appears to be the most fruitful of the ones considered exactly because it leaves room for such issues to be sensibly addressed, given his account of learning (the way concepts and theories are acquired by individuals).

What all three accounts of scientific theories share is the implicit claim that one account will do for all theories. This claim is overly ambitious. The many ways that any one theory is manifest, at the same time or over periods of time, indicate that a universal account of "the nature of scientific theories" will be illusory. We may have to be satisfied with several different accounts of theories, each having their own particular appropriate ranges of application, as well as a variety of objects that count as theories.

I now turn to the general issue of the social nature of science. I first consider the objection that one is tempted to a social account of science merely because it has become so large scale. Of course, the objection runs, science is *prima facie* a social activity in the present day, because of the advent of large research teams, and this claim is unobjectionable. What is objectionable is the claim that scientific cognition in general is a social product.

Shapin and Schaffer (1985) present a strong case that scientific cognition was an essentially social product at the time of Boyle. Boyle required technicians to set up and operate his famous airpump experiments, but the crucial point is that these technicians were not able to report the results of experiments. The way that experimental results were established was by the invitation of a selection of appropriate witnesses, members of Boyle's own social peer group, to observe the experiments. This procedure was required to corroborate the experimental results. Any experiments carried out in the absence of suitable witnesses would simply have to be repeated in their presence.[41] This case reveals the depth of the claim that scientific cognition is a social product. The claim is not simply that science is usually practiced by groups of people, who given the right conditions may be dispensable, rather that scientific knowledge, the final validated outcome of scientific practice, is a

---

[41] A corollary to this case is provided by Shapin and Schaffer's observation that Hobbes believed that there could be no science in the state of nature, as without a properly ordered society there could be no communication.

result of certain crucial social interactions. I now provide some conceptual arguments supporting the claim that scientific cognition is a social product.

Asking the question: "What examples are there of lone scientists?" can help to sharpen intuitions about the social nature of science. The following thought experiments are based on this question. Imagine two women identically dressed in white coats both busily at work performing identical tasks in identical rooms. The tasks involved operations performed with various test-tubes full of liquid, the pouring of the liquid into a larger machine, and the writing down of figures displayed on a screen at the side of the machine. It turns out that one woman is involved in an important scientific experiment and the other is rehearsing a role in a play which depicts a post-holocaust society who perform the actions of experimental scientists as a religious ritual that invokes their pre-holocaust past. Neither individual woman defines her own context by her activities, it is only her involvement in the greater picture, the first in the biochemistry community and the second in the theater group, that provides the context for these activities.

This thought experiment addresses both the issue of the scale of investigations of science and the inescapably social nature of science. The context is what makes sense of the activity as a contribution to science; the context also provides answers to most of our questions about the significance of each of the two identical types of activity.

Consider another scenario in which two identical computer programs are running on identical machines connected to two larger identical machines in two identical rooms, both produce a print out of numbers and sentence fragments in English. One set of computer printout is used by an artist to form a continuous frieze down the entire spiral length of the Guggenheim Museum, and the other is taken by a group of scientists, who interpret its numbers and sentence fragments and submit them in the form of a paper to *Science* claiming that the computer has made a discovery. We are only led to conclude that the second computer is doing science by our discovery of its place within the scientific community (it turns out that the paper is published). The computer does not do science by itself, nor is it a representation of how science might be done by oneself.[42]

---

[42] A related point can be made by imagining how one could stop the women in the first example from being scientists. As we can only establish that they are scientists from their context, we can only prevent them from being scientists by manipulating their context. Also the computer's "work" can only be considered

There are at least two ways in which social factors can be involved in the production of knowledge each can be illustrated by a simple example. At a coronation the crowning of the queen only makes sense against the background of the beliefs of the nation in which she is crowned. The honor of being queen is bestowed not by the crown itself, but by a complex set of social norms that prescribe the place of the queen relative to her subjects. This example invokes a notion of social interaction that is shared with the thought experiments about the women who appear to be scientists. The case of the discovery of the Mendelian Laws of inheritance is a case in which this notion of social interaction comes into play. In many cases of scientific discovery it is the acceptance of a particular discovery, and the place it is given within a body of communally shared knowledge that marks it from similar episodes that are not accepted as scientific discoveries.[43]

There is a different sense of social interaction that underlies the ornithine cycle case (and the case of jury decision making). This is the following kind of interaction: If it were the case that a particular output could be provided by a group of people and yet it could not be analyzed in terms of the contributions of the individuals in a group (say by adding each of their contributions together), we could say that the relevant product was a social product. My claim about the ornithine cycle is that it is a social product in this sense. An illuminating analogy can be drawn here between this kind of non-additive social product and the non-additive nature of gene inheritance in biology.[44]

The following thought experiment captures this notion of social interaction: Marching bands often perform routines that leave their members distributed in such a way on the football field as to write the name of their team. The "output" of the band, the name of the team read by the fans in the stands, is not something that can be analyzed in terms of the "outputs" of individual

for publication once it is treated in the proper context. These types of considerations not only cause problems for cognitive individualists, but also for naive laboratory anthropologists/ethnographers, who have no theoretical considerations to fall back on concerning wider scale issues in science such as validation of discoveries and the honorific nature of facts.

[43] There is a similarity between the social interaction that leads to scientific cognition, and the social interaction that Putnam claimed contributed to meaning in "The Meaning of 'Meaning'" (Putnam 1975a). Putnam's notion of the "division of linguistic labor" influenced my thinking on the division of cognitive labor in science.

[44] This analogy was pointed out by Richard Burian.

band-members. Imagine a case in which the band wrote an obscene word on the field. We could only interpret this as a case of group defiance, or mischief. Perhaps extreme coercion on behalf of a small subgroup might produce the word, but the relevant output still cannot be analyzed in terms of the individual band members contributions.

To sharpen this example imagine what idealization, or idealizations, would enable an individual band member to produce the desired effect. Recall my arguments in section 1 against the possible appeal to idealization to defend the cognitive individualist against the attacks from a socializer of cognition. The point of this extension of the thought experiment is to highlight the difficulty of appealing to idealization, or in principle formulation. Arguments claiming that in principle an individual could produce a certain effect may not hold for cases of social interaction that conform to this model. An appeal to idealization has to be useful and informative, but if the idealization proposed can shed no light on the original phenomena it was intended to elucidate it is not useful.[45]

In this section I have highlighted several aspects of the social nature of science. First, I pointed to public embodiment of theories, which was necessary for their transmission across communities and across time. Notice that part of this aspect of the social nature of scientific theories is one that need not refer to people themselves, as theories are in part the books and computer programs they are embodied in. The notion of social at work here relies on the idea that such embodied theories are public, or communal property. Second, I presented an example of seventeenth-century science that relied on social interaction to produce scientific knowledge. Third, I developed some thought experiments to motivate the idea of a social context being important for the validation of scientific discoveries. Finally, I used a thought experiment to assist in the development of intuitions about social interaction leading directly to cognitive products. These considerations do not amount to a case for the claim that all scientific cognition is socially produced, but they do give convincing support to the weaker claim that the social aspects of science should not be ignored by any thoroughgoing empirical account of the nature of scientific cognition.

---

[45] Bach and Harnish make a related point in a different context when they argue that the idealization of a "frictionless plane" has obvious payoff and usefulness in physics, but a "surfaceless plane" is not a useful or informative idealization (Bach & Harnish 1982).

4. Role of Normative Theories of Scientific Inference In Cognitive Science of Science.

When the positivists, Popper, and the logical empiricists turned their attention from traditional epistemological concerns to the philosophy of science there was an important break from first person epistemology. Epistemologists were (and still are) attempting to answer questions such as "How are my beliefs justified?" and "Can I have certain knowledge of x?" The philosophers of science became (and still are) interested in questions such as "How is scientific knowledge justified?" and "What sets rational inquiry apart from other kinds of inquiry?" Cognitive science of science researchers attempt to provide accounts of the objectivity of scientific knowledge, and scientific rationality in terms of individual cognitive agents. They argue that cognitive science provides models of the thought processes of individuals and these models can be applied to scientific cognition in an attempt to explain it. The cognitive individualist's mistake is to take on the task of developing models of scientific inference (directly inherited from the above mentioned philosophers of science) and turn it into the task of producing models of individual cognition, understood as psychological processing. The cognitive individualist runs into trouble partly due to the conflation of two enterprises: the attempt to produce both a normative and a descriptive account of scientific inference. Further the cognitive individualist addresses issues that are more germane to first person epistemology, than to philosophy of science.

Traditional philosophers of science present various abstract models of scientific reasoning, for example those that guarantee the confirmation of scientific hypotheses and those that guarantee the deduction of predictions from a given theory. Formal mechanisms guarantee results in the same way that the formal mechanisms of proof theory guarantee that truth is preserved. A recent contribution to this research is Glymour's *Theory and Evidence* (1981), in which a theory of confirmation as "bootstrapping" is proposed. This work exemplifies the philosopher's lack of concern for descriptive issues, the aim is to produce a model of the abstract reasoning that would

lead to the confirmation of hypotheses. Facts about how such confirmation is achieved in practice are not brought into the discussion.

The issue of whether to take descriptive issues into account or not is an important one. If descriptive issues about actual scientific practice are not taken into account, then we are free to talk in terms of theories and hypotheses as formal structures. The forms such structures take need not be constrained by the limited capacities of the system that instantiates the theory, particularly the human brain or a scientific community. Such restrictions of instantiation may put unworkable restrictions on the development of formal theories. As we saw in Chapter 2 Glymour's way around this problem is to suggest that computers, rather than humans, present the ideal way to instantiate theories (Glymour 1987), because they lack the very kinds of limitations we have (they have larger memories and are faster at carrying out routine problem solving tasks). The important point to notice is that concrete limitations need not get in the way of these philosophical accounts, intended to guarantee the best results in ideal conditions.

Cognitive science of science researchers inherit the normative tasks of philosophy of science, but with a new aim in mind, the normative theory must be instantiated in individual cognitive systems. The idea is that hypotheses are confirmed by the psychological processes of individuals. As has been pointed out in the preceding chapters, this move poses serious problems for cognitive science of science researchers. Before I provide a general account of these difficulties, I briefly outline the descriptive task for cognitive scientists of science.

The kinds of phenomena philosophers of science try to explain are the construction, testing and evaluation of scientific theories, scientific theory change, and scientific discovery. When normative theories of scientific inference such as the logical empiricists' deductive nomological model of explanation are presented to explain these phenomena it is often objected that such theories are at odds with facts from case studies of actual scientific practice. Philosophers influenced by historians of science turned their attention to these case studies of actual scientific practice. For these philosophers normative issues played a subordinate role to descriptive issues, or at most held equal status with descriptive issues. Extreme conclusions from such case study work included Feyerabend's (1975) claim that Galileo's success relied on his behaving irrationally according to

philosophers' norms of rationality. Historical studies of science were augmented by sociological studies and psychological studies of science, also taking up the descriptive challenge. I focus on psychological studies.

Psychologists of science aim to investigate how scientists actually think when they produce a particular discovery, or confirm a particular hypothesis (see Tweney, Doherty, & Mynant 1981). Scientific discoveries were an early focus of psychology of science partly because philosophers had argued that scientific discovery was "merely" a psychological process, and so were of no interest to philosophers. Attempts to understand various cases of scientific thought led to the attempt to produce a theory that would explain all scientific thought. At this stage of the inquiry, case studies of scientific practice become data points for the various theories of the psychology of science. In Cognitive Science of Science work some of the objectives of psychology of scinece have been taken over, but there has been a tendency (explicit in some cases, e.g. Thagard 1988) to focus on successful scientific thinking, leading to the production of theories of successful scientific thought.[46]

The focus on successful scientific thought in cognitive science of science is analogous to the focus of cognitive scientists on correct inference in their attempt to provide a theory that explains human thought. Attempts have been made to base such theories on deductive logic, a theory of correct inference (see Sternberg & Smith eds. 1988). Cognitive science of science researchers following suit account for scientific thought in terms of theories of correct scientific inference (see Thagard 1988 Chs. 7 & 8, and Langley et al. Ch. 2). I now investigate the general problems caused by running together the normative and descriptive projects that occur in cognitive science of science.

One possible interpretation of the cognitive science of science researcher's agenda is encapsulated in the following hypothesis: H1: "Each individual scientist has psychological processes that embody methods of scientific inference that produce correct results." But this hypothesis has problematic consequences such as the following: If H1 is true, why do even expert scientists regularly violate canons of correct inference (Faust 1984)? Or why do scientists often take many years to make a

---

[46] In many sociological accounts of scientific practice an attempt is made to account for both successful and unsuccessful science, below I suggest an adoption of a new version of the Interest Theorists' "symmetry principle" (Barnes & Bloor 1981) that guides this approach.

seemingly trivial step in a chain of inference? The fact is that such a strong hypothesis does not match up with the facts of everyday scientific practise.

Thagard points out that a slightly different version of hypothesis H1 and its consequences have been discussed in the philosophy of logic. H1 is a version of what has been called "strong psychologism" (Haack 1978), in which the principles of logic are descriptive of our psychological processes. Analogous problems follow from this version of the hypothesis. An oft cited example is the difficulty in teaching deductive logic. More convincing evidence against the hypothesis comes from psychological experiments on deductive logic (see e.g. Wason 1977). The experiments show that people systematically fail to reason according to even the simplest of deductive rules.

Cognitive science of science researchers deny that they support H1, for example Thagard does explicitly and Simon implicitly. They both claim to support a weaker formulation of the relationship between correct reasoning and the psychological process of reasoning. Giere also holds that psychological processes approximate to some extent or other the processes that would be produced by principles of correct inference. Giere cannot hold H1, because by adopting Simon's notion of satisficing he has weakened the standard of what is to count as correct.

This weaker position is captured by the following hypothesis: H2: "There is considerable overlap between the principles of correct scientific inference and the psychological processes that produce such inference." Hypothesis H2 also runs into problems. Some of these indicate that H2 is no weaker than H1 in terms of dealing with empirical consequences.

H2 is challenged by the psychological evidence that trained experts fail to reason according to even the most basic principles of inference (see e.g. Tversky & Kahneman 1984, Faust 1984). How can these considerations be handled? An argument against the empirical results on bad reasoning practices is to argue that given the right conditions, people will reason correctly. But such an "in principle" argument appealing to "ideal conditions" defends H1 equally well, for example when reasoning correctly is treated as a level of "competence," contrasted with the level of performance reached in normal situations (cf. L.J.Cohen 1981). The strongest way to pose this argument is to postulate that logic is the representation of competent reasoning, but then this is to defend H1 and not the weaker H2, which was proposed to avoid the strong consequences of H1.

A second defense of H2 is to argue that people's actual reasoning practices improve with exposure to correct reasoning practises. This is a kind of reflective equilibrium defence, championed by Thagard (see Thagard 1988, Chapter 3 above). The projected culmination of such an equilibrium is that individuals reason according to principles of right reason. Even if this were to occur, it is not obvious that the ability to make correct inferences in one context can be transferred to another context (Tversky & Kahneman 1982, Nisbett & Ross 1980). To take this discussion seriously in the context of the attempt to explain scientific cognition is a mistake, as it presupposes that an account of scientific practice is exhausted by an account of scientific inference. This is the position of the logical empiricists, which cognitive science of science researchers and naturalizers in philosophy of science specifically want to reject. What I turn to now is the relationship of this discussion to issues in first person epistemology.

When considering the first person case, providing individuals (hypothetical or actual) with canons of right reason enables them to make correct inferences. For example canons of deductive reasoning ensure the truth of conclusions, given true premises. So the possession of canons of right reason can ensure the accruing of true beliefs by the individual. But when considering epistemic warrant appealing to canons of right reason is inappropriate. Particularly in the case of the warrant of scientific claims, as this approach fails to take into account the fact that the institutions of modern science are set up to corroborate, check, confirm and replicate scientific results (to say nothing of the fact that *no* canons of inductive inference lead to the truth). Now if it is argued that such institutions are required because individuals can be unreliable, then we can conclude that theories of right reason are descriptively inadequate as theories of the psychological processes involved in scientific inference. This therefore forces us to reject H1. For to explain the phenomena, such a theory requires a set of mechanisms to account for the varying reliability of the individual reasoners, but if all individuals reason correctly according to certain canons no such variation is possible. If individuals are claimed to be entirely reliable reasoners, then the implications for science policy are great: remove review boards, put an end to collaborative projects, and cancel all projects that replicate established experimental results.

The problem isolated in the above scenario is that much of normative task of philosophy of science is reconstructed by cognitive science of science of science researchers in terms of a task for first person epistemology. And further the normative burden, for example to explain the justification of theories, is being transferred to a descriptive account, which is presented exclusively in terms of the internal processes of individuals. Thagard's evaluation of hypotheses by inference to the best explanation, and Churchland's evaluation of theories in terms of conceptual unity are both examples of this type of reconstruction.

But the cognitive individualist approach (accounting for the canons of right reason by proposing that they are psychological processes) also poses problems from the perspective of first person epistemology itself. If our psychological processes guarantee the results of our reasoning, then there need be no more question about the certainty of our knowledge, such certainty is guaranteed by our psychological mechanisms. Epistemologists would be quick to point out that this approach is misguided. For example the "evil genius" argument works just as well at the level of psychological processes.

Perhaps those who attempt to account for scientific reasoning would have more success if they aped the sociologists' of science "symmetry principle" (Barnes & Bloor 1981) in the following manner: if it is the case that psychological processes produce scientific cognition (they are at least necessary conditions), then the same, or similar processes must underlie both "good" and "bad" science (assessed *post facto*). Otherwise we might be led to the conclusion that people who produce "bad" science are defective psychologically in some way, when the fact that such science is judged as "bad" is only with the benefit of historical hindsight. position given that Newton for example, produced as much "good" as "bad" science (physics and alchemy respectively). Adopting the new version of the symmetry principle allows us to look for psychological processes that underlie both good and bad science, and relieves the psychological processes of the burden of providing warrant for various beliefs. Churchland's account of representation already conforms to this new symmetry principle. Assuming, as I argued, that there is no special relation between our representations and the "real world."

The issue of warrant arises when we notice that all psychological processes are context dependent. The very notion of scientific cognition cannot be understood unless the social context of science is taken into account. The warrant of scientific claims only makes sense in the context of the scientific community. This social context involves not only the interelations of scientists, but the instruments, techniques for applying instruments, and interpretations of data produced that they share. As I argued in section 3 there is little sense to the notion of a lone scientist with a set of psychological processes that produce scientific cognition. Scientific cognition is produced by the community of scientists for the community of scientists and the greater community. The issue of warrant for beliefs only makes sense against this background of intersubjectivity and communication. If we were not so concerned that scientific beliefs contribute to our joint understanding of the world, we would be far less concerned with their justification.

Let me finally forestall one possible misconception. The argument so far should not imply straightforward division of labor between belief warrant and belief generation, the former being a social issue and the latter an individual psychological issue. The two processes are closely interwoven, as has been amply demonstrated by philosophers who exploded the distinction between discovery and justification (see e.g. Nickles 1980b), and by sociologists and historians who have demonstrated that discovery is both a psychological and a sociological process (see e.g. Brannigan 1981, Holmes 1985).

At the beginning of the section I proposed that the move from first person epistemology to the warrant of scientific beliefs set the framework for philosophy of science. Naturalized philosophy of science has the task of confronting the issue of warrant within a descriptive framework that does justice to scientific practice. Cognitive individualist cognitive science of science is not adequate to this task, and runs the further risk of reducing the issue of the warrant of scientific beliefs to issues of first person epistemology, a position philosophy of science, for example in its attempts to provide accounts for the success of whole theories, explicitly attempts to avoid.[47]

---

[47] My argument that we can only gain an account of the psychological processes of scientists by rejecting cognitive individualism also implies a rejection of cognitivism in cognitive science (Cf. Haugeland 1981 and Cummins 1983). This is the position that to be rational is to reason according to accepted canons of inference. This is a larger implication of my position that I acknowledge without attempting to defend, the defense of this larger implication will have to be left for another occasion.

5. Conclusion.

In the preceding chapters I have argued that the four bodies of work in cognitive science of science all fail to provide adequate accounts of scientific cognition because of their cognitive individualism. They attempt to account for scientific cognition that results from social relations in terms of the psychological processes of individual scientists. I have offered three reasons for the failure of the cognitive individualist accounts. First, there are *prima facie* cases of scientific cognition resulting from social interactions that are not reducible into individual psychological components. Second, the descriptive theories of human cognition attribute too much to individual human agents; psychologists have demonstrated humans to be cognitively deficient in ways that render such accounts descriptively implausible. Third, in making abstract reasoning to correct conclusions the object of inquiry the investigators have been misled into producing normative theories that are instantiable only in computers or other ideal cognitive agents. The result is that such theories do not account for any psychological processes that lead to failure or mistakes in science.

The arguments in the preceding chapters imply several theses that could provide the basis for a naturalized philosophy of science and more generally a naturalized epistemology:

Thesis 1: An adequate empirical account of scientific practice should highlight its social nature.

Two possible implications of this thesis are: 1a. The cognitive science of science is simply one part of a future comprehensive study of science that will include an account of the social dimension science. But this does not take into account the stronger implication of the thesis: 1b. The cognitive science of science research evaluated in this dissertation is for the most part irreconcilable with present sociological accounts of science, and sociologically influenced history of science due to its cognitive individualism.

Thesis 2: Factors currently understood as "cognitive" may consistently be understood as "social," calling into question the merit of a distinction between social and cognitive factors in science.

The evidence for claims about the cognitive nature of particular phenomena, does not establish them as specially "cognitive" in a sense that would distinguish them from social phenomena, without some sort of a priori theoretical commitment, such as commitment to cognitive

individualism. I have argued at length that many cognitive phenomena in science are social phenomena. Such phenomena are best explained, not in terms of the internal processes of individual scientists, but in terms of social interaction. Cognitive individualism places an explanatory burden on the internal processes of individuals that is simply not necessary given their high degree of social interaction.

> Thesis 3: Naturalized epistemology ought to be based on the assumption that humans are deficient reasoners measured by the standard of established canons of right reason.

Much work by cognitive psychologists indicates that human agents are cognitively deficient, to assume that science will be best explained by presupposing highly competent (or ideally competent) individual scientists goes against these findings. The model of science required must allow for the demonstrated inabilities of individual human agents, and yet account for the great success of the overall endeavor of science.

> Thesis 4: Features that are often deemed cognitive capacities of individuals are socially constructed and socially manifested. Specifically this is the case for logic, both inductive and deductive.

This thesis is consistent with parallel distributed processing accounts of representation in the cognitive sciences, and a radical form of eliminativism, and so it is consistent with Churchland's account of representation. As I have argued above the criticisms of cognitive science of science do not lead us to conclude that individual humans possess no internal representational structures, rather such representational structures probably do not mirror the structure of the features in the world that they represent. Churchland's view that our representational structures are highly plastic is consistent with this picture, assuming the way in which the internal representations are formed is conceded to be affected largely by social interaction as I argued in Chapter 5.

The thesis is consistent with parallel distributed processing accounts of representation, because artificial networks can generate certain kinds of cognitive behavior without prior familiarity with specific rules characterizing that kind of behavior. (For example deductive inference can be performed by a system with no rules of deductive inference built in (Bechtel, Forthcoming).) The thesis is consistent with eliminativism, because beliefs, theories, concepts, and "sentences in the head" understood as mental entities are not tenable on this view.

The arguments in the preceding chapters also point to ways research on scientific cognition might fruitfully progress. One suggestion that is consistent with my arguments, but which is not necessarily implied by them, is due to Latour, and this is to put a moratorium on all cognitive studies of science (Latour 1987a, p. 247). By cognitive studies of science Latour means to include all studies that focus on scientists' minds. He recommends a study of science that concentrates on what are commonly referred to as social aspects of science. My arguments do not force one to Latour's conclusion, as the thesis is not that the individual's psychological processes contribute nothing to scientific cognition, but rather that any theory that attributes all that is involved in scientific cognition to individual cognizers is mistaken. Individual scientists' psychological processes are capable of producing correct and incorrect scientific reasoning depending on the specific context. I conclude that to provide an account of scientists' psychological processes we require an inquiry that follows my revised symmetry principle, and takes into account the context of each scientist and group of scientists.

It may be the case that social studies of science currently provide the best descriptive accounts of various facets of scientific cognition, because psychology of science lacks a coherent research agenda once denied the cognitive individualist approach. A psychology of science that is consistent with the arguments of this dissertation would not challenge the central tenets of current social studies of science. These two latter areas of research will provide the empirical background for a naturalized epistemology that is not cognitive individualist.

BIBLIOGRAPHY

J.R.Anderson (1980), *Cognitive Psychology and its Implications*, W.H.Freeman, San Francisco.

R.J.Ackermann (1985), *Data, Instruments and Theory*, Princeton University Press, Princeton.

B.J.Baars (1986), *The Cognitive Revolution in Psychology*, Guildford Press, New York.

K.Bach & R.M.Harnish (1982), "Katz as Katz Can," *Journal of Philosophy*, Vol. 79: pp. 168-171.

B.Barnes & D.Bloor (1982), "Relativism, Rationalism and the Sociology of Knowledge," in Hollis and Lukes (eds.), pp. 21-47.

L.W.Barsalou (1987), "The Instability of Graded Structure: Implications for the Structure of Concepts," in Neisser (ed.), pp. 101-139.

W.Bectel (forthcoming), *Connectionist Models of Mind.*

W.E.Bijker, T.P.Hughes & T.Pinch (eds.) (1987), *The Social Construction of Technological Systems*, MIT Press, Cambridge, Mass.

A.Brannigan (1981), *The Social Basis of Scientific Discoveries*, Cambridge University Press, Cambridge, UK.

D.Bloor (1976), *Knowledge and Social Imagery*, Routledge Kegan Paul, London.

J.R.Brown (1984), *Scientific Rationality: The Sociological Turn*, D.Reidel, Dordrecht.

B.G.Buchanan & T.M.Mitchell (1978), "Model Directed Learning of Production Rules," in D.A.Waterman & F.Hayes-Roth (eds.).

T.Burge (1986), "Individualism and Psychology," *The Philosophical Review*, XCV, No.1.: pp. 3-45.

T.Burge (1979), "Individualism and the Mental," in *Midwest Studies in Philosophy*, University of Minnesota Press, Minneapolis, pp. 73- 121.

C.Cherniak (1986), *Minimal Rationality*, MIT Press, Cambridge, Mass.

M.T.H.Chi, P.J.Feltovitch & R.Glaser (1981), "Categorization and Representation of Physics Problems by Experts and Novices," in *Cognitive Science*, 5: pp. 121-152.

P.M.Churchland (1990), *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*, MIT Press, Cambridge, Mass.

P.M.Churchland (1984), *Matter and Consciousness*, MIT Press, Cambridge, Mass.

P.M.Churchland (1979), *Scientific Realism and the Plasticity of Mind*, Cambridge University Press, Cambridge, UK.

L.J.Cohen (1986), *The Dialogue of Reason*, Oxford University Press, Oxford.

L.J.Cohen (1981), "Can Human Irrationality be Experimentally Demonstrated?" in *Behavioral and Brain Sciences* 4: pp. 317-370.

R.G.Colodny (ed.) (1966), *Mind and Cosmos*, University of Pittsburgh Press, Pittsburgh.

R.C.Cummins (1989), *Meaning and Mental Representation*, MIT Press, Cambridge, Mass.

R.C.Cummins (1985), "Review of Fodor's The Modularity of Mind," The Philosophical Review, Vol. 94, No.1: pp. 101-108.

R.C.Cummins (1983a), *The Nature of Psychological Explanation*, MIT Press, Cambridge, Mass.

R.C.Cummins (1983b), "SOFT," in *The Proceedings of the Conference on Artificial Intelligence*, Oakland University.

D.Dennett (1988), "When Philosophers Encounter Artificial Intelligence," *Daedalus*: pp. 283-295.

D.Dennett (1987), *The Intentional Stance*, MIT Press, Cambridge, Mass.

D.Dennett (1978), *Brainstorms*, MIT Press, Cambridge, Mass.

E.Dietrich (1990), "Computationalism," *Social Epistemology*, Vol.4, No.2: pp. 135-164.

S.M.Downes, "Cognitive Individualism and the History of Science," manuscript.

K.A.Ericsson & W.L.Oliver (1988), "Methodology for Laboratory Research on Thinking: Task Selection, Collection of Observations, and Data Analysis," in Sternberg & Smith (eds.), pp. 392-428.

D.Faust (1984), *The Limits of Scientific Reasoning*, University of Minnesota Press, Minneapolis.

J.H.Fetzer (ed.) (1988), *Aspects of Artificial Intelligence*, Kluwer, Dordrecht.

P.Feyerabend (1975), *Against Method*, Verso, London.

J.A.Fodor (1987), *Psychosemantics*, MIT Press, Cambridge, Mass.

J.A.Fodor (1983), *The Modularity of Mind*, MIT Press, Cambridge, Mass.

J.A.Fodor (1981), *Representations*, MIT Press, Cambridge, Mass.

J.A.Fodor (1975), *The Language of Thought*, Thomas Y. Crowell, New York.

J.A.Fodor, T.G.Bever & M.F.Garrett (1974), *The Psychology of Language*, McGraw-Hill, New York.

S.W.Fuller (1989), *Philosophy of Science and its Discontents*, Westview Press, Boulder.

S.W.Fuller (1988), *Social Epistemology*, Indiana University Press, Bloomington.

S.W.Fuller, M.De Mey, T.Shinn & S.Woolgar (eds.) (1989), *The Cognitive Turn: Sociological and Psychological Perspectives on Science*, Nijhoff, Dordrecht.

P.Galison (1987), *How Experiments End*, University of Chicago Press, Chicago.

H.Gardner (1987), *The Mind's New Science*, Basic Books, New York.

A.Gellatly, D.Rogers & J.A.Sloboda (1989), *Cognition and Social Worlds*, Clarendon Press, Oxford.

B.Gholson, W.R.Shadish, R.A.Neimeyer & A.C.Houts (eds.) (1989), *Psychology of Science: Contributions to Metascience*, Cambridge University Press, Cambridge, U.K.

R.Giere (1989), "The Units of Analysis of Science Studies," in Fuller et al. (eds.), pp. 3-11.

R.Giere (1988), *Explaining Science: A Cognitive Approach*, University of Chicago Press, Chicago.

R.Giere (1979), *Understanding Scientific Reasoning*, Holt Reinhardt Winston, New York.

C.Glymour (1988), "Philosophy is Artificial Intelligence," in J.H.Fetzer (ed.), pp. 195-207.

C.Glymour (1987), "Android Epistemology and the Frame Problem," in Pylyshyn (ed.) pp. 65-75.

C.Glymour (1981), *Theory and Evidence*, Princeton University Press, Princeton.

C.Glymour & K.Kelly (1989), "Convergence to the Truth and Nothing but the Truth," in *Philosophy of Science*, Vol. 56, No. 2: pp. 185-220.

C.Glymour, K.Kelly & P.Spirtes (1988), "Philosophy of Science and the Logic of Doscovery," unpublished manuscript.

A.I.Goldman (1986), *Epistemology and Cognition*, Harvard University Press, Cambridge, Mass.

D.Gooding & F.A.J.L.James (eds.) (1985), *Faraday Reconsidered*, Stockton Press, New York.

M.E.Gorman & W.B.Carlson (1989), "Can Experimants be Used to Study Science?" *Social Epistemology*, Vol 3, No. 2: pp. 89-106.

H.E.Gruber (1981), *Darwin on Man*, University of Chicago Press, Chicago.

H.Guerlac (1981), "Lavoisier," in Dictionary of Scientific Biography, Scribner's, New York, pp. 66-91.

S.Haack (1978), *Philosophy of Logics*, Cambridge University Press, Cambridge, U.K.

I.Hacking (1988), "The Participant Irrealist at Large in the Laboratory," *British Journal for the Philosophy of Science* Vol. 39, No. 3: pp. 277-294.

I.Hacking (1983), *Representing and Intervening*, Cambridge University Press, Cambridge, UK.

I.Hacking (ed.) (1981), *Scientific Revolutions*, Oxford University Press, Oxford.

N.R.Hanson (1958), *Patterns of Discovery*, Cambridge University Press, Cambridge, U.K.

J.Haugeland (1985), *Artificial Intelligence: The Very Idea*, MIT Press, Cambridge, Mass.

J.Haugeland (1981a), "The Nature and Plausibility of Cognitivism," in Haugeland (1981b), pp. 243-281.

J.Haugeland (ed.) (1981b), *Mind Design*, MIT Press, Cambridge, Mass.

J.H.Holland, K.J.Holyoak, R.E.Nisbett & P.R.Thagard (1986), *Induction*, MIT Press, Cambridge, Mass.

M.Hollis & S.Lukes (eds.) (1982), *Rationality and Relativism*, MIT Press, Cambridge, Mass.

F.L.Holmes (1985), *Lavoisier and the Chemistry of Life: An Exploration of Scientific Creativity*, University of Wisconsin Press, Madison.

F.L.Holmes (1980), "Hans Krebs and the Discovery of the Ornithine Cycle," *Federation Proceedings*, 39: pp. 216-225.

G.Holton (1978), *The Scientific Imagination*, Cambridge University Press, Cambridge, U.K.

P.N.Johnson-Laird (1983), *Mental Models*, Harvard University Press, Cambridge, Mass.

P.N.Johnson-Laird & P.C.Wason (eds.) (1977), *Thinking*, Cambridge University Press, Cambridge U.K.

D.Kahneman & D.T.Miller (1986), "Norm Theory: Comparing Reality to Its Alternatives," in *Psychological Review*, Vol. 93, No. 2.

D.Kahneman, P.Slovic and A.Twersky (eds.) (1982), *Judgement Under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge, U.K.

W.Kintsch, J.R.Miller & P.G.Polson (eds.) (1984), *Method and Tactics in Cognitive Science*, Lawrence Earlbaum Associates, New Jersey.

P.Kitcher (1990), "The Division of Cognitive Labor," in *Journal of Philosophy*, Vol. 87: pp. 5-22.

P.Kitcher (1981), "Explanatory Unification," in *Philosophy of Science*, 48: pp. 507-531.

H.Kornblith (ed.) (1985), *Naturalizing Epistemology*, MIT Press, Cambridge,Mass.

T.Kuhn (1970), *The Structure of Scientific Revolutions*, University of Chicago Press, Chicago.

T.Kuhn (1977a), "Second Thoughts on Paradigms," in Kuhn (1977b), pp. 293-319.

T.Kuhn (1977b), *The Essential Tension*, University of Chicago Press, Chicago.

D.Kulkarni & H.A.Simon (1988), "The Process of Scientific Discovery: The Strategy of Experimentation," *Cognitive Science* 12: pp. 139-176.

I.Lakatos (1981), "History of Science and its Rational Reconstructions," in Hacking (ed), pp. 107-127.

I.Lakatos and A.Musgrave (eds.) (1970), *Criticism and the Growth of Knowledge*, Cambridge University Press, Cambridge, U.K.

P.Langley, G.L.Bradshaw, H.A.Simon & J.M.Zytgow (1987), *Scientific Discovery*, MIT Press, Cambridge, Mass.

J.Larkin, et al. (1980), "Expert and Novice Performance in Solving Physics Problems," *Science* 208: 1335-42.

B.Latour (1987a), *Science in Action*, Harvard University Press, Cambridge, Mass.

B.Latour (1987b), "The Author Responds: Latour to Oldroyd," in *Social Epistemology*, Vol. 1, No. 4.

B.Latour & S.Woolgar (1986), *Laboratory Life*, Second Edition, Princeton, New Jersey.

B.Latour & S.Woolgar (1979), *Laboratory Life*, Sage, London. L.Laudan (1977), *Progress and its Problems*, University of California Press, Berkeley.

J.Leplin (ed.) (1984), *Scientific Realism*, University of California Press, Berkeley.

P.T.Manicas (1987), *A History and Philosophy of Social Science*, Basil Blackwell, Oxford.

M.McCloskey (1983), "Naive Theories of Motion," in D.Gentner & A.L.Stevens (eds.) *Mental Models*, Erlbaum, New Jersey, pp. 299-324.

R.K.Merton (1973), *The Sociology of Science*, University of Chicago Press, Chicago.

M.J.Mulkay & K.D.Knorr-Cetina (eds.) (1983), *Science Observed*, Sage, London.

U.Neisser (ed.) (1987), *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*, Cambridge Uniiversity Press, Cambridge, U.K.

N.Nersessian (ed.) (1987), *The Process of Science*, Nihjoff, Dordrecht.

N.Nersessian (1985), *Faraday to Einstein: Constructing Meaning in Scientific Theories*, Nijhoff, Dordrecht.

T.Nickles (1987), "Twixt Method and Madness," in Nersessian (ed.) (1987), pp. 41-68.

T.Nickles (1980a), "Scientific Discovery and the Future of Philosophy of Science," in Nickles (1980b), pp. 1-63.

T.Nickles (ed.) (1980b), *Scientific Discovery*, 2 Volumes, Reidel, Dordrecht.

R.E.Nisbett and L.Ross (1980), *Human Inference: Strategies and Shortcomings of Social Judgement*, Prentice-Hall, New Jersey.

R.Nisbett & T.D.Wilson (1977), "Telling More than we can Know: Verbal Reports on Mental Processes," *Psychological Review*, Vol.84, No.3: pp. 231-259.

A.Ortony (ed.) (1979), *Metaphor and Thought*, Cambridge University Press, Cambridge, UK.

P.B.Paulus (ed.) (1989), *Psychology of Group Influence*, Lawrence Erlbaum, New Jersey.

A.Pickering (1984), *Constructing Quarks: A Sociological History of Particle Physics*, University of Chicago Press, Chicago.

K.R.Popper (1972), *Objective Knowledge; an Evolutionary Approach*, Clarendon Press, Oxford.

K.R.Popper (1968), *The Logic of Scientific Discovery*, Second Edition, Harper Row, New York.

H.Putnam (1988), "Much Ado About Not Very Much," *Daedalus*: pp. 269-281.

H.Putnam (1981), "Reductionism and the Nature of Psychology," in Haugeland (ed.) (1981b), pp, 205-219.

H.Putnam (1975a), *Mind, Language, and Reality*, Cambridge University Press, Cambridge, U.K.

H.Putnam (1975b), "The Meaning of 'Meaning'," in Putnam 1975a, pp. 215-271.

Z.W.Pylyshyn (1984), *Computation and Cognition*, MIT Press, Cambridge, Mass.

Z.W.Pylyshyn (ed.) (1987), *The Robot's Dilemma*, Ablex, New Jersey.

A.Rosenberg (1988), *Philosophy of Social Science*, Westview Press, Boulder.

E.H.Rosch (1978), "Principles of Categorization," in Rosch & Lloyd (eds.).

E.H.Rosch & B.B.Lloyd (eds.) (1978), *Cognition and Categorization*, Erlbaum, New Jersey.

P.A.Roth (1987), *Meaning and Method in the Social Sciences*, Cornell University Press, Ithaca.

B.Russell (1959), *The Problems of Philosophy*, Oxford University Press, New York.

H.Sarkar (1983), *A Theory of Method*, University of California Press, Berkeley.

S.Shapin & S.Schaffer (1985), *Leviathon and the Air-pump*, Princeton University Press, New Jersey.

R.C.Shank & R.P.Abelson (1977), *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*, Lawrence Erlbaum, New Jersey.

H.A.Simon & A.Ericsson (1984), *Protocol Analysis: Verbal Reports as Data*, MIT Press, Cambridge, Mass.

H.A.Simon (1977), *Models of Discovery*, D.Reidel, Dordrecht.

H.A.Simon & A.Newell (1972), *Human Problem Solving*, Prentice-Hall, New Jersey.

H.A.Simon (1969), *The Sciences of the Artificial*, MIT Press, Cambridge, Mass.

H.A.Simon (1966), "The Psychology of Scientific Problem Solving," in R.G.Colodny (ed.).

H.A.Simon (1957), *Models of Man*, John Wiley & Sons, Inc., New York.

H.A.Simon (1945), *Administrative Behavior*, Free Press, New York.

P.Slezak (1990), "Scientific Discovery by Computer as Empirical Refutation of the Strong Programme," *Social Studies of Science*, Vol.19, No.4.: pp. 563-695.

P.Slezak, Manuscript, "Bloor's Bluff."

R.J.Sternberg & E.E.Smith (eds.) (1988), *The Psychology of Human Thought*, Cambridge University Press, Cambridge, U.K.

S.P.Stich (1988), "Reflective Equilibrium, Analytic Epistemology and the Problem of Cognitive Diversity," *Synthese* 74: pp. 391-413.

S.P.Stich (1985), "Could Man Be an Irrational Animal?" *Synthese* 64: pp. 115-135.

S.P.Stich (1983), *From Folk Psychology to Cognitive Science*, MIT Press, Cambridge, Mass.

S.P.Stich & R.E.Nisbett (1980), "Justification and the Psychology of Human Reasoning," in *Philosophy of Science* 47: 188-202.

P.R.Thagard (1989a), "Scientific Cognition: Hot or Cold?" in Fuller et al. (eds.), pp. 71-82.

P.R.Thagard (1989b), "Explanatory Choherence," in *Behavioral and Brain Sciences*, 12: pp. 435-502.

P.R.Thagard (1988), *A Computational Philosophy of Science*, MIT Press, Cambridge, Mass.

P.R.Thagard & G.Nowak (1988), "Explanatory Coherence and Continental Drift," PSA Vol 1, Michigan.

C.Trevarthen & K.Logotheti (1989), "Child and Culture: Genesis of Co-operative Knowing," in Gellatly et al. (eds.), pp. 38-56.

R.D.Tweney (1989), "A Framework for the Cognitive Psychology of Science," in Gholson et al. (eds.) pp. 342-366.

R.D.Tweney (1985), "Faraday's Discovery of Induction: A Cognitive Approach," in Gooding and James (eds.), pp. 189-209.

R.D.Tweney, M.E.Doherty and C.R.Mynant (eds.) (1981), *On Scientific Thinking*, Columbia University Press, New York.

P.C.Wason (1977), "Self-contradictions," in P.N.Johnson-Laird & P.C.Wason (eds.), pp. 114-128.

D.A.Waterman & F.Hayes-Roth (eds.) (1978), *Pattern Directed Inference Systems*, Academic Press, New York.

W.C.Wimsatt (1980), "Reductionist Research Strategies and their Biases in the Units of Selection Controversy," in Nickles (1980b), pp. 213-259.

S.Woolgar (1988), *Science: The Very Idea*, Tavistock, London.

J.M.Zytgow & H.A.Simon (1988), "Normative Systems of Discovery and Logic of Search," *Synthese* 74: pp. 65-90.

Vita

Curriculum Vitae

STEPHEN MATTHEW DOWNES

Born: 21 Dec 1960, Bebington, Britain.

Academic Record:

1988-90 PhD Candidate, Science and Technology Studies, VPI & SU, (Awarded July 1990).

1985-88 Graduate Student, Philosophy, University of Colorado at Boulder, (A.B.D. 1988).

1984-85 M.A. candidate, Philosophy, University of Warwick, U.K., (Awarded July 1986).

1984 B.A.Hons (1st Class) Philosophy, University of Manchester, U.K.

Appointed as Visiting Assistant Professor of Philosophy, University of Cincinnati, beginning September 1990.