What Should a Connectionist Philosophy of Science Look Like?[1]

William Bechtel
Department of Philosophy
Washington Univeristy in St. Louis

The reemergence of connectionism[2] has profoundly altered the philosophy of mind. Paul Churchland has argued that it should equally transform the philosophy of science. He proposes that connectionism offers radical and useful new ways of understanding theories and explanations.

An individual's overall theory of the world, Churchland proposes, is not a set of propositions, but

> a specific point in that individual's synaptic weight space. It is a configuration of connection weights, a configuration that partitions the system's activation-vector space(s) into useful divisions and subdivisions relative to the inputs typically fed the system. 'Useful' here means 'tends to minimize the error messages'. (1989, p. 177)

In a connectionist network, it is the weights on the various connections that determines the response of the network to a particular input that is supplied by providing activations to a set of units. The response of a network is a pattern of activation over a designated set of units.[3] As a

[1] Support for this research was provided by the American Society for Cell Biology, the National Endowment for the Humanities (RH-21013-91) and Georgia State University, all of which are gratefully acknowledged. I also thank Robert McCauley for his comments on an earlier draft of this manuscript.

[2] This is not the place to present an introduction to connectionism, an approach to modeling cognitive phenomena that was first developed in the 1950s and once again gained prominence in the 1980s, in part with the publication of Rumelhart, McClelland, and the PDP Research Group (1986). Both Paul and Patricia Churchland have presented introductions to connectionism in various of their writings. See also Bechtel and Abrahamsen (1991) for an introduction. The understanding of connectionism varies significantly among authors. For both of the Churchlands the importance of connectionism seems to be that the parallels between connectionist networks and real neural networks allows our emerging understanding of brain function to inform cognitive modeling. For many others the importance lies not so much in the similarity of connectionist networks to neural architecture as in the fact that connectionist models seem to exhibit features of cognition lacking in other approaches to cognitive modeling.

[3] Often the units in a connectionist network will be layered so that a set of input units sends activations to one or more layers of other units, known as *hidden units*, which in turn send activations to a final layer of units known as *output units*. Such a network is known as a feedforward network. However, increasingly connectionists are pursuing architectures in which there are connections between units of a layer or feeding from a higher layer back to a lower layer. Moreover, there need not be any layering of units; units might be connected to all other units in the network or to a subset of them.

result of acquiring a certain set of weights (reaching a specific point in weight space), a network will learn to categorize input patterns into different groups, each member of which generates either the same or a similar response. Thus, what Churchland is terming a *theory* in a connectionist network determines the response categories of the network.

Within the range of responses that a trained network gives to members of a category, there is one pattern that constitutes the central or prototypical response for that category. No actual input the network has yet received may trigger this response, but it represents the central tendency amongst the response. Responses to actual inputs will tend to cluster around this prototypical responses. It is the activation of a response close to the prototypical response that Churchland proposes constitutes the system's explanatory understanding of its input circumstances. New examples are *explained* as they activate the same (or nearly the same) pattern: "I wish to suggest that those prototype vectors, when activated, constitute the creature's recognition and concurrent *understanding* of its objective situation, an understanding that is reflected in the creature's subsequent behavior". (208)

I shall briefly review Churchland's case for these radical construals of theories and explanatory understanding in part 1. What makes Churchland's view of theories and explanatory understanding novel is that it by-passes the sentential paradigm. One of my concerns in the rest of this paper will be with whether it is wise to by-pass the sentential paradigm so completely. While I am no fan of the sentential paradigm as an approach to explaining human cognitive activity and concur with Churchland that connectionism has much to offer philosophy of science, I contend that Churchland is mistaken in localizing the focus of philosophy of science exclusively in activities occurring in the heads of scientists.[4] While representations are key to scientific activity, the representations that matter are not exclusively mental representations. They are also external representations such as are found in sentences of natural language as well as in tables, figures, and diagrams. In part 2 I will argue that it is in terms of these representations that we need to understand the notions of *theory* and *explanation* and will explore how recognizing the role of these external symbolic representations in science changes Churchland's conception of the role of connectionism in modeling the cognitive activities of scientists. In parts 3 and 4 I will then attempt to illustrate this revised conception of a connectionist philosophy of science, showing how it might apply to actual cases of scientific research.

1. Churchland's Case for a Connectionist Philosophy of Science

Churchland advances his case for a connectionist philosophy of science partly by pointing to what he takes to be failures in sentential approaches that connectionism can overcome and partly by showing how a connectionist perspective provides positive accounts of such central notions in recent philosophy of science as simplicity, theory-ladenness of observation, and paradigm.

[4] On the other hand, making the reasoning activities of scientists part of the concern of philosophy of science is certainly an advance. In *Discovering Complexity* (Bechtel and Richardson, 1993) Richardson and I bemoan the fact that the standard models in philosophy of science have excluded the scientists and their ways of understanding and working to solve problems. Below I will show how one of the case studies we presented in that book lends itself to analysis within a connectionist framework. But the remedy to ignoring the minds of scientists is not to focus there exclusively.

Among shortcomings of the sentential approach Churchland identifies a number of well-known problems such as the paradoxes of confirmation which afflict the hypothetico-deductive framework, the problem of determining which among many propositions used to make a prediction are falsified in a Popperian framework, and the inability of probabilistic accounts of theory choice to account for the rationality of large-scale conceptual change. Except for some comments relevant to the last point, Churchland does not make it clear how a connectionist perspective overcomes these problems. Large scale conceptual change arises, in Churchland's connectionist framework, when a network gets trapped in a local minimum and must be bumped out of the minimum by an infusion of noise which significantly alters the current weights in a network. With luck, the network will then be able to find a deeper minimum. But there is no guarantee that a network will find a deeper minimum. While this account may characterize what occurs within a scientist as he or she undergoes large-scale conceptual change, it neither explains what causes the change (the sort of noise that will bump a network out of a local minimum) nor its rationality (especially since most such bumps fail to lead to deeper minimums).[5]

In "Explanation: A PDP Approach" Churchland raises two objections directed specifically at the deductive-nomological (D-N) model of explanation. According to the D-N model, explanation involved deduction from laws. But, Churchland argues, people often cannot articulate the laws on which their explanatory understanding is supposed to reside. Thus, explanation does not seem to require sententially stated laws. Further, people arrive at an understanding of phenomena for which they seek an explanation in much less time than it would likely take them to perform a deduction. Thus, Churchland questions both the need for laws and the appeal to logic as the way of relating laws to the phenomena to be explained.

The other objections Churchland raises against the sentential framework have to do with features of science that it cannot address. For example, the sentential view offers no account of learning to make perceptual discriminations or of learning to use the propositional system itself. Likewise, it cannot account for the learning of skills which, as Kuhn has argued, is just as important as learning the facts of a discipline. From a connectionist perspective, these are accounted for in the same manner as all other learning: through the adjustment of weights within the network. Further, the sentential perspective cannot explain how we retrieve relevant information in the process of reasoning about theories. This, it might seem, would require a massive search through all the propositions stored within the system. From a connectionist perspective, all stored knowledge is stored in the weights through which processing will occur. These weights *coarse code* the stored knowledge so that a particular weight will figure in the network's response to many inputs and does not provide a representation of a discrete bit of knowledge. But when a set of inputs similar to one on which a given response was learned is presented, these weights will cause the network to generate a similar response. Hence, the knowledge stored in connection weights automatically is brought into play whenever relevant.

[5] Further, as historians of science have pointed out, many major conceptual changes have resulted from scientists who had not accepted or acquired the previous conceptual framework. Often the revolutionary scientist comes from a different conceptual field and brings ideas from that field to the new field, ideas which are often in radical conflict with those of the existing conceptual framework. As a result in Churchland's terms, the revolutionary scientist never encountered the local minimum.

Finally, Churchland contends that sentential perspectives cannot explain the progress of science. In some sense it seems that current theories are closer to the truth than previous ones, and so many sentential theorists have tried to explain progress in terms of how new theories are better approximations of the truth. But Churchland contends that no adequate account has been developed of what it is to be closer to the truth. Indeed, the notion of *truth* has itself become problematic.

Among the positive features Churchland cites for a connectionist approach is that it provides an account of why simplicity is a cognitive virtue. He construes simplicity in terms of the number of hidden units in a network and points out that the ability of networks to generalize depends upon their utilizing the minimal number of hidden units needed for a particular problem. Networks with greater numbers of hidden units usually fail to develop weights on connections that generalize well. Rather, they will use different weights to generate the same response to different inputs that are supposed to be members of the same category, and thus will neither acquire the common category nor be able to extend it to new cases. Thus, he argues that the preference for simplicity can be understood from a connectionist perspective not simply as an aesthetic nicety, but as an important epistemic virtue: The point in weight space (i.e., the theory) found in simpler networks generalizes better to new cases.

Churchland also links the virtue of simplicity with the virtue of explanatory unity. He proposes that explanatory unity arises not from arranging theories in a deductive hierarchy, but from finding one set of weights that enables a single network to solve a multitude of problems. The virtue of this arrangement over using distinct subsets of weights to deal with each sub-type of problem is that this set of weights will allow the network to generalize to many new problems. These are problems whose inputs lie within the region (not necessarily contiguous) in the input space that generates that response. Generally this region will be larger than the sum of the regions defined by the separate sets of weights that might otherwise determine the common response; thus, a side benefit of using just one set of weights is that new cases (for which the network otherwise might have no solution) now fall within the region that will generate the same response as well.

Another feature that Churchland contends points to the strength of the connectionist framework is that it can explain features of science to which Kuhn (1970) had drawn attention: the theory-ladenness of observation and the role of paradigms in science. Since all processing in a network is determined by the weights, and these are constituents of the network's global theory, it follows directly that any processing of inputs by a network will be governed by its theory. Theory-laden observations is the expected case, not something requiring further justification. The notion of a paradigm was central to Kuhn's account. Normal research, for Kuhn, is directed by the paradigm with the goal of filling in the general perspective on phenomena encoded in the paradigm. But, as Churchland notes, Kuhn was severely criticized for the vagueness of this notion. Churchland contends, however, that connectionism provides a way both to make the notion both more specific and to explain why paradigms often seem vague:

> For a brain to command a paradigm is for it to have settled into a weight configuration that produces some well-structured similarity space whose central hypervolume locates the prototypical application(s). And it is only to be expected that even the most reflective person will be incompletely articulate on what dimensions constitute this highly complex and abstract space, and even less articulate on what metric distributes examples along

> each dimension. A complete answer to these questions would require a microscopic examination of the person's brain. (1989, p. 191)

Churchland thus sees connectionism as providing for a radical advance in philosophy of science. By understanding theories in terms of weights in a network and explanations in terms of prototypical responses of a network, he claims that we can overcome some of the problems that have afflicted the classical approaches that took theories to be sets of propositions and explanations to involve derivations from theories or laws.

2. The Use of Symbolic Representations in Science

An interesting feature of Churchland's construal of the deductive-nomological form of explanation as a major element in the sentential paradigm is his portrayal of its use in the mental lives of scientists. Thus, in commenting on the failures of the classical approach described above, he says:

> Those failures suggest to me that what is defective in the classical approach is its fundamental assumption that languagelike structures of some kind constitute the basic or most important form of representation in cognitive creatures, and the correlative assumption that cognition consists in the manipulation of those representations by means of structure-sensitive rules. (p. 154)

However, most of the positivists who formulated the classical conception of theories and explanation had very little interest in the representations used *in* cognitive creatures or in the nature of cognition generally. When they construed laws and theories as involving specific kinds of universal generalizations and explanation as a matter of deduction from these laws or theories, they seemed to be focused on the representation of laws in natural language inscriptions and deductions that could be carried out in terms of such inscriptions. Many of them took the extreme position of denying any interest in how scientists actually justified laws and theories or generated explanations, focusing rather on how these activities might be logically reconstructed to show the warrant of laws and theories or the adequacy of proffered explanations. (See Lycan's paper in this volume).

Churchland has moved the positivists' account of explanation inside the head of scientists and argues that it fails to give an adequate account of cognition. I concur that it fails to give an adequate account of the cognitive activities of scientists, but want to resist the idea that laws and theories are primarily representations in the head and that explanatory understanding is localized in internal cognitive activity. The common view is that scientists "write up" their ideas in papers for publication, and that they often find it most useful to present their ideas in figures and diagrams. What I want to suggest, however, is that these natural language representations and figures and diagrams are *not* translations of representations in the head (except insofar as scientists and other people often rehearse privately the sentences they will speak publically or image the diagrams they will draw, an activity that depends upon their mastery of the external representational systems). Rather, constructing natural language accounts of the phenomena they are studying and creating diagrams is part of developing theories and acquiring explanatory understanding of the phenomena. Constructing an explanation is an interactive activity involving both the cognitive agent and various external representational systems.

The view that I am advancing is suggested by Rumelhart, Smolensky, McClelland, & Hinton (1986) in their account of multiplying two three digit numbers. For most people, this is too complex a task to carry out in one's head. So, we "work on paper". That is, we represent the problem in a canonical form such as

$$\begin{array}{r}343\\ \underline{822}\end{array}$$

We then proceed in a step-wise manner, having learned to decompose the overall task into component tasks that are much simpler than the overall task. We begin with the problem 2 x 3, whose answer we have already memorized (have trained up our networks to solve). As a result, we write *6* directly beneath the *2* and the *3*:

$$\begin{array}{r}343\\ \underline{822}\\ 6\end{array}$$

We then proceed to the next step, multiplying 2 x 4, etc. What is important here is that we solve the problem by working interactively with external representations. In fact, as long as we have learned the procedures for approaching such a problem, we do not need any internal representations to solve this problem.

Elsewhere I have argued both that our ability to make logical inferences (Bechtel & Abrahamsen, 1991, Bechtel, in press a) and to use natural languages (Bechtel, in press b) have a similar character. Focusing for now on language, what I propose is that we learn to use the symbols of a language as they are initially embodied in an external medium (sound, print, etc.) as representational devices. These symbols afford concatenation in various ways and so we also learn the ways of concatenating these symbols employed in our native language as we participate in interpersonal communication. According to this view, our knowledge of grammar, for example, may consist in knowledge of *procedures* for comprehending and producing sentences in spoken or written speech. The grammar is not, as often thought, a set of rules for operating on internally represented strings of symbols. Just as a Turing machine consists of a finite state device, which employs a finite set of rules to determine responses to different possible situations, supplemented by a potentially infinite tape on which symbols are written and from which they are read, so the cognitive system may possess sets of procedures which enable the system to produce and comprehend spoken or written symbols.

Can syntactically correct speech by processed by a system that does not process internal symbols according to explicit rules? In addressing this question we should bear in mind that we are only addressing linguistic performance, not competence. Moreover, performance is generally far better in written language, in which it is possible for the user to backtrack and correct errors, than it is in spoken language. Finally, sophisticated language users are able to rehearse their performance privately, storing a trace in echoic memory, before actually producing it.[6] This process does allow some error correction. So, in evaluating whether a system that does not manipulate internal symbols in rule governed ways can account for linguistic capacities, we need to be careful not to overestimate what is required. Nonetheless, accounting for linguistic abilities

[6] This private use of language is, as Vygotsky (1962) argued, dependent upon having mastered the public use of language, and involves going through most of the steps of actually producing a public utterance, and then responding to the product that is generated (the sounds of inner speech) as input.

constitutes a major challenge for connectionism. St. John and McClelland's (1990) network for sentence comprehension, however, suggests how a network might learn to respect grammatical constraints in a natural language. It learned to develop case role representations from an impressive variety of sentences in which the correct interpretation often depended on grammatical structure as well as semantic structure.

What I want to emphasize here is that, if this account is correct, we can view natural language as providing a powerful extension to our cognitive capacities. It provides an external representational system which allows us to encode information that can be useful in guiding our actions. For example, we can write down the steps in a procedure we do not execute very often (e.g., a recipe for chicken marbella) and extract from the written representation the information we need when it comes time to execute the procedure again. We can also record descriptions of events that have transpired so that we can consult these records again. These written documents are not transcriptions of our mental representations, but specifically constructed representations with which we have learned to interact. In fact, these linguistic representations possess features that may not be found in our internal cognitive representations. For example, written records can endure unchanged for extended periods of time whereas our internal "memory" appears to rely on reconstruction, not retrieval of stored records. Moreover, through the various syntactical devices provided by language, relations between pieces of information can be kept straight (e.g., that a tree fell and a person jumped) which might otherwise become confused (e.g., when linked only in an associative structure such as a simple connectionist network). Thus, by acquiring language a system acquires capacities to represent information and use that in its own activities which it would otherwise like. (In this discussion I have abstracted from the use of language in coordinating behavior amongst agents. This is a not insignificant further role that language plays that is crucial to the operation of science.)

Given the potential usefulness of language as a representational system, it is natural that scientists should avail themselves of it. And of course they do in their conversations and publications. But I would contend that language plays a much greater use than merely a medium for transmitting ideas to other scientists. It also is the medium in which laws, theories, and models are often developed. The first two of these ideas are familiar from positivist philosophy of science. Laws are universal counter-factual supporting generalizations, while theories also consist of universal statements that are supposed to provide a more fundamental account (i.e., in terms of more basic entities and processes) from which laws can be derived. Recently a number of philosophers have charged that laws do not play quite the role that they have generally been taken to play in positivist philosophy of science. Cartwright (1983) for example, argues that strictly interpreted, most laws of physics, let alone the special sciences, are false. They present an idealized structure that is never quite realized in the natural world. Other philosophers such as Giere (1988) have argued that most science involves less abstract entities which he calls *models*. In part, models are more concrete as a result of filling in specifications of parameters left unspecified in more general laws.

Starting not from physics but from the life sciences, one is impressed with how infrequently scientists appeal to structures having the character of laws. The challenge is not to subsume individual cases under generalizations. Rather, scientists frequently find themselves confronted with a system in nature which generates a certain kind of output. What they want to know is how the system uses the inputs it receives to generate this output. For example, biochemists seek to detail the substrates and the enzymes that operate upon them in the course of performing a physiological function. Thus, scientists construe the systems they are confronting as machines, and their goal is to determine how these machines work. This involves identifying the parts of the machine and the contributions each makes to the performance that is of interest. The result is not a set of laws, but rather a blueprint or model of a possible machine that is thought to be capable of carrying out the activities of the system in nature. There are some notable respects in which models are different from laws. First, they are often quite particular, describing how specific components behave in specific circumstances (e.g, the enzymes and cofactors involved in fermentation under specific conditions in a particular species of yeast cells), not general. Second, they are often incomplete, and sometimes known to be incomplete. For example, Mitchell proposed the basic chemiosmotic mechanism of oxidative phosphorylation in the 1960s and it was widely accepted by the 1970s despite the recognition that parts of the account were incomplete and required further investigation. Third, characterization of entities and processes in models are sometimes known to be false but to nonetheless provide a useful basis for initially understanding and reasoning further about a phenomenon, as in Kauffman's (1993) models of gene control networks (see also Wimsatt, 1987).

I will return to this notion of a model in the next section when I consider how connectionism can help us understand the process through which scientists construct such models. What I want to stress is that scientists frequently formulate their models linguistically: they name the various actors in the process and describe the transactions they envision as occurring. Language provides one way of representing these actors and processes, diagrams provide another. (We should note that diagrams are often labelled linguistically, and are many times uninterpretable except with linguistic commentary.) Therefore, we should not attempt to develop our whole account of what scientists do by focusing on what is going on in their heads, but consider as well how they manipulate these external representations that are crucial to science.

Churchland is not unaware of the relevance of public discourse in science. Indeed, in his closing remarks in "On the Nature of Theories" he comments:

> It remains for this approach to comprehend the highly discursive and linguistic dimensions of human cognition, those that motivated the classical view of cognition. We need not pretend that this will be easy, but we can see how to start. We can start by exploring the capacity of networks to manipulate the structure of existing language, its syntax, its semantics, its pragmatics, and so forth. (1989, p. 195)

The approach Churchland outlines for dealing with this public discourse seems to be close in spirit to what I have sketched above. What I want to contend, though, is that this is not a task we can put off until we have worked out notions of theories and explanatory understanding. Scientific theorizing and explanation depends upon this external representational capacity. The weights on connections and activation patterns in the head are only part of the story. They potentially can explain how scientists are able to employ these representations and specifically how they relate them to particular phenomena to be explained. If the connectionist account of cognition is correct, it has a major role to play in our understanding of scientists. But so do linguistic structures, diagrams, and other external representational systems.

Before leaving this issue there is one objection that Churchland makes to a related position that needs to be addressed. Namely, he argues that focusing on the public level fails to account for important differences between individuals. Churchland himself raises this when he considers whether, in his connectionist account, one should identify a theory with a particular point in weight space or with a set of points that would produce the same classifications of inputs. He objects that the latter fails to account for the differences in dynamics between agents/systems. The details of how a particular network solved a problem are important for determining how that network will generalize and learn in the future (1989, p. 177). This information is lost when we move to the external level, either by focusing on how the network partitions inputs or by looking to the external representations that the network might use. I grant part of Churchland's contention: the process of conceptual change is partly determined by the details of internal processing. Change those details and a different trajectory will be followed. But this is not all that will determine future trajectories. External representations also affect future trajectories. The same model of a system can be presented in different ways even within language. These different ways will highlight different aspects of the model. What is highlighted is particularly important when the model must be revised, for those are the features one is likely to change. Further, different sorts of external representational systems afford different manipulations and restrict others. Stereoscopic representations of a system may make salient possible alterations that are not apparent in non-stereoscopic representations. For example, since one is aware of three dimensional relations, one may see multiple ramifications of altering a part that would not otherwise be noticed. The process of theory revision is thus controlled at many points, both at points within the mental processes of scientists and at points in a theory's external representations.

3. Model Building as Multiple Soft Constraint Satisfaction

Having allowed a central place for symbolic representations (albeit external, not internal ones), one might wonder what role connectionism might still play in altering our philosophy of science. To see this, we need to look not at the representations themselves, but at the way in which these representations are used. In the positivists' account, logical relations, especially deduction, were the model of relations between propositions. While it is certainly possible for humans and connectionist networks to use logical relations in connecting propositions (e.g., to construct a natural deduction out of propositions, Bechtel, in press a), this is not the only or even the primary way of relating propositions. In fact, most inferences people make are not logically valid. People often seem to be quite good at determining what would likely be the case when certain propositions are true even when what is inferred does not follow logically. Thus, it seems that the way in which people interact with even propositionally encoded information is not primarily through logical deduction.

An alternative view is that what people seek to do is fit information together into coherent wholes. Some of the information available is viewed as having especially high prior credibility. This information serves as a constraint on the whole pattern of information which is constructed. (For example, knowledge of a person's previous behavior and expressed moral views may make it extremely unlikely that the person would act in a certain way.) Constraints, however, cannot always be respected. Sometimes the only ways that can be envisioned for fitting pieces together requires rejecting some of the constraints.

The idea of constraint satisfaction is one that resonates with a connectionist perspective. When a connectionist network tries to solve a problem, it too seeks to find a solution that respects the various constraints presented to it (either in the form of inputs to the network or weights on connections). But the best solution it can generate sometimes violates these constraints. The constraints are thus often referred to as *soft constraints*. In what follows I will suggest that this is quite characteristic of scientific reasoning, and that it is by encouraging us to adopt this perspective on scientific reasoning that connectionism may make its major contribution to philosophy of science.

In the previous section I suggested that, at least in the life sciences, models of mechanisms rather than deductively organized sets of laws were the primary explanatory tool. Richardson and I (Bechtel & Richardson, 1993) have developed an account of the processes by which scientists construct such models. What I want to emphasize here is that when scientists engage in such model construction, they are typically confronted with a number of constraints. Some of these constraints stem from data they have about the behavior of the system to be explained, others from knowledge of possible mechanisms. Their task is one of constructing a description of a mechanism or a diagram of a mechanism that is compatible with these constraints. The problem is that often the researcher cannot find any way to do this. Any approach violates one or more of the constraints. The reasoning task turns into one of doing least damage to these constraints.

An example will make this clear. After Buchner's (1897) discovery that fermentation could be carried out in cell extracts, it became clear to investigators that fermentation did not depend upon special properties of living cells but was a chemical process. Their goal was to explain the reaction. The overall chemical change was well known. In the case of alcoholic fermentation, sugar was transformed to alcohol:

$$C_6H_{12}O_6 \rightarrow C_2H_5OH$$

Since this is not a simple or basic chemical reaction, the task was to determine the intermediate reactions and their products. The researchers were constrained by beliefs about what reactions were possible and what chemical formulae corresponded to actually occurring substances. Since it was fairly clear that fermentation involved the splitting a molecule of sugar (a compound with six carbon atoms) into 2 molecules of a compound with three carbon atoms, researchers focused their attention on a number of three-carbon compounds known from organic chemistry: lactic acid, methylglyoxal, glyceraldehyde, dihydroxyacetone, and pyruvic acid. The initial question was which of these might be intermediates in fermentation.

In evaluating whether these compounds could be inserted into their proposed pathways, researchers imposed two constraints. They required evidence (a) that the substances be found in fermenting cells and (b) that they metabolize[7] as rapidly as sugar. Both of these are reasonable, if ultimately problematic.

First, if it was not possible to find any evidence that a substance ever occurred in living cells, then it was not plausible to assume that it was indeed an intermediate in a chemical reaction in the cell. This criterion can nonetheless be problematic: in tightly linked systems such as those found in cells, the product of one reaction is rapidly employed in other reactions and so there will actually be little build up of the substance. Thus, one must often develop ingenious ways to interrupt the normal processes in the cell in order to find evidence of the intermediary. If the substance appears when normal processes are interrupted, however, it is always possible that it does so as a consequence of these perturbed conditions and is not generated under normal conditions in the cell.

[7] The researchers generally spoke of the intermediates themselves as *fermenting*. This reflects the conception, discussed below, that fermentation was simply a chain of independent reactions, not the result of a highly integrated system.

Second, if something is an intermediary, it is reasonable that the reaction from that point on take no longer than the reaction from the initial starting point. However, this is plausible only if the reactions are independent and comprise a linear chain of reactions. If later reactions are coupled to earlier ones, then an intermediate might not react as rapidly as the initial substance since the early coupled reactions will not occur. Thus, in imposing the second condition, researchers are implicitly adopting the idea that a mechanism is decomposable into a linear chain of reactions. This, however, is such a natural and powerful constraint on human thinking that it is natural to impose it as an additional constraint on possible models.

Pyruvic acid ($C_3H_4O_3$) satisfied both of the constraints on possible intermediates: it was found in living cells and reacted rapidly to yield alcohol. It was, therefore, provisionally assumed to be part of the pathway. The challenge was to fill in the rest of the pathway. One proposal stemmed from Carl Neuberg (Neuberg & Kerb, 1913). He proposed (Figure 1) that a molecule of sugar was first scissioned into two molecules of methylglyoxal (Step 1). The course of the subsequent reactions depended upon whether a supply of aldehyde was available. Before it was available, two molecules of methylglyoxal would react with two molecules of water to generate a molecule of glycerol and of pyruvic acid (Step 2a). A molecule of carbon dioxide would then be removed from the pyruvic acid, yielding aldehyde (Step 3). The molecule of aldehyde would then react with another molecule of methylglyoxal to yield pyruvic acid and alcohol (Step 2b). After a supply of aldehyde was created, Step 2a would drop out and the aldehyde generated at stage 3 in a previous cycle of the reaction would react with methylglyoxal generated in step 2b to create alcohol and pyruvic acid.

(1) $C_6H_{12}O_6 \Rightarrow 2C_3H_4O_2 + 2H_2O$
[Hexose] [Methylglyoxal + Water]

(2a) $2C_3H_4O_2 + 2H_2O \Rightarrow C_3H_8O_3 + C_3H_4O_3$
[Methylglyoxal + Water] [Glycerol + Pyruvic acid]

(2b) $C_3H_4O_2 + C_2H_4O + 2H_2O \Rightarrow C_3H_4O_3 + C_2H_5OH$
[Methylglyoxal + Aldehyde + Water] [Pyruvic acid + Alcohol]

(3) $C_3H_4O_3 \Rightarrow C_2H_4O + CO_2$
[Pyruvic acid] [Aldehyde + Carbon Dioxide]

Figure 1. Neuberg and Kerb's (1913) model of the chemical reactions involved in alcoholic fermentation.

Neuberg's account is often cited as the first coherent model of fermentation (Fruton, 1972 and Florkin, 1975). It does provide an account in terms of known intermediates and known reactions. One thing to note about this proposal is that the constraint of linearity has already been violated. After the initial stage, a reaction earlier in the pathway (2b) depends upon a reaction later in the pathway (3). The constraint of linearity was sacrificed not for principled reasons, but because a non-linear organization provided the only way to account for the overall reaction using known intermediates and known reactions.

There were two other features of Neuberg's pathway that were problematic. First, the investigations of Harden and Young in the first years of the century had indicated that in order to sustain fermentation in extracts from which all whole cells had been removed, inorganic phosphate had to be added to the extract (Harden & Young, 1906). Although the added phosphate would induce a spurt of rapid fermentation, it was soon taken up into a hexosediphosphate ester, a stable compound of sugar and phosphate, and the reaction rate slowed dramatically. The fact that adding phosphates caused a spurt of rapid fermentation suggested that phosphates played some role in the reaction. But Neuberg had provided no role for them. Neuberg's exclusion of phosphates, however, is quite understandable. Hexosediphosphate reacted very slowly in yeast extract. Thus, it failed to satisfy the criterion that any intermediate must ferment as rapidly as sugar itself. Neuberg explained away the need for adding phosphate to cell free extracts as an artifact of the experimental procedure.

The second problematic feature, however, was that methylglyoxal, when added to yeast extract, would itself not ferment. This would seem to be a telling evidence against Neuberg's proposal, especially since he has used failure to ferment as grounds for rejecting hexosediphosphate from the pathway. But Neuberg did not so regard it. He assumed that the failure of methylglyoxal to ferment was due to the experimental arrangement, and he and others continued to seek evidence that, if supplied in the correct way, it would metabolize in normal yeast cells. He also considered the possibility that the form of methylglyoxal occurring in normal fermentation was different from laboratory methylglyoxal. (Another factor Neuberg cited as supporting the role of methylglyoxal in the pathway was that it could account for the methyl that was found in the yeast extract.)

About twenty years later Neuberg's model of fermentation was supplanted by one in which phosphorylated compounds figured throughout the pathway and methylglyoxal was removed from the pathway (for details of this history, see Bechtel & Richardson, 1993). The reason hexosediphosphate seemed to build up and not ferment in cell-free extracts also came to be understood: the overall reaction requires a supply of ADP to which to transfer the high-energy phosphate bonds that are developed through the fermentation reaction (forming ATP). In normal cells, this ADP is made available by the breakdown of ATP in the course of cell work, but this process was not available in the extracts. The cell is not the nearly decomposable system that early researchers assumed, but a highly integrated system. Nonetheless, for about twenty years Neuberg's model was regarded as the most plausible model of fermentation.

What is of interest here is the fact that Neuberg's model was pieced together in an attempt to satisfy a number of constraints. It succeeded in satisfying many of these constraints. It used known intermediates and known reactions to construct a coherent pathway. This success in satisfying constraints largely explains its acceptance. But the constraints it failed to satisfy are also noteworthy: it could not explain the need for phosphates and methylglyoxal failed to satisfy the criterion of metabolizing as rapidly as sugar. Thus, it satisfied some constraints, but not others. This suggests that in science as in ordinary life, many constraints are soft constraints and these can be violated if the overall result is the best that can be obtained. Insofar as this is a mode of reasoning connectionism leads us to expect, connectionism may have an important role in helping us understand the reasoning processes scientists employ in developing such models. (Moreover, it suggests that connectionism may eventually provide a useful framework for modeling such reasoning.)

4. Evaluating Research Techniques and Data as Soft Constraint Satisfaction

One of the virtues Churchland cites for his connectionist account of theories is that it can account for Kuhn's (1970) contention that observation is theory-laden. The theory-ladenness claim opposed the view that scientific knowledge was built on foundations such as observation reports that were immune to challenge. But the traditional account of theory-ladenness only focuses on the role of theories in fixing our characterization of what we observe. In many scientific disciplines, what is observed is not just the product of our unaided senses and our theories, but rather depends upon a variety of instruments and research techniques. Outside of Hacking (1983), the development of these instruments and techniques for using them has not received much discussion in philosophy. (Sociologists of science such as Latour, 1987, on the other hand, have made much of the role of the development of instruments and research techniques, seeing in them further support for the view the science is a social construction not guided by epistemic considerations.) But explaining the reasoning involved in development and evaluation of instruments and research techniques should be a major objective for philosophy of science since these techniques play a pivotal role in determining what is taken to be the evidence for scientific models or theories.

Although philosophers of science have said little about how scientists reason about instruments and research techniques, discussion of instruments and research techniques is a major part of the ongoing discussions of many sciences. Some of the greatest controversies in the scientific literature are not about theories, but rather about the instruments and techniques which give rise to data. A constant challenge is that what is taken to be data by one scientist is not informative about the natural system under study but is a product only of the instruments or techniques used to study those systems. Reasoning about instruments and research techniques is therefore a central activity of scientists.

If we cannot give an account of how scientists reason about and decide upon instruments and research techniques, then we can give no account of why they should put such faith in their evidence. Insofar as we maintain our focus on logical relations between linguistic propositions, moreover, we may not be able account for reasoning about instruments since much of the development of instruments and research techniques is not grounded on theories or propositions, but on physical explorations with the instruments. Instruments and techniques are not justified because they are built on already justified principles. Churchland briefly suggests that a connectionist framework can help us understand the development of skills as an important part of learning to be a scientist since, like all knowledge, knowledge of skills for connectionists consists in weights in a network. But, as in the last section, I want to urge that if we are to understand the development of knowledge scientists have of instruments and research techniques, we cannot, as Churchland suggests, focus just on what occurs inside the head. Rather, there is a crucial interaction between scientist and physical parts of the world, including both the physical instrument and the physical actions of scientists and technicians. Further, as with models themselves, as scientists develop instruments and techniques they are frequently attempting to satisfy multiple soft constraints.

The modern discipline of cell biology emerged with the development of new instruments and research techniques which made it possible to identify structures within the cell and determine their contributions to cell life. Very important among these was cell fractionation, which provided a tool for isolating cell organelles to determine their biochemical function. The potential for artifact presented by this technique is obvious: it involves subjecting cellular materials to forces several thousand times that of gravity in order to effect the separation. But one hopes that the cellular components themselves will not be adversely affected by the process. In the early development of cell fractionation techniques various approaches to fractionation were developed, many of which succeeded in some part of the task but then failed in others, requiring constant compromise. This is clearly brought about by Allfrey (1959):

> The ideal isolation procedure is easy to define: it is the method which yields the desired intracellular components as they exist in the cell, unchanged, uncontaminated, and in quantitative yield. Unfortunately, all cell fractionation techniques known at present fall short of this ideal, and some compromise becomes necessary. Methods which give quantitative yields often involve serious alterations of form, structure, or composition, and procedures which preserve the morphology often destroy activity and function. Purity and homogeneity of the product are rarely, if ever, achieved. (p. 198)

I will focus on three aspects of the overall fractionation process, in each case showing how researchers were trying to satisfy competing constraints.

A first challenge in cell fractionation is to break the cell membrane. This requires subjecting the cell to disruptive forces. At the same time, however, one seeks to do the least damage to the internal components. For example, one wants to prevent them from releasing enzymes contained within them into the medium in which the fractionation is occurring. A host of instruments were employed by researchers in breaking the cell: the Waring Blendor, mortar and pestle, various piston-type homogenizers, colloid mills, and sonic vibrators. In fact during the 1940s and 50s scientists were actively exploring various techniques. Allfrey (1959) presents a diagram of nine different designs for homogenizers alone, each of which produced the shearing force required to break the cell membrane in somewhat different ways, and hence contributed to different overall results. In each case, researchers faced competing goals: to insure that all cell membranes were broken, but not to disrupt internal membranes. If one failed to break all the cells, then their materials would be deposited in the first fraction (together with nuclei), distorting the analysis of that fraction and the attempt to secure quantitative information about what substances appeared at each location in the cell. On the other hand, too violent a technique would break up the internal organelles as well.

Typically techniques for breaking cells were evaluated in terms of their products: did the technique break all the cells without harming the organelles (e.g., altering their biochemistry). Whether all the cells were broken could be evaluated by microscopic evaluation, but evaluating whether organelles were harmed was far more problematic. The process of fractionation was the primary tool for determining the nature of these organelles, so there was no independent standard by which to judge harm. Here researchers relied on two strategies. The first was to compare the results of one technique for breaking the cell with others that were already held in some repute. But if the technique was to be judged an improvement, it should not just produce results that can already be obtained otherwise. The second criterion therefore was crucial: did the fractionation process as a whole yielded results concurrent with an emerging theory? While this appeal to theories, which are supposed to be grounded on the evidence provided by the technique, to evaluate the technique may seem circular to a foundationalist, it is in fact the sort of reasoning that often figures in scientists' evaluations of their techniques, as we shall see by turning to the other components of the cell fractionation process.

As the cell membrane is broken, its contents must be released into a fluid medium. The challenge was to find a medium that would cause least disruption of the cell organelles. Among the variables considered were the substances to include (salt, sucrose, citric acid, etc.), the concentration of each of these (hypotonic, isotonic, or hypertonic), and the pH for the medium. Each of these would affect the organelles differently. Claude's early fractionation work used a solution consisting of a few drops of NaOH at a pH of about 9.0 (Claude, 1943); he later used a saline solution buffered at a pH of 9.0 to 9.5. This led to serious distortion of the shape of organelles such as mitochondria, but did provided a basis for identifying the mitochondrion as the locus of some of the crucial enzymes in cellular respiration (Hogeboom, Claude, & Hotchkiss, 1946). Subsequently, Hogeboom, Schneider, and Palade (1948) employed hypertonic sucrose solution, which resulted in mitochondria retaining their normal rod-like shape and staining, but failing to synthesize ATP. Schneider (1948) then introduced isotonic sucrose, which preserves the ATP synthesis, but compromises shape. We should note that once again what was critical to judging whether a new medium represented an improvement was whether the results cohered with those using other investigatory strategies (microscopy for determining the shape of particles, staining for identifying particles) and with developing theories (e.g., that the mitochondrion was the "powerplant" of the cell). Moreover, a perfect solution was not possible.

After the cell membranes have been broken and the contents released into an aqueous medium, the preparation is ready for centrifugation. Here the underlying principle was clear: depending upon their density and shape, particles will travel at different speeds in a centrifuge. Drawing on Stokes Law, it was possible to specify the rate at which different particles should move, but a variety of factors generated results different from the ideal. For example, particles would hit the side of the vesicle, and often slide down it at a very different rate. Also, particles might agglutinate, resulting in different size and shaped particles traveling together. Finally, some of the lighter materials will start closer to the bottom than some of the heavier materials, and will sediment out in the same time as it takes for the heavier materials (de Duve and Berthet, 1954). The result is that if one prolongs centrifugation to secure all the heavier particles, one increases the amount of contamination in that layer. If one shortens the time, then the heavier particles will contaminate other fractions.

In his early fractionation work, Claude (1940) discovered a technique (three or four alternate long and short runs of a high-speed centrifuge under 18,000 g) for producing a fraction of small particles, which he compared with particles he had isolated from viruses. Finding them in all cells, normal and pathological, he developed the idea that they might be mitochondria, or pieces of mitochondria. He contrasted them with somewhat larger particles previously separated and identified as mitochondria by Bensley and Hoerr (1934) and argued that Bensley's particles were really secretory. Claude subsequently reversed his judgment, deciding that the larger particles, which were sedimented faster, were the mitochondria, and identified the new particles as small particles or microsomes. As the techniques were refined (establishment of different speeds and different times for different fractions, and use of washing and resedimenting to increase purity), it became clear that the two fractions have very different chemical make-ups. Two other fractions were also distinguished: a nuclear fraction was sedimented even more quickly or at lesser speed, and a supernate consisting of the fluid materials left after the last sedimentation. The separation of these four fractions became the standard approach for many years.

What justifies these four fractions? One factor is that the particles were of significantly different sizes, Further, they were clearly different in chemical make-up. Claude (1948, p. 127-8) comments: "It should be pointed out that division of the cell in this manner is not arbitrary and is not based on size differences alone since . . . these various fractions are also distinct in chemical constitution, in biochemical functions, and even in color." While many enzymes and other compounds were common to all fractions, some were predominately recovered in a single fraction: DNA and DPN-synthesizing enzyme in the nuclear fraction, succinic dehydrogenase, cytochrome oxidase and cytochrome *c* in the mitochondrial fraction, and glucose-6-phosphatase in the microsomal fraction. Some researchers developed the strategy of trying to demonstrate that different enzymes each originated in a different fraction of the cell, and used the ability to produce fractions of this sort as a criterion of the correctness of their approach (Claude, 1948, de Duve & Berthet, 1954). Accepting the principle as a constraint provided the basis for differentiating yet more fractions and arguing that they originated in different loci in the normal cell, an approach that turned out to be very fruitful. Other researchers (e.g., Dounce, 1954) strongly resisted this move, arguing that it was highly plausible that the same enzyme or chemical compound might function in different parts of the cell.

The first thing to note is that cell fractionation is a complex process. Small variations in the technique could yield very different results. de Duve and Berthet make this point clearly: "Differential centrifugation is a delicate method, and small modifications in the procedures applied may in many cases alter quite significantly the manner in which a preparation is finally fractionated". (1954, p. 226) No approach achieved the ideal described in the passage by Allfrey above. And yet some of the techniques were taken to provide authoritative information about the natural state of cells. How did the scientists determine which results were authoritative? As in the building of models described in the previous section, alternatives were developed and evaluated against multiple constraints such as: How did the results of this technique correspond to those achieved with other approaches? How well did the results fit with an emerging understanding of cell function? Since no technique satisfied all criteria, some had to be sacrificed. Thus, once again scientists seem to be engaged in soft constraint satisfaction, the sort of processing characteristic of connectionist networks.

In developing their instruments and techniques, we should note that scientists are typically not engaged in manipulating propositions. They use propositions and diagrams in communicating their techniques to others, and in presenting reasons why a technique should or should not be respected, but generally not in the process of developing them. Rather, they interact directly with the physical objects. While I have not mentioned it, the physical instruments and the scientist's body themselves provide additional constraints. Only some modifications of the material of the instrument or ways of maneuvering a person's body are possible. In this case it does not make sense to think of the reasoning as carried out in language. Linguistically encoded information provides one source of constraint, but directly apprehended physical factors provide others. Amongst these constraints, though, the chief cognitive activity is one of satisfying as many of the constraints as possible.

5. Conclusion

Churchland argues that connectionism can play an important role in helping us reconceptualize philosophy of science. I agree with this contention, but claim that the role for connectionism is somewhat different than Churchland presents. For Churchland, the contribution is to move us away from a sentential construal of theories and of explanation. In their stead, he proposes that theories consists of weights within a network and that explanatory understanding involves the activation of prototype vectors in the network. I have argued that sentential representations do have a role to play, but that their primary locus is not in the heads of scientists. Scientists use representations in natural language as well as figures and diagrams to encode their models. These representations are not translations of what is in the heads of the scientists, rather they are devices used by scientists. Scientific theories may take a sentential form even if, in using these theories, scientists rely on weights on connections within their heads. Consequently, we should not seek to localize the story of scientific development in representations and processes occurring in the head. Rather, we need to take seriously the fact that scientists are situated cognizers whose cognitive processes involve interactions with external representations as well as physical devices.

While I do not foresee connectionism supplanting the sentential framework, when that framework is restricted to external representations, I have suggested it can fundamentally alter our conception of what scientists do in their interactions with theories or with their research instruments. Connectionist networks treat connections and inputs as constraints; in solving a problem they seek a state that maximally satisfies these constraints. Some of the constraints must be overridden in the process. Accordingly, the constraints in a network are *soft*. I have argued that this is the kind of reasoning scientists often engage in when developing explanatory models. They seek a representation that maximally satisfies the various empirical and conceptual constraints on the model. I have also argued that in their reasoning about new instruments and research techniques scientists likewise seek to satisfy multiple soft constraints. Thus, connectionism can make an important contribution to philosophy of science as it moves us away from deductive and inductive logic as the model of scientific reasoning to a model of soft constraint satisfaction performed in the context of interacting with external representations and physical devices.

References

Allfrey, V. (1959). The isolation of subcellular components. In. J. Brachet and A. E. Mirsky (eds.), *The cell: Biochemistry, physiology, and morphology*, pp. 193-290. New York: Academic.

Bechtel, W. (in press a). Natural deduction in connectionist systems. *Synthese.*

Bechtel, W. (in press b). What knowledge must be in the head that we might acquire language? In B. Velichkovsky and D. M. Rumbaugh (ed.), *Naturally human: Origins and destiny of language*.

Bechtel, W. and Abrahamsen, A. (1991). *Connectionism and the mind: An introduction to parallel processing in networks*. Oxford: Basil Blackwell.

Bechtel, W. and Richardson, R. C. (1993). *Discovering complexity: Decomposition and localization as scientific research strategies*. Princeton: Princeton University Press.

Bensley, R. R. and Hoerr, N. (1934). Studies on cell structure by the freeze-drying method. VI. The preparation and properties of mitochondria. *Anatomical Record*, *60*, 449-55.

Buchner, E. (1897). Alkoholische Gährung ohne Hefezellen Vorläufige Mittheilung. *Berichte der* deutschen chemischen Gesellschaft, *37*, 417-28.

Cartwright, N. (1989). *How the laws of physics lie*. Oxford: Clarendon Press.

Churchland, P. M. (1989). *A neurocomputational perspective: The nature of mind and the structure of science*. Cambridge, MA: MIT Press.

Claude, A. (1940). Particulate components of normal and tumor cells. *Science*, *91*, 77-8.

Claude, A. (1943). Distribution of nucleic acids in the cell and the morphological constitution of cytoplasm." In J. Cattell (ed.) *Biological symposium. Volume X. Frontiers of cytochemistry*, pp. 111-29. Lancaster, PA: Jacques Cattell Press.

Claude, A. (1948). Studies on cells. Morphology, chemical constitution, and distribution of biochemical fractions. *Harvey Lectures*, *43*, 121-64.

de Duve, C. and Berthet, J. (1954). The use of differential centrifugation in the study of tissue enzymes. *International Review of Cytology*, *3*, 225-275.

Dounce, A. L. (1954). The significance of enzyme studies on isolated cell nuclei. *International Review of Cytology*, *3*, 199-223.

Florkin, M. (1975). *Comprehensive biochemistry.* Volume 31, *A history of biochemistry.* Part III. *History of the sources of free energy in organisms*. Amsterdam: Elsevier.

Fruton, J. (1972). *Molecules and life: Historical essays on the interplay of chemistry and biology*. New York: Wiley Interscience.

Giere, R. (1988). *Explaining science*. Chicago: University of Chicago Press.

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science.* Cambridge: Cambridge University Press.

Harden, A. and Young, W. J. (1906). The alcoholic fermentation of yeast-juice. *Proceedings of the Royal Society, London, B77*, 405-420.

Hogeboom, G. H., Claude, A. and Hotchkiss, R. (1946). The distribution of cytochrome oxidase and succinoxidase in the cytoplasm of the mammalian liver cell. *Journal of Biological Chemistry*, *165*, 615-629.

Hogeboom, G. H., Schneider, W. C., and Palade, G. E. (1948). Cytochemical studies of mammalian tissues. I. Isolation of intact mitochondria from rat liver: Some biochemical properties of mitochondria and submicroscopic particulate material. *Journal of Biological Chemistry,* Kauffman, S. A. (1993): *The origins of order: Self-organization and selection in evolution.* Oxford: Oxford University Press.

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

Latour, B. (1987). *Scientists in action: How to follow scientists and engineers through society.* Cambridge, MA: Harvard University Press.

Neuberg, C. and Kerb, J. (1914). Über zuckerfreie Hefegärungen. XII. Über Vorgänge bei der Hefegärung. *Biochemische Zeitschrift*, *53*, 406-19.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., and Hinton, G.E. (1986). Schemas and sequential thought processes in PDP models. In J. L. McClelland, D. E. Rumelhart, and the PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 2: Psychological and biological models*, pp. 7-57. Cambridge, MA: MIT Press.

Schneider, W. C. (1948). The intracellular distribution of enzymes. III. The oxidation of octanoic acid by rat liver fractions. *Journal of Biological Chemistry*, 176, 259-266.

St. John, M. F. & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, *46*, 217-57.

Wimsatt, W. C. (1987). False models as means to truer theories. In M. Nitecki and A. Hoffman (eds.) *Neutral models in biology*, pp. 23-55. London: Oxford University Press.

Vygotsky, L. S. (1962). *Language and thought*. Cambridge, MA: MIT Press.