Philosophical Perspectives, 4
Action Theory and Philosophy of Mind, 1990

CONNECTIONISM, ELIMINATIVISM AND THE FUTURE OF FOLK PSYCHOLOGY[1]

William Ramsey
University of Notre Dame
Stephen Stich
Rutgers University
Joseph Garon
La Jolla, CA

1. Introduction

In the years since the publication of Thomas Kuhn's *Structure of Scientific Revolutions*, the term "scientific revolution" has been used with increasing frequency in discussions of scientific change, and the magnitude required of an innovation before someone or other is tempted to call it a revolution has diminished alarmingly. Our thesis in this paper is that if a certain family of connectionist hypotheses turn out to be right, they will surely count as revolutionary, even on stringent pre-Kuhnian standards. There is no question that connectionism has already brought about major changes in the way many cognitive scientists conceive of cognition. However, as we see it, what makes certain kinds of connectionist models genuinely revolutionary is the support they lend to a thoroughgoing eliminativism about some of the central posits of common sense (or "folk") psychology. Our focus in this paper will be on beliefs or propositional memories, though the argument generalizes straightforwardly to all the other propositional attitudes. If we are right, the consequences of this kind of connectionism extend well beyond the confines of cognitive science, since these models, if successful, will require a radical reorientation in the way we think about ourselves.

Here is a quick preview of what is to come. Section 2 gives a brief account of what eliminativism claims, and sketches a pair of premises that eliminativist arguments typically require. Section 3 says a bit

about how we conceive of common sense psychology, and the propositional attitudes that it posits. It also illustrates one sort of psychological model that exploits and builds upon the posits of folk psychology. Section 4 is devoted to connectionism. Models that have been called "connectionist" form a fuzzy and heterogeneous set whose members often share little more than a vague family resemblance. However, our argument linking connectionism to eliminativism will work only for a restricted domain of connectionist models, interpreted in a particular way; the main job of Section 4 is to say what that domain is and how the models in the domain are to be interpreted. In Section 5 we will illustrate what a connectionist model of belief that comports with our strictures might look like, and go on to argue that if models of this sort are correct, then things look bad for common sense psychology. Section 6 assembles some objections and replies. The final section is a brief conclusion.

Before plunging in we should emphasize that the thesis we propose to defend is a *conditional* claim: *If* connectionist hypotheses of the sort we will sketch turn out to be right, so too will eliminativism about propositional attitudes. Since our goal is only to show how connectionism and eliminativism are related, we will make no effort to argue for the truth or falsity of either doctrine. In particular, we will offer no argument in favor of the version of connectionism required in the antecedent of our conditional. Indeed our view is that it is early days yet—too early to tell with any assurance how well this family of connectionist hypotheses will fare. Those who are more confident of connectionism may, of course, invoke our conditional as part of a larger argument for doing away with the propositional attitudes.[2] But, as John Haugeland once remarked, one man's ponens is another man's tollens. And those who take eliminativism about propositional attitudes to be preposterous or unthinkable may well view our arguments as part of a larger case against connectionism. Thus, we'd not be at all surprised if trenchant critics of connectionism, like Fodor and Pylyshyn, found both our conditional and the argument for it to be quite congenial.[3]

2. Eliminativism and Folk Psychology

'Eliminativism', as we shall use the term, is a fancy name for a simple thesis. It is the claim that some category of entities, processes

or properties exploited in a common sense or scientific account of the world do not exist. So construed, we are all eliminativists about many sorts of things. In the domain of folk theory, witches are the standard example. Once upon a time witches were widely believed to be responsible for various local calamities. But people gradually became convinced that there are better explanations for most of the events in which witches had been implicated. There being no explanatory work for witches to do, sensible people concluded that there were no such things. In the scientific domain, phlogiston, caloric fluid and the luminiferous ether are the parade cases for eliminativism. Each was invoked by serious scientists pursuing sophisticated research programs. But in each case the program ran aground in a major way, and the theories in which the entities were invoked were replaced by successor theories in which the entities played no role. The scientific community gradually came to recognize that phlogiston and the rest do not exist.

As these examples suggest, a central step in an eliminativist argument will typically be the demonstration that the theory in which certain putative entities or processes are invoked should be rejected and replaced by a better theory. And that raises the question of how we go about showing that one theory is better than another. Notoriously, this question is easier to ask than to answer. However, it would be pretty widely agreed that if a new theory provides more accurate predictions and better explanations than an old one, and does so over a broader range of phenomena, and if the new theory comports as well or better with well established theories in neighboring domains, then there is good reason to think that the old theory is inferior, and that the new one is to be preferred. This is hardly a complete account of the conditions under which one theory is to be preferred to another, though for our purposes it will suffice.

But merely showing that a theory in which a class of entities plays a role is inferior to a successor theory plainly is not sufficient to show that the entities do not exist. Often a more appropriate conclusion is that the rejected theory was wrong, perhaps seriously wrong, about some of the properties of the entities in its domain, or about the laws governing those entities, and that the new theory gives us a more accurate account *of those very same entities*. Thus, for example, pre-Copernican astronomy was very wrong about the nature of the planets and the laws governing their movement. But it would be something of a joke to suggest that Copernicus and Galileo showed

that the planets Ptolemy spoke of do not exist.[4]

In other cases the right thing to conclude is that the posits of the old theory are reducible to those of the new. Standard examples here include the reduction of temperature to mean molecular kinetic energy, the reduction of sound to wave motion in the medium, and the reduction of genes to sequences of polynucleotide bases.[5] Given our current concerns, the lesson to be learned from these cases is that even if the common sense theory in which propositional attitudes find their home is replaced by a better theory, that would not be enough to show that the posits of the common sense theory do not exist.

What more would be needed? What is it that distinguishes cases like phlogiston and caloric, on the one hand, from cases like genes or the planets on the other? Or, to ask the question in a rather different way, what made phlogiston and caloric candidates for elimination? Why wasn't it concluded that phlogiston is oxygen, that caloric is kinetic energy, and that the earlier theories had just been rather badly mistaken about some of the properties of phlogiston and caloric?

Let us introduce a bit of terminology. We will call theory changes in which the entities and processes of the old theory are retained or reduced to those of the new one *ontologically conservative* theory changes. Theory changes that are not ontologically conservative we will call *ontologically radical*. Given this terminology, the question we are asking is how to distinguish ontologically conservative theory changes from ontologically radical ones.

Once again, this is a question that is easier to ask than to answer. There is, in the philosophy of science literature, nothing that even comes close to a plausible and fully general account of when theory change sustains an eliminativist conclusion and when it does not. In the absence of a principled way of deciding when ontological elimination is in order, the best we can do is to look at the posits of the old theory—the ones that are at risk of elimination—and ask whether there is anything in the new theory that they might be identified with or reduced to. If the posits of the new theory strike us as deeply and fundamentally different from those of the old theory, in the way that molecular motion seems deeply and fundamentally different from the "exquisitely elastic" fluid posited by caloric theory, then it will be plausible to conclude that the theory change has been a radical one, and that an eliminativist conclusion is in order. But

since there is no easy measure of how "deeply and fundamentally different" a pair of posits are, the conclusion we reach is bound to be a judgment call.[6]

To argue that certain sorts of connectionist models support eliminativism about the propositional attitudes, we must make it plausible that these models are not ontologically conservative. Our strategy will be to contrast these connectionist models, models like those set out in Section 5, with ontologically conservative models like the one sketched at the end of Section 3, in an effort to underscore just how ontologically radical the connectionist models are. But here we are getting ahead of ourselves. Before trying to persuade you that connectionist models are ontologically radical, we need to take a look at the folk psychological theory that the connectionist models threaten to replace.

3. Propositional Attitudes and Common Sense Psychology

For present purposes we will assume that common sense psychology can plausibly be regarded as a theory, and that beliefs, desires and the rest of the propositional attitudes are plausibly viewed as posits of that theory. Though this is not an uncontroversial assumption, the case for it has been well argued by others.[7] Once it is granted that common sense psychology is indeed a theory, we expect it will be conceded by almost everyone that the theory is a likely candidate for replacement. In saying this, we do not intend to disparage folk psychology, or to beg any questions about the status of the entities it posits. Our point is simply that folk wisdom on matters psychological is not likely to tell us all there is to know. Common sense psychology, like other folk theories, is bound to be incomplete in many ways, and very likely to be inaccurate in more than a few. If this were not the case, there would be no need for a careful, quantitative, experimental science of psychology. With the possible exception of a few die hard Wittgensteinians, just about everyone is prepared to grant that there are many psychological facts and principles beyond those embedded in common sense. If this is right, then we have the first premise needed in an eliminativist argument aimed at beliefs, propositional memories and the rest of the propositional attitudes. The theory that posits the attitudes is indeed a prime candidate for replacement.

Though common sense psychology contains a wealth of lore about beliefs, memories, desires, hopes, fears and the other propositional attitudes, the crucial folk psychological tenets in forging the link between connectionism and eliminativism are the claims that propositional attitudes are *functionally discrete*, *semantically interpretable*, states that play a *causal role* in the production of other propositional attitudes, and ultimately in the production of behavior. Following the suggestion in Stich (1983), we'll call this cluster of claims *propositional modularity*.[8] (The reader is cautioned not to confuse this notion of propositional modularity with the very different notion of modularity defended in Fodor (1983).)

There is a great deal of evidence that might be cited in support of the thesis that folk psychology is committed to the tenets of propositional modularity. The fact that common sense psychology takes beliefs and other propositional attitudes to have semantic properties deserves special emphasis. According to common sense:

> i) when people see a dog nearby they typically come to believe *that there is a dog nearby*;
> ii) when people believe *that the train will be late if there is snow in the mountains*, and come to believe *that there is snow in the mountains*, they will typically come to believe *that the train will be late*;
> iii) when people who speak English say 'There is a cat in the yard,' they typically believe *that there is a cat in the yard.*

And so on, for indefinitely many further examples. Note that these generalizations of common sense psychology are couched in terms of the *semantic* properties of the attitudes. It is in virtue of being the belief *that p* that a given belief has a given effect or cause. Thus common sense psychology treats the predicates expressing these semantic properties, predicates like 'believes *that the train is late*', as *projectable* predicates—the sort of predicates that are appropriately used in nomological or law-like generalizations.

Perhaps the most obvious way to bring out folk psychology's commitment to the thesis that propositional attitudes are *functionally discrete* states is to note that it typically makes perfectly good sense to claim that a person has acquired (or lost) a single memory or belief. Thus, for example, on a given occasion it might plausibly be claimed that when Henry awoke from his nap he had completely forgotten that the car keys were hidden in the refrigerator, though he had

forgotten nothing else. In saying that folk psychology views beliefs as the sorts of things that can be acquired or lost one at a time, we do not mean to be denying that having any particular belief may presuppose a substantial network of related beliefs. The belief that the car keys are in the refrigerator is not one that could be acquired by a primitive tribesman who knew nothing about cars, keys or refrigerators. But once the relevant background is in place, as we may suppose it is for us and for Henry, it seems that folk psychology is entirely comfortable with the possibility that a person may acquire (or lose) the belief that the car keys are in the refrigerator, while the remainder of his beliefs remain unchanged. Propositional modularity does not, of course, deny that acquiring one belief often leads to the acquisition of a cluster of related beliefs. When Henry is told that the keys are in the refrigerator, he may come to believe that they haven't been left in the ignition, or in his jacket pocket. But then again he may not. Indeed, on the folk psychological conception of belief it is perfectly possible for a person to have a long standing belief that the keys are in the refrigerator, and to continue searching for them in the bedroom.[9]

To illustrate the way in which folk psychology takes propositional attitudes to be functionally discrete, *causally active* states let us sketch a pair of more elaborate examples.

i) In common sense psychology, behavior is often explained by appeal to certain of the agent's beliefs and desires. Thus, to explain why Alice went to her office, we might note that she wanted to send some e-mail messages (and, of course, she believed she could do so from her office). However, in some cases an agent will have several sets of beliefs and desires each of which *might* lead to the same behavior. Thus we may suppose that Alice also wanted to talk to her research assistant, and that she believed he would be at the office. In such cases, common sense psychology assumes that Alice's going to her office might have been caused by either one of the belief/desire pairs, or by both, and that determining which of these options obtains is an empirical matter. So it is entirely possible that on *this* occasion Alice's desire to send some e-mail played no role in producing her behavior; it was the desire to talk with her research assistant that actually caused her to go to the office. However, had she not wanted to talk with her research assistant, she might have gone to the office anyhow, because the desire to send some e-mail, which was causally inert in her actual decision making, might then have become actively

involved. Note that in this case common sense psychology is prepared to recognize a pair of quite distinct semantically characterized states, one of which may be causally active while the other is not.

ii) Our second illustration is parallel to the first, but focuses on beliefs and inference, rather than desires and action. On the common sense view, it may sometimes happen that a person has a number of belief clusters, any one of which might lead him to infer some further belief. When he actually does draw the inference, folk psychology assumes that it is an empirical question what he inferred it from, and that this question typically has a determinate answer. Suppose, for example, that Inspector Clouseau believes that the butler said he spent the evening at the village hotel, and that he said he arrived back on the morning train. Suppose Clouseau also believes that the village hotel is closed for the season, and that the morning train has been taken out of service. Given these beliefs, along with some widely shared background beliefs, Clouseau might well infer that the butler is lying. If he does, folk psychology presumes that the inference might be based either on his beliefs about the hotel, or on his beliefs about the train, or both. It is entirely possible, from the perspective of common sense psychology, that although Clouseau has long known that the hotel is closed for the season, this belief played no role in his inference on this particular occasion. Once again we see common sense psychology invoking a pair of distinct propositional attitudes, one of which is causally active on a particular occasion while the other is causally inert.

In the psychological literature there is no shortage of models for human belief or memory which follow the lead of common sense psychology in supposing that propositional modularity is true. Indeed, prior to the emergence of connectionism, just about all psychological models of propositional memory, save for those urged by behaviorists, were comfortably compatible with propositional modularity. Typically, these models view a subject's store of beliefs or memories as an interconnected collection of functionally discrete, semantically interpretable states which interact in systematic ways. Some of these models represent individual beliefs as sentence-like structures—strings of symbols which can be individually activated by transferring them from long term memory to the more limited memory of a central processing unit. Other models represent beliefs as a network of labeled nodes and labeled links through which patterns of activation may spread. Still other models represent beliefs

as sets of production rules.[10] In all three sorts of models, it is generally the case that for any given cognitive episode, like performing a particular inference or answering a question, some of the memory states will be actively involved, and others will be dormant.

In Figure 1 we have displayed a fragment of a "semantic network" representation of memory, in the style of Collins & Quillian (1972). In this model, each distinct proposition in memory is represented by an oval node along with its labeled links to various concepts. By adding assumptions about the way in which questions or other sorts of memory probes lead to activation spreading through the network, the model enables us to make predictions about speed and accuracy in various experimental studies of memory. For our purposes there are three facts about this model that are of particular importance. First, since each proposition is encoded in a functionally discrete way, it is a straightforward matter to add or subtract a *single* proposition from memory, while leaving the rest of the network unchanged. Thus, for example, Figure 2 depicts the result of removing one proposition from the network in Figure 1. Second, the model treats predicates expressing the semantic properties of beliefs or memories as *projectable*.[11] They are treated as the sorts of predicates that pick out scientifically genuine *kinds*, rather than mere accidental conglomerates, and thus are suitable for inclusion in the statement of lawlike regularities. To see this, we need only consider the way in which such models are tested against empirical data about memory acquisition and forgetting. Typically, it will be assumed that if a subject is told (for example) that the policeman arrested the hippie, then the subject will (with a certain probability) remember *that the policeman arrested the hippie*.[12] And this assumption is taken to express a nomological generalization—it captures something lawlike about the way in which the cognitive system works. So while the class of people who *remember that the policeman arrested the hippie* may differ psychologically in all sorts of ways, the theory treats them as a psychologically natural kind. Third, in any given memory search or inference task exploiting a semantic network model, it makes sense to ask which propositions were activated and which were not. Thus, a search in the network of Figure 1 might terminate without ever activating the proposition that cats have paws.

4. A Family of Connectionist Hypotheses

Our theme, in the previous section, was that common sense psychology is committed to propositional modularity, and that many models of memory proposed in the cognitive psychology literature are comfortably compatible with this assumption. In the present section we want to describe a class of connectionist models which, we will argue, are *not* readily compatible with propositional modularity. The connectionist models we have in mind share three properties:

> i) their encoding of information in the connection weights and in the biases on units is *widely distributed*, rather than being *localist*;
> ii) individual hidden units in the network have no comfortable symbolic interpretation; they are *subsymbolic*, to use a term suggested by Paul Smolensky;
> iii) the models are intended *as cognitive models*, not merely as *implementations* of cognitive models.

A bit later in this section we will elaborate further on each of these three features, and in the next section we will describe a simple example of a connectionist model that meets our three criteria. However, we are under no illusion that what we say will be sufficient to give a sharp-edged characterization of the class of connectionist models we have in mind. Nor is such a sharp-edged characterization essential for our argument. It will suffice if we can convince you that there is a significant class of connectionist models which are incompatible with the propositional modularity of folk psychology.

Before saying more about the three features on our list, we would do well to give a more general characterization of the sort of models we are calling "connectionist," and introduce some of the jargon that comes with the territory. To this end, let us quote at some length from Paul Smolensky's lucid overview.

> Connectionist models are large networks of simple, parallel computing elements, each of which carries a numerical *activation value* which it computes from neighboring elements in the network, using some simple numerical formula. The network elements or *units* influence each other's values through connections that carry a numerical strength or *weight*...

> In a typical...model, input to the system is provided by imposing activation values on the *input units* of the network; these numerical values represent some encoding or *representation* of the input. The activation on the input units propagates along the connections until some set of activation values emerges on the *output units*; these activation values encode the output the system has computed from the input. In between the input and output units there may be other units, often called *hidden units*, that participate in representing neither the input nor the output.
>
> The computation performed by the network in transforming the input pattern of activity to the output pattern depends on the set of connection strengths; *these weights are usually regarded as encoding the system's knowledge*.[13] In this sense, the connection strengths play the role of the program in a conventional computer. Much of the allure of the connectionist approach is that many connectionist networks *program themselves*, that is, they have autonomous procedures for tuning their weights to eventually perform some specific computation. Such *learning procedures* often depend on training in which the network is presented with sample input/output pairs from the function it is supposed to compute. In learning networks with hidden units, the network itself "decides" what computations the hidden units will perform; because these units represent neither inputs nor outputs, they are never "told" what their values should be, even during training ....[14]

One point must be added to Smolensky's portrait. In many connectionist models the hidden units and the output units are assigned a numerical "bias" which is added into the calculation determining the unit's activation level. The learning procedures for such networks typically set both the connection strengths and the biases. Thus in these networks the system's knowledge is usually regarded as encoded in *both* the connection strengths and the biases.

So much for a general overview. Let us now try to explain the three features that characterize those connectionist models we take to be incompatible with propositional modularity.

(i) In many non-connectionist cognitive models, like the one illustrated at the end of Section 3, it is an easy matter to locate a

functionally distinct part of the model encoding each proposition or state of affairs represented in the system. Indeed, according to Fodor and Pylyshyn, "conventional [computational] architecture requires that there be distinct symbolic expressions for each state of affairs that it can represent."[15] In some connectionist models an analogous sort of functional localization is possible, not only for the input and output units but for the hidden units as well. Thus, for example, in certain connectionist models, various individual units or small clusters of units are themselves intended to represent specific properties or features of the environment. When the connection strength from one such unit to another is strongly positive, this might be construed as the system's representation of the proposition that if the first feature is present, so too is the second. However, in many connectionist networks it is not possible to localize propositional representation beyond the input layer. That is, there are no particular features or states of the system which lend themselves to a straightforward semantic evaluation. This can sometimes be a real inconvenience to the connectionist model builder when the system as a whole fails to achieve its goal because it has not represented the world the way it should. When this happens, as Smolensky notes,

> [I]t is not necessarily possible to localize a failure of veridical representation. Any particular state is part of a large causal system of states, and failures of the system to meet goal conditions cannot in general be localized to any particular state or state component."[16]

It is connectionist networks of this sort, in which it is not possible to isolate the representation of particular propositions or states of affairs within the nodes, connection strengths and biases, that we have in mind when we talk about the encoding of information in the biases, weights and hidden nodes being *widely distributed* rather than *localist*.

(ii) As we've just noted, there are some connectionist models in which some or all of the units are intended to represent specific properties or features of the system's environment. These units may be viewed as the model's symbols for the properties or features in question. However, in models where the weights and biases have been tuned by learning algorithms it is often not the case that any single unit or any small collection of units will end up representing a specific feature of the environment in any straightforward way.

As we shall see in the next section, it is often plausible to view such networks as collectively or holistically encoding a set of propositions, although none of the hidden units, weights or biases are comfortably viewed as *symbols*. When this is the case we will call the strategy of representation invoked in the model *subsymbolic*. Typically (perhaps always?) networks exploiting subsymbolic strategies of representation will encode information in a widely distributed way.

(iii) The third item on our list is not a feature of connectionist models themselves, but rather a point about how the models are to be interpreted. In making this point we must presuppose a notion of theoretical or explanatory level which, despite much discussion in the recent literature, is far from being a paradigm of clarity.[17] Perhaps the clearest way to introduce the notion of explanatory level is against the background of the familiar functionalist thesis that psychological theories are analogous to programs which can be implemented on a variety of very different sorts of computers.[18] If one accepts this analogy, then it makes sense to ask whether a particular connectionist model is intended as a model at the psychological level or at the level of underlying neural implementation. Because of their obvious, though in many ways very partial, similarity to real neural architectures, it is tempting to view connectionist models as models of the implementation of psychological processes. And some connectionist model builders endorse this view quite explicitly. So viewed, however, connectionist models are not *psychological* or *cognitive* models at all, any more than a story of how cognitive processes are implemented at the quantum mechanical level is a psychological story. A very different view that connectionist model builders can and often do take is that their models are at the psychological level, not at the level of implementation. So construed, the models are in competition with other psychological models of the same phenomena. Thus a connectionist model of word recognition would be an alternative to—and not simply a possible implementation of—a non-connectionist model of word recognition; a connectionist theory of memory would be a competitor to a semantic network theory, and so on. Connectionists who hold this view of their theories often illustrate the point by drawing analogies with other sciences. Smolensky, for example, suggests that connectionist models stand to traditional cognitive models (like semantic networks) in much the same way that quantum mechanics stands to classical mechanics. In each case

the newer theory is deeper, more general and more accurate over a broader range of phenomena. But in each case the new theory and the old are competing at the same explanatory level. If one is right, the other must be wrong.

In light of our concerns in this paper, there is one respect in which the analogy between connectionist models and quantum mechanics may be thought to beg an important question. For while quantum mechanics is conceded to be a *better* theory than classical mechanics, a plausible case could be made that the shift from classical to quantum mechanics was an ontologically *conservative* theory change. In any event, it is not clear that the change was ontologically *radical*. If our central thesis in this paper is correct, then the relation between connectionist models and more traditional cognitive models is more like the relation between the caloric theory of heat and the kinetic theory. The caloric and kinetic theories are at the same explanatory level, though the shift from one to the other was pretty clearly ontologically radical. In order to make the case that the caloric analogy is the more appropriate one, it will be useful to describe a concrete, though very simple, connectionist model of memory that meets the three criteria we have been trying to explicate.

5. A Connectionist Model of Memory

Our goal in constructing the model was to produce a connectionist network that would do at least some of the tasks done by more traditional cognitive models of memory, and that would perspicuously exhibit the sort of distributed, sub-symbolic encoding described in the previous section. We began by constructing a network, we'll call it Network A, that would judge the truth or falsehood of the sixteen propositions displayed above the line in Figure 3. The network was a typical three tiered feed-forward network consisting of 16 input units, four hidden units and one output unit, as shown in Figure 4. The input coding of each proposition is shown in the center column in Figure 3. Outputs close to 1 were interpreted as 'true' and outputs close to zero were interpreted as 'false'. Back propagation, a familiar connectionist learning algorithm was used to "train up" the network thereby setting the connection weights and biases. Training was terminated when the network consistently gave an output higher than .9 for each true proposition and lower than .1 for each false

proposition. Figure 5 shows the connection weights between the input units and the leftmost hidden unit in the trained up network, along with the bias on that unit. Figure 6 indicates the connection weights and biases further upstream. Figure 7 shows the way in which the network computes its response to the proposition *Dogs have fur* when that proposition is encoded in the input units.

There is a clear sense in which the trained up Network A may be said to have stored information about the truth or falsity of propositions (1)-(16), since when any one of these propositions is presented to the network it correctly judges whether the proposition is true or false. In this respect it is similar to various semantic network models which can be constructed to perform much the same task. However, there is a striking difference between Network A and a semantic network model like the one depicted in Figure 1. For, as we noted earlier, in the semantic network there is a functionally distinct sub-part associated with each proposition, and thus it makes perfectly good sense to ask, for any probe of the network, whether or not the representation of a specific proposition played a causal role. In the connectionist network, by contrast, there is no distinct state or part of the network that serves to represent any particular proposition. The information encoded in Network A is stored holistically and distributed throughout the network. Whenever information is extracted from Network A, by giving it an input string and seeing whether it computes a high or a low value for the output unit, *many* connection strengths, *many* biases and *many* hidden units play a role in the computation. And any particular weight or unit or bias will help to encode information about *many* different propositions. It simply makes no sense to ask whether or not the representation of a particular proposition plays a causal role in the network's computation. It is in just this respect that our connectionist model of memory seems radically incongruent with the propositional modularity of common sense psychology. For, as we saw in Section 3, common sense psychology seems to presuppose that there is generally some answer to the question of whether a particular belief or memory played a causal role in a specific cognitive episode. But if belief and memory are subserved by a connectionist network like ours, such questions seem to have no clear meaning.

The incompatibility between propositional modularity and connectionist models like ours can be made even more vivid by contrasting Network A with a second network, we'll call it Network

B, depicted in Figures 8 and 9. Network B was trained up just as the first one was, except that one additional proposition was added to the training set (coded as indicated below the line in Figure 3). Thus Network B encodes all the same propositions as Network A plus one more. In semantic network models, and other traditional cognitive models, it would be an easy matter to say which states or features of the system encode the added proposition, and it would be a simple task to determine whether or not the representation of the added proposition played a role in a particular episode modeled by the system. But plainly in the connectionist network those questions are quite senseless. The point is not that there are no differences between the two networks. Quite the opposite is the case; the differences are many and widespread. But these differences do not correlate in any systematic way with the functionally discrete, semantically interpretable states posited by folk psychology and by more traditional cognitive models. Since information is encoded in a highly distributed manner, with each connection weight and bias embodying information salient to many propositions, and information regarding any given proposition scattered throughout the network, the system lacks functionally distinct, identifiable sub-structures that are semantically interpretable as representations of individual propositions.

The contrast between Network A and Network B enables us to make our point about the incompatibility between common sense psychology and these sorts of connectionist models in a rather different way. We noted in Section 3 that common sense psychology treats predicates expressing the semantic properties of propositional attitudes as projectable. Thus 'believes that dogs have fur' or 'remembers that dogs have fur' will be projectable predicates in common sense psychology. Now both Network A and Network B might serve as models for a cognitive agent who believes that dogs have fur; both networks store or represent the information that dogs have fur. Nor are these the only two. If we were to train up a network on the 17 propositions in Figure 3 plus a few (or minus a few) we would get yet another system which is as different from Networks A and B as these two are from each other. The moral here is that though there are *indefinitely* many connectionist networks that represent the information that dogs have fur just as well as Network A does, these networks have no projectable features in common that are describable in the language of connectionist theory. From the

point of view of the connectionist model builder, the class of networks that might model a cognitive agent who believes that dogs have fur is not a genuine kind at all, but simply a chaotically disjunctive set. Common sense psychology treats the class of people who believe that dogs have fur as a psychologically natural kind; connectionist psychology does not.[19]

6. Objections and Replies

The argument we've set out in the previous five sections has encountered no shortage of objections. In this section we will try to reconstruct the most interesting of these, and indicate how we would reply.

> Objection (i): Models like A and B are not serious models for human belief or propositional memory.

Of course, the models we've constructed are tiny toys that were built to illustrate the features set out in Section 4 in a perspicuous way. They were never intended to model any substantial part of human propositional memory. But various reasons have been offered for doubting that *anything like* these models could ever be taken seriously as psychological models of propositional memory. Some critics have claimed that the models simply will not scale up—that while teaching a network to recognize fifteen or twenty propositions may be easy enough, it is just not going to be possible to train up a network that can recognize a few thousand propositions, still less a few hundred thousand.[20] Others have objected that while more traditional models of memory, including those based on sentence-like storage, those using semantic networks, and those based on production systems, all provide some strategy for *inference* or *generalization* which enables the system to answer questions about propositions it was not explicitly taught, models like those we have constructed are incapable of inference and generalization. It has also been urged that these models fail as accounts of human memory because they provide no obvious way to account for the fact that suitably prepared humans can easily acquire propositional information one proposition at a time. Under ordinary circumstances, we can just *tell* Henry that the car keys are in the refrigerator, and he can readily record this fact in memory. He doesn't need anything

like the sort of massive retraining that would be required to teach one of our connectionist networks a new proposition.

Reply: If this were a paper aimed at defending connectionist models of propositional memory, we would have to take on each of these putative shortcomings in some detail. And in each instance there is at least something to be said on the connectionist side. Thus, for example, it just is not true that networks like A and B don't generalize beyond the propositions on which they've been trained. In Network A, for example, the training set included:

Dogs have fur	Cats have fur.
Dogs have paws	Cats have paws.
Dogs have fleas	Cats have fleas.

It also included

Dogs have legs.

but not

Cats have legs.

When the network was given an encoding of this last proposition, however, it generalized correctly and responded affirmatively. Similarly, the network responded negatively to an encoding of

Cats have scales

though it had not previously been exposed to this proposition.

However, it is important to see that this sort of point by point response to the charge that networks like ours are inadequate models for propositional memory is not really required, given the thesis we are defending in this paper. For what we are trying to establish is a *conditional* thesis: *if* connectionist models of memory of the sort we describe in Section 4 are right, *then* propositional attitude psychology is in serious trouble. Since conditionals with false antecedents are true, we win by default if it turns out that the antecedent of our conditional is false.

Objection (ii): Our models do not really violate the principle of propositional modularity, since the propositions the system has learned are coded in functionally discrete ways, though this may not be obvious.

We've heard this objection elaborated along three quite different lines. The first line—let's call it Objection (iia)— notes that functionally discrete coding may often be *very* hard to notice, and can not be expected to be visible on casual inspection. Consider, for example, the way in which sentences are stored in the memory of a typical von Neuman architecture computer—for concreteness we might suppose that the sentences are part of an English text and are being stored while the computer is running a word processing program. Parts of sentences may be stored at physically scattered memory addresses linked together in complex ways, and given an account of the contents of all relevant memory addresses one would be hard put to say where a particular sentence is stored. But nonetheless each sentence is stored in a *functionally discrete* way. Thus if one knew enough about the system it would be possible to erase any particular sentence it is storing by tampering with the contents of the appropriate memory addresses, while leaving the rest of the sentences the system is storing untouched. Similarly, it has been urged, connectionist networks may in fact encode propositions in functionally discrete ways, though this may not be evident from a casual inspection of the trained up network's biases and connection strengths.

Reply (iia): It is a bit difficult to come to grips with this objection, since what the critic is proposing is that in models like those we have constructed there *might* be some covert functionally discrete system of propositional encoding that has yet to be discovered. In response to this we must concede that indeed there might. We certainly have no argument that even comes close to demonstrating that the discovery of such a covert functionally discrete encoding is impossible. Moreover, we concede that if such a covert system were discovered, then our argument would be seriously undermined. However, we're inclined to think that the burden of argument is on the critic to show that such a system is not merely possible but *likely*; in the absence of any serious reason to think that networks like ours do encode propositions in functionally discrete ways, the mere logical possibility that they might is hardly a serious threat.

The second version of Objection (ii)—we'll call it Objection (iib)—makes a specific proposal about the way in which networks like A and B might be discretely, though covertly, encoding propositions. The encoding, it is urged, is to be found in the pattern of activation of the hidden nodes, when a given proposition is presented to the

network. Since there are four hidden nodes in our networks, the activation pattern on presentation of any given input may be represented as an ordered 4-tuple. Thus, for example, when network A is presented with the encoded proposition *Dogs have fur*, the relevant 4-tuple would be (21, 75, 73, 12), as shown in Figure 7. Equivalently, we may think of each activation pattern as a point in a four dimensional hyperspace. Since each proposition corresponds to a unique point in the hyperspace, that point may be viewed as the encoding of the proposition. Moreover, that point represents a functionally discrete state of the system.[21]

Reply (iib): What is being proposed is that the pattern of activation of the system on presentation of an encoding of the proposition *p* be identified with the belief that *p*. But this proposal is singularly implausible. Perhaps the best way to see this is to note that in common sense psychology beliefs and propositional memories are typically of substantial duration; and they are the sorts of things that cognitive agents generally have lots of even when they are not using them. Consider an example. Are kangaroos marsupials? Surely you've believed for years that they are, though in all likelihood this is the first time today that your belief has been activated or used.[22] An activation pattern, however, is not an enduring state of a network; indeed, it is not a state of the network at all except when the network has had the relevant proposition as input. Moreover, there is an enormous number of other beliefs that you've had for years. But it makes no sense to suppose that a network could have many activation patterns continuously over a long period of time. At any given time a network exhibits at most one pattern of activation. So activation patterns are just not the sorts of things that can plausibly be identified with beliefs or their representations.

Objection (iic): At this juncture, a number of critics have suggested that long standing beliefs might be identified not with activation patterns, which are transient states of networks, but rather with *dispositions to produce activation patterns*. Thus, in network A, the belief that dogs have fur would not be identified with a location in activation hyperspace but with the network's *disposition* to end up at that location when the proposition is presented. This *dispositional state* is an enduring state of the system; it is a state the network can be in no matter what its current state of activation may be, just as a sugar cube may have a disposition to dissolve in water even when there is no water nearby.[23] Some have gone on to suggest that the

familiar philosophical distinction between dispositional and occurrent beliefs might be captured, in connectionist models, as the distinction between dispositions to produce activation patterns and activation patterns themselves.

Reply (iic): Our reply to this suggestion is that while dispositions to produce activation patterns are indeed *enduring* states of the system, they are not the right sort of enduring states—they are not the discrete, independently causally active states that folk psychology requires. Recall that on the folk psychological conception of belief and inference, there will often be a variety of quite different underlying causal patterns that may lead to the acquisition and avowal of a given belief. When Clouseau says that the butler did it, he may have just inferred this with the help of his long standing belief that the train is out of service. Or he may have inferred it by using his belief that the hotel is closed. Or both long standing beliefs may have played a role in the inference. Moreover, it is also possible that Clouseau drew this inference some time ago, and is now reporting a relatively long standing belief. But it is hard to see how anything like these distinctions can be captured by the dispositional account in question. In reacting to a given input, say p, a network takes on a specific activation value. It may also have dispositions to take on other activation values on other inputs, say q and r. But there is no obvious way to interpret the claim that these further dispositions play a causal role in the network's reaction to p—or, for that matter, that they do not play a role. Nor can we make any sense of the idea that on one occasion the encoding of q (say, the proposition that the train is out of service) played a role while the encoding of r (say, the proposition that the hotel is closed) did not, and on another occasion, things went the other way around. The propositional modularity presupposed by common sense psychology requires that belief tokens be functionally discrete states capable of causally interacting with one another in some cognitive episodes and of remaining causally inert in other cognitive episodes. However, in a distributed connectionist system like Network A, the dispositional state which produces one activation pattern is functionally inseparable from the dispositional state which produces another. Thus it is impossible to isolate some propositions as causally active in certain cognitive episodes, while others are not. We conclude that reaction pattern dispositions won't do as belief tokens. Nor, so far as we can see, are there any other states of networks like A and B that will fill the bill.

7. Conclusion

The thesis we have been defending in this paper is that connectionist models of a certain sort are incompatible with the propositional modularity embedded in common sense psychology. The connectionist models in question are those which are offered as models at the *cognitive* level, and in which the encoding of information is widely distributed and subsymbolic. In such models, we have argued, there are no *discrete*, *semantically interpretable* states that play a *causal role* in some cognitive episodes but not others. Thus there is, in these models, nothing with which the propositional attitudes of common sense psychology can plausibly be identified. If these models turn out to offer the best accounts of human belief and memory, we will be confronting an *ontologically radical* theory change—the sort of theory change that will sustain the conclusion that propositional attitudes, like caloric and phlogiston, do not exist.

Notes

1. Thanks are due to Ned Block, Paul Churchland, Gary Cottrell, Adrian Cussins, Jerry Fodor, John Heil, Frank Jackson, David Kirsh, Patricia Kitcher and Philip Kitcher for useful feedback on earlier versions of this paper. Talks based on the paper have been presented at the UCSD Cognitive Science Seminar and at conferences sponsored by the Howard Hughes Medical Foundation and the University of North Carolina at Greensboro. Comments and questions from these audiences have proved helpful in many ways.
2. See, for example, Churchland (1981) & (1986), where explicitly eliminativist conclusions are drawn on the basis of speculations about the success of cognitive models similar to those we shall discuss.
3. Fodor, J. & Pylyshyn, Z. (1988).
4. We are aware that certain philosophers and historians of science have actually entertained ideas similar to the suggestion that the planets spoken of by pre-Copernican astronomers do not exist. See, for example, Kuhn (1970), Ch. 10, and Feyerabend (1981), Ch. 4. However, we take this suggestion to be singularly implausible. Eliminativist arguments can't be that easy. Just what has gone wrong with the accounts of meaning and reference that lead to such claims is less clear. For further discussion on these matters see Kuhn (1983), and Kitcher (1978) & (1983).
5. For some detailed discussion of scientific reduction, see Nagel (1961); Schaffner (1967); Hooker (1981); and Kitcher (1984). The genetics case is not without controversy. See Kitcher (1982) & (1984).

6. It's worth noting that judgments on this matter can differ quite substantially. At one end of the spectrum are writers like Feyerabend (1981), and perhaps Kuhn (1962), for whom relatively small differences in theory are enough to justify the suspicion that there has been an ontologically radical change. Toward the other end are writers like Lycan, who writes:

> I am at pains to advocate a very liberal view...I am entirely willing to give up fairly large chunks of our commonsensical or platitudinous theory of belief or of desire (or of almost anything else) and decide that we were just wrong about a lot of things, without drawing the inference that we are no longer talking about belief or desire .... I think the ordinary word "belief" (qua theoretical term of folk psychology) points dimly toward a natural kind that we have not fully grasped and that only mature psychology will reveal. I expect that "belief" will turn out to refer to some kind of information bearing inner state of a sentient being..., but the kind of state it refers to may have only a few of the properties usually attributed to beliefs by common sense. (Lycan (1988), pp. 31-2.)

   On our view, both extreme positions are implausible. As we noted earlier, the Copernican revolution did not show that the planets studied by Ptolemy do not exist. But Lavosier's chemical revolution *did* show that phlogiston does not exist. Yet on Lycan's "very liberal view" it is hard to see why we should not conclude that phlogiston really does exist after all—it's really oxygen, and prior to Lavosier "we were just very wrong about a lot of things".
7. For an early and influential statement of the view that common sense psychology is a theory, see Sellars (1956). More recently the view has been defended by Churchland (1970) & (1979), Chs. 1 & 4; and by Fodor (1988), Ch. 1. For the opposite view, see Wilkes (1978); Madell (1986); Sharpe (1987).
8. See Stich (1983), pp. 237 ff.
9. Cherniak (1986), Ch. 3, notes that this sort of absent mindedness is commonplace in literature and in ordinary life, and sometimes leads to disastrous consequences.
10. For sentential models, see John McCarthy (1968), (1980), & (1986); and Kintsch (1974). For semantic networks, see Quillian (1969); Collins & Quillian (1972); Rumelhart, Lindsay & Norman (1972); Anderson & Bower (1973); and Anderson (1976) & (1980), Ch. 4. For production systems, see Newell & Simon (1972); Newell (1973); Anderson (1983); and Holland, et. al. (1986).
11. For the classic discussion of the distinction between projectable and non-projectable predicates, see Goodman (1965).
12. See, for example, Anderson & Bower (1973).
13. Emphasis added.
14. Smolensky (1988), p. 1.
15. Fodor & Pylyshyn (1988), p. 57.

16. Smolensky (1988), p. 15.
17. Broadbent, D. (1985); Rumelhart & McClelland (1985); Rumelhart & McClelland (1986), Ch. 4; Smolensky (1988); Fodor & Pylyshyn (1988).
18. The notion of program being invoked here is itself open to a pair of quite different interpretations. For the right reading, see Ramsey (1989).
19. This way of making the point about the incompatibility between connectionist models and common sense psychology was suggested to us by Jerry Fodor.
20. This point has been urged by Daniel Dennett, among others.
21. Quite a number of people have suggested this move, including Gary Cottrell, & Adrian Cussins.
22. As Lycan notes, on the common sense notion of belief, people have lots of them "even when they are asleep." (Lycan (1988), p. 57.)
23. Something like this objection was suggested to us by Ned Block and by Frank Jackson.

References

Andersón, J. & Bower, G. (1973). *Human Associative Memory*, Washington, D. C., Winston.

Anderson, J. (1976). *Language, Memory and Thought*, Hillsdale, N.J., Lawrence Erlbaum Associates.

Anderson, J. (1980). *Cognitive Psychology and Its Implications* San Francisco, W. H. Freeman & Co.

Anderson, J. (1983). *The Architecture of Cognition*, Cambridge, MA, Harvard University Press.

Broadbent, D. (1985). "A Question of Levels: Comments on McClelland and Rumelhart," *Journal of Experimental Psychology: General*, 114.

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

Churchland, P. (1970). "The Logical Character of Action Explanations," *Philosophical Review*, 79.

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

Churchland, P. (1981). "Eliminative Materialism and Propositional Attitudes," *Journal of Philosophy*, 78, 2.

Churchland, P. (1986). "Some Reductive Strategies in Cognitive Neurobiology," *Mind*, 95.

Collins, A. & Quillian, M. (1972). "Experiments on Semantic Memory and Language Comprehension," in L. Gregg, ed., *Cognition in Learning and Memory*, New York, Wiley.

Feyerabend, P. (1981). *Realism, Rationalism and Scientific Method: Philosophical Papers Vol. 1*, Cambridge, Cambridge University Press.

Fodor, J. & Pylyshyn, Z. (1988). "Connectionism and Cognitive Architecture: A Critical Analysis," *Cognition*, 28.

Fodor, J. (1987). *Psychosemantic: The Problem of Meaning in the Philosophy of Mind*, Cambridge, MA, Bradford Books/MIT Press.

Goodman, N. (1965). *Fact, Fiction and Forecast*, Indianapolis, Bobbs-Merrill.
Holland, J., Holyoak, K., Nisbett, R. & Thagard, P. (1986). *Induction: Processes of Inference, Learning and Discovery*, Cambridge, MA, Bradford Books/MIT Press.
Hooker, C. (1981). "Towards a General Theory of Reduction," Parts I, II & III, *Dialogue*, 20.
Kintsch, W. (1974). *The Representation of Meaning in Memory*, Hillsdale, N.J., Lawrence Erlbaum Associates.
Kitcher, P. (1978). "Theories, Theorists and Theoretical Change," *Philosophical Review*, 87.
Kitcher, P. (1982). "Genes," *British Journal for the Philosophy of Science*, 33.
Kitcher, P. (1983). "Implications of Incommensurability," *PSA 1982*, (Proceedings of the 1982 Biennial Meeting of the Philosophy of Science Association) Vol. 2, ed. by P. Asquith & T. Nickles, East Lansing, Philosophy of Science Association.
Kitcher, P. (1984). "1953 and All That: A Tale of Two Sciences," *Philosophical Review*, 93.
Kuhn, T. (1962). *The Structure of Scientific Revolutions*, Chicago, University of Chicago Press. 2nd Edition (1970).
Kuhn, T. (1983). "Commensurability, Comparability, Communicability," *PSA 1982*, (Proceedings of the 1982 Biennial Meeting of the Philosophy of Science Association) Vol. 2, ed. by P. Asquith & T. Nickles, East Lansing, Philosophy of Science Association.
Lycan, W. (1988). *Judgement and Justification*, Cambridge, Cambridge University Press.
Madell, G. (1986). "Neurophilosophy: A Principled Skeptic's Response," *Inquiry*, 29.
McCarthy, J. (1968). "Programs With Common Sense," in M. Minsky, ed., *Semantic Information Processing*, Cambridge, MA, MIT Press.
McCarthy, J. (1980). "Circumscription: A Form of Non-Monotonic Reasoning," *Artificial Intelligence*, 13.
McCarthy, J. (1986). "Applications of Circumscription to Formalizing Common-Sense Knowledge," *Artificial Intelligence*, 28.
Nagel, E. (1961). *The Structure of Science*, New York, Harcourt, Brace & World.
Newell, A. & Simon, H. (1972). *Human Problem Solving*, Englewood Cliffs, N.J., Prentice Hall.
Newell, A. (1973). "Production Systems: Models of Control Structures," in W. Chase, ed., *Visual Information Processing*, New York, Academic Press.
Quillian, M. (1966). *Semantic Memory*, Cambridge, MA, Bolt, Branak & Newman.
Ramsey, W. (1989). "Parallelism and Functionalism," *Cognitive Science*, 13.
Rumelhart, D. & McClelland, J. (1985). "Level's Indeed! A Response to Broadbent," *Journal of Experimental Psychology: General*, 114.
Rumelhart, D., Lindsay, P. & Norman, D. (1972). "A Process Model for Long Term Memory," in E. Tulving & W. Donaldson, eds., *Organization of Memory*, New York, Academic Press.

Rumelhart, D., McClelland, J. & the PDP Research Group (1986). *Parallel Distributed Processing*, Volumes I & II, Cambridge, MA, Bradford Books/MIT Press.

Sellars, W. (1956). "Empiricism and the Philosophy of Mind," *Minnesota Studies in the Philosophy of Science*, Vol. I, H. Feigl & M. Scriven, eds., Minneapolis, University of Minnesota Press.

Schaffner, K. (1967). "Approaches to Reduction," *Philosophy of Science*, 34.

Sharp, R. (1987). "The Very Idea of Folk Psychology," *Inquiry*, 30.

Smolensky, P. (1988). "On the Proper Treatment of Connectionism," *The Behavioral & Brain Sciences*, 11.

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

Wilkes, K. (1978). *Physicalism*, London, Routledge and Kegan Paul.

FIGURE 1

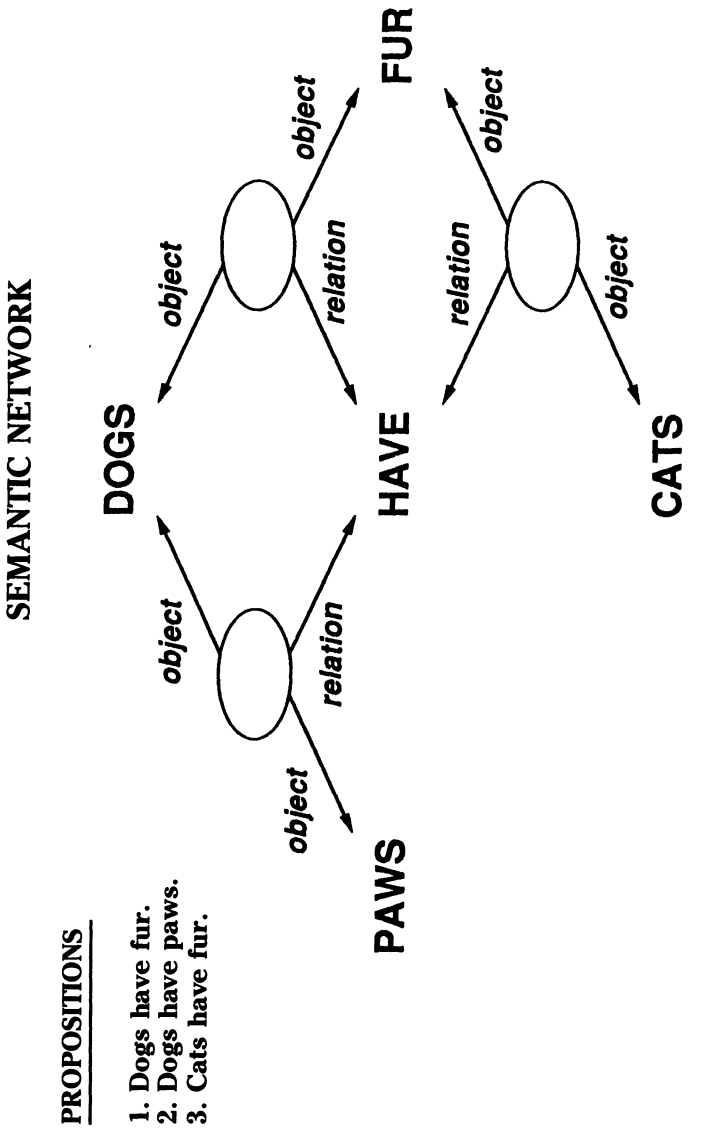


FIGURE 2

	Proposition	Input	Output
1	Dogs have fur.	11000011 00001111	1 true
2	Dogs have paws.	11000011 00110011	1 true
3	Dogs have fleas.	11000011 00111111	1 true
4	Dogs have legs.	11000011 00111100	1 true
5	Cats have fur.	11001100 00001111	1 true
6	Cats have paws.	11001100 00110011	1 true
7	Cats have fleas.	11001100 00111111	1 true
8	Fish have scales.	11110000 00110000	1 true
9	Fish have fins.	11110000 00001100	1 true
10	Fish have gills.	11110000 00000011	1 true
11	Cats have gills.	11001100 00000011	0 false
12	Fish have legs.	11110000 00111100	0 false
13	Fish have fleas.	11110000 00111111	0 false
14	Dogs have scales.	11000011 00110000	0 false
15	Dogs have fins.	11000011 00001100	0 false
16	Cats have fins.	11001100 00001100	0 false
	Added Proposition		
17	Fish have eggs.	11110000 11001000	1 true

FIGURE 3

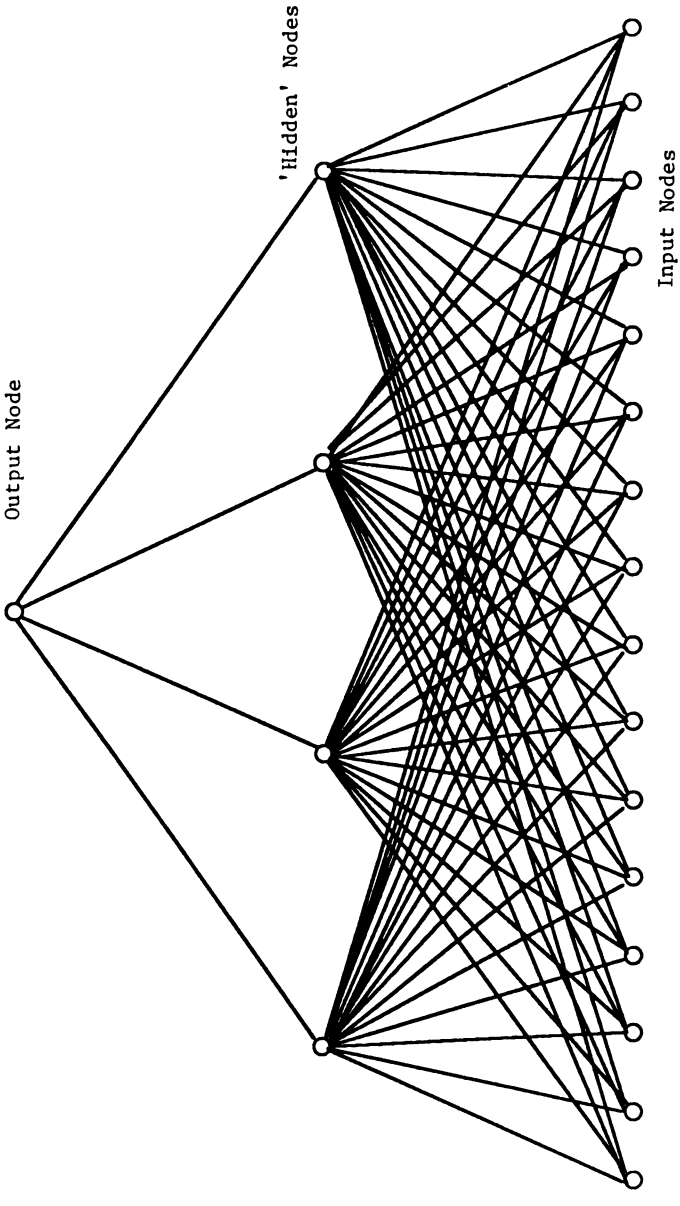


FIGURE 4

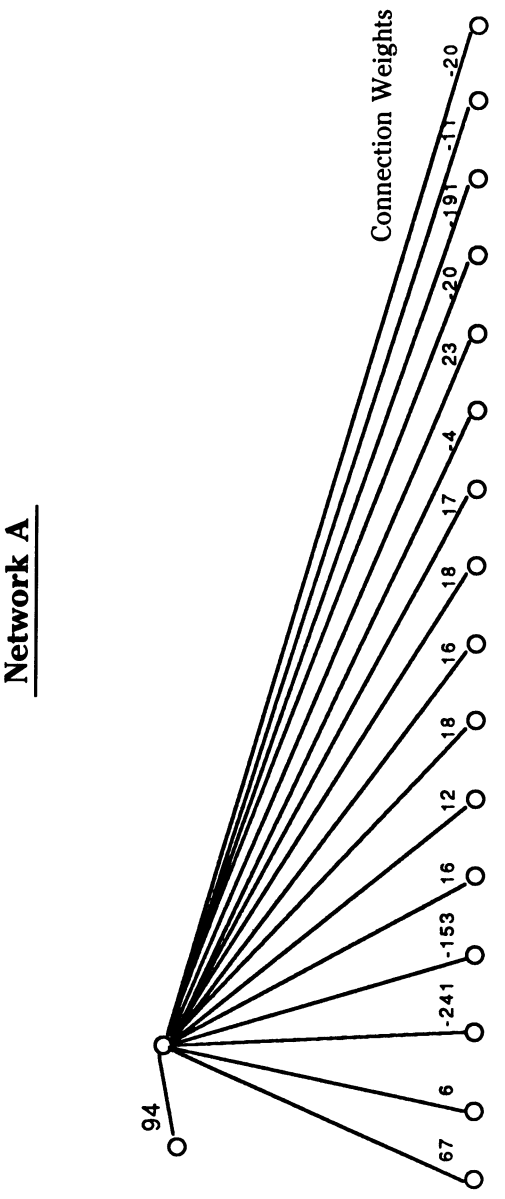
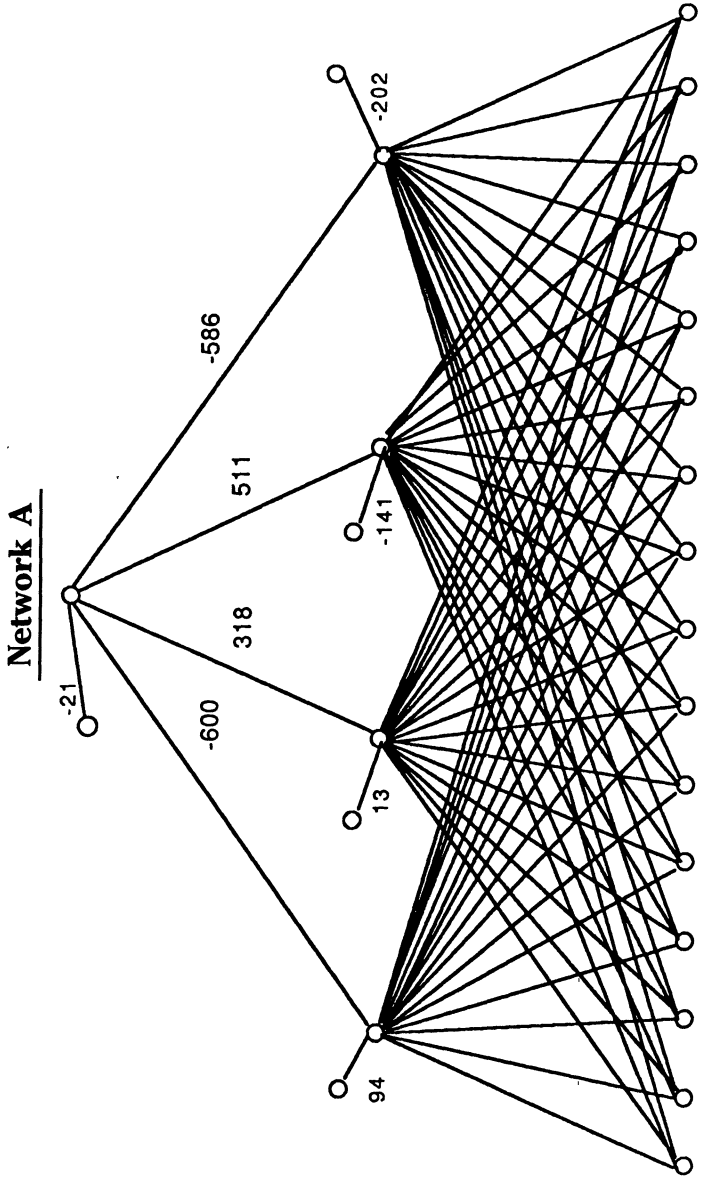


FIGURE 5



Weights and biases in network with 16 propositions.

FIGURE 6

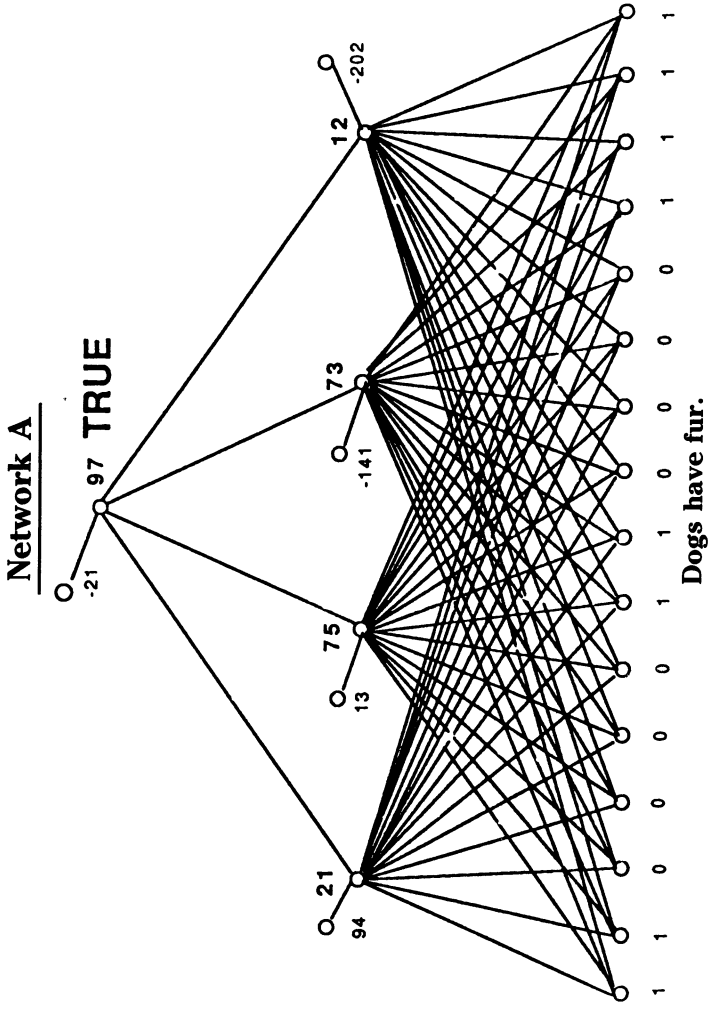
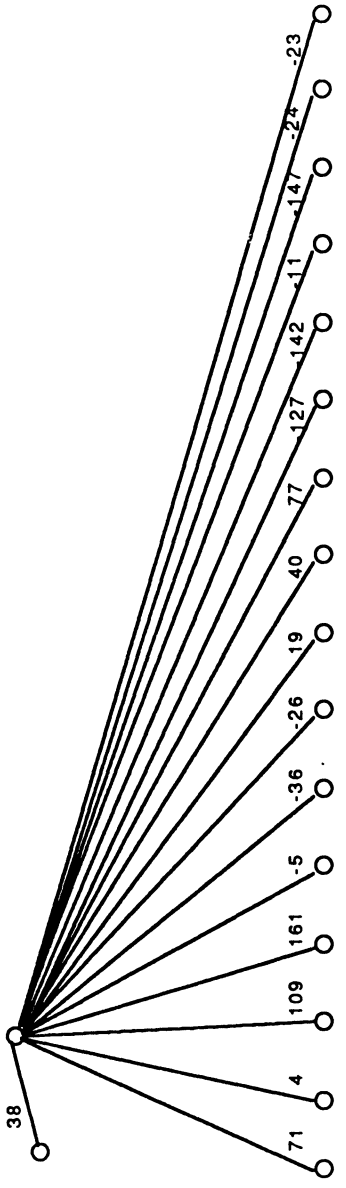
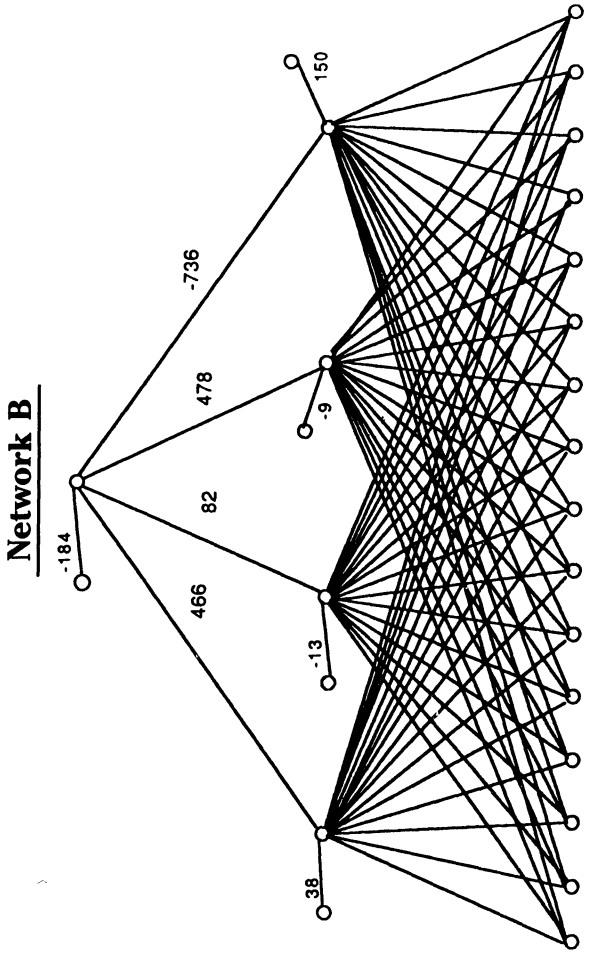


FIGURE 7



Input weights and bias to first hidden node in network with 17 propositions.

FIGURE 8



Weights and biases in network with 17 propositions.

FIGURE 9