Networks with Attitudes

Paul Skokowski
Symbolic Systems, Stanford University
Stanford CA 94305-2150
paulsko-atsign-turing.stanford.edu

Abstract

Does connectionism spell doom for folk psychology? I examine the proposal that cognitive representational states such as beliefs can play no role if connectionist models -- interpreted as radical new cognitive theories -- take hold and replace other cognitive theories. Though I accept that connectionist theories are radical theories that shed light on cognition, I reject the conclusion that neural networks do not *represent*. Indeed, I argue that neural networks may actually give us a better working notion of cognitive representational states such as beliefs, and in so doing give us a better understanding of how these states might be instantiated in neural wetware.

1. Introduction

An important article by Ramsey, Stich and Garon (1991) offers the tempting thesis that connectionist networks support eliminativism with regards to folk psychological notions such as belief or memory. They buttress their arguments with examples of neural networks they have trained to encode propositions. I say their thesis is tempting because I agree with them that connectionism does offer a radically different way of interpreting cognition and cognitive states, and because I agree with them that many existing theories which purport to be foundational for the propositional attitudes fall far short of that goal. But even given this level of agreement I am not yet ready to conclude that the death knell for belief has been rung by the neural network theorists. The reason for this is that, notwithstanding the arguments given by Ramsey, Stich and Garon, I believe neural networks may actually give us a better working notion of cognitive states such as beliefs or memories, and in so doing, give us a better understanding of how these states might be instantiated in neural wetware.

Though the article was first published in 1991, and produced a small cottage-industry of critical and supportive papers in its aftermath (O'Brien 1991, Stich 1991, Forster and Seidel 1994, Ramsey 1994, Clark 1995, Stich and Warfield 1995), I believe it is worth revisiting, for two reasons. First, because the implication of eliminativism is intriguing by posing a threat to psychology, which relies heavily on the notion of belief in the explanation of human and social behaviour; and second, because I feel that the above-mentioned papers missed the most salient points about connectionism – points that are crucial to understanding how connectionism provides a satisfying explanation for belief in biological neural systems – systems like *us*.

In this article I will concentrate on countering Ramsey, Stich and Garon's (1991, 217) claim that connectionist models have "...no *discrete*, *semantically interpretable* states that play a *causal role* in some cognitive episodes but not others." The states that they refer to here are belief or memory states, and I will follow their lead in using these terms interchangeably. I will therefore consider my job done if I manage to save a working notion of belief from their frontal assaults. Doing this will, I believe, throw doubt upon their entire eliminativist program for the propositional attitudes.

2. The Problem

Ramsey, Stich and Garon start with the assumption that folk psychology is a theory, and that states such as beliefs, desires and so on are posits of the theory.[1] They argue that this theory is a prime candidate for replacement because it can't possibly be telling us all there is to know about psychology. What is crucial to the Folk Psychologist's program, they argue, is the claim of *Propositional Modularity* (1991, 204). Propositional Modularity holds that the propositional attitudes are:

1. functionally discrete
2. semantically interpretable, and
3. play a causal role (in mental and behavioral output).

Thus, in classical folk psychological models it is clear when a functionally distinct representation such as a belief plays a causal role. But Ramsey, Stich and Garon point out that there are classes of connectionist models that fly in the face of Propositional Modularity. These connectionist models become candidates to replace their folk psychological counterparts.

They offer as examples two networks of their own design: Network A and Network B. Both encode simple propositions, such as 'Cats have fur' and 'Dogs have legs' with binary strings of length 16. These binary strings count as input to the 16 input nodes of the network. Four units comprise the hidden unit layer, and there is a single output node which (after training) will register a '1' (or very close to it) for a true proposition and a '0' (or very close to it) for a false proposition. Network A was trained on 16 propositions using back propagation until it was accurate at distinguishing true from false in the training set. It then was seen to 'generalize', for it gave an affirmative answer to the new proposition (not in its training set) 'Cats have legs', and negatively to the proposition 'Cats have scales'. Network B was just like Network A in architecture, but its training set included one additional proposition, 'Fish have eggs', for a total of 17 propositions in the training set. Network B performed similarly after training to Network A in accuracy and generalization.

In contrast with classical models, say Ramsey, Stich and Garon, connectionist networks like Networks A and B have no distinct states or parts that serve to represent particular propositional contents. Information storage is distributed across the network and is holistic. Following Smolensky (1988), this sort of representation is termed *subsymbolic*. Thus, any particular unit or weight value can encode information about many different contents.

Connectionist models of this sort, they claim, have three properties:(Ramsey, Stich and Garon, 1991, 207)

- their encoding of information in the weights is *widely distributed*, not localist
- the individual units have no symbolic interpretation -- they are *subsymbolic*
- the models are not intended as implementations, but as true (and ontologically radical) cognitive theories that compete with traditional cognitive theories.

Given the stark contrast between propositional modularity and connectionist models, Ramsey, Stich and Garon remark that:

> 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 ... 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.(Ramsey, Stich and Garon 1991, 212)

Since connectionist networks lack modular propositional states, they won't have the discrete features required to make them fall under psychological generalizations. Classically, seeing an F generally leads to me believing B, that F. A law connects the object F with my belief B. But according to the analysis given by Ramsey, Stich and Garon, there are no discrete, functionally distinct, belief states or structures like B that are implemented by all the networks that appear to exemplify such beliefs. Thus Network A's belief that F will differ from Network B's belief that F, since the individual weights and unit activations, and hence their internal representations, are necessarily different. They go on to claim that "these networks have no projectable features in common that are describable in the language of connectionist theory."(Ramsey 1991, 213)

In what follows, I will argue against Ramsey, Stich and Garon to show that neural networks do have states which satisfy the claims 1-3 of Propositional Modularity. These states can include a conjunction of input, hidden unit, and weight states, as will be discussed below in sections 3 and 4. I will then show in section 5 how a proper interpretation of these states relates to the existing literature on Ramsey, Stich and Garon's paper, and avoids eliminativism with respect to belief.

3. Neural Correlates of Belief

There is a lot I agree with in the account offered by Ramsey, Stich and Garon above. I agree that encoding in connectionist networks is distributed, and that individual units rarely have a symbolic interpretation. But I don't think that accepting such things means that such networks don't represent *at all*. There are, I believe, connectionist -- and ultimately, *neural* -- correlates of belief.

Let me start by debunking a myth. The myth is that neural networks don't have distinct states or parts that serve to represent particular contents. Consider again Networks A and B which were claimed above to lack distinct states that represented particular propositional contents. We seem to have conveniently forgotten the input units here. These represent the propositional contents in a distinct and straightforward way, and they have the added convenience that they are part of the network in question. So it appears the network does represent propositions. It is even a distributed representation, but then the English sentence "Cats have fur" is also a distributed representation -- distributed across letters -- of the proposition, or content, *cats have fur.*[2]

This might appear to be cheating, but it's not. Presumably what connectionist models are going to be useful for is explaining human (and other animal) cognitive phenomena. But humans have their analogues of input units too: the senses and their wiring into the brain. These inputs vary nomically with outside conditions. When an infant and an adult look at a flower, they both have nearly identical retinal, optical tract, optical chiasm, and cortical (V1-V4, say) stimulations. Their retinas, and the rest of their sensory delivery systems, then, carry the same information. This information is distributed: the entire retina may be stimulated, and similarly for the bundles of neurons delivering the signals further downstream in a parallel (and distributed) fashion. The sensory systems for both the infant and adult vary in lawlike, and very similar, ways with the external environment. It is what happens after that information has been picked up -- a story to do with *learning* -- which determines what content is available for behavioral output for the agent.[3] The adult believes the object is a flower, and he can behave in appropriate ways. The infant, on the other hand, lacks the appropriate beliefs; she hasn't learned about flowers yet. Similarly for a neural network. The information at the input level is a lot like sensory information. It is only after learning that the network can distinguish categories in the training set.

It is important to understand the difference between two kinds of physical properties in a neural network. The first kind of property is that which occurs when the units are *activated*, say by the presentation of an **F**. This is a property that occurs at a time. Electrical signals pass through the network upon such a presentation. Consider a network which has *learned* to recognize **F**'s. The input units will exemplify a characteristic activation pattern, call it **I**, corresponding to an **F** when presented with one. Similarly, the output units exemplify a characteristic output pattern, call it **O**, upon such a presentation. But also notice that the hidden units exemplify an activation pattern after learning, call it **H**. The property **H** exemplified by the hidden units is different from the (learned) final weight configuration, call it **W**. The former property **H** is a transient one, it occurs at a time. The latter property is stable; it lasts for more than the moment over which electrical signals stimulate the network to be activated.

After learning, the network's hidden units may exemplify two sorts of properties, **H** and **W**. Note that these are like the properties of real neurons instantiated in wetware. A collection of neurons may fire in some manner, thus exhibiting a property which is the analogue of an activation pattern **H** in a neural network. A collection of neurons also has the fairly stable property of intersynaptic connections. This property is the analogue of the property **W** in neural networks.

We know that after learning, the weight structure **W** becomes a stable, permanent feature of a neural network. Since it is stable, it won't change with different incoming signals. It also isn't a property that is exemplified only when an input arrives -- it is there over long periods of time whether signals are coming in or not. As such, this is not the sort of property which we would tend to associate with a regularity, such as the regularity which occurs between the input units and outside conditions **F**: When an **F** is presented under the right conditions, a pattern **I** will occur on the input units. As pointed out above, this is the sort of regularity we normally associate with the carrying of information.

According to Ramsey, Stich and Garon (1991, 215-217), neither of the states **H** or **W** occurring in networks can be considered to be beliefs or memories. The activation pattern **H** won't do because it is transient, and beliefs are supposed to be enduring. Thus John believes that kangaroos are marsupials even when he isn't thinking about kangaroos. The weight structure **W** won't do because they find it extremely implausible that weights encode content in functionally discrete ways. That is, it is unlikely that **W** has discrete encoding properties corresponding to properties in the environment (or a training set). However, they do remark that there might indeed be some system of encoding in the weights that they are unfamilar with. And, "Moreover we concede that if such a covert system were discovered, then our argument would be seriously undermined."(Ramsey, Stich and Garon 1991, 215) We will return to this point below.

Since neither activation patterns nor weight states fit Ramsey, Stich and Garon's criteria for representational states such as beliefs or memories, there are no representational states in neural networks. They have been eliminated in the brave new world of connectionism. As I have said before, though I am sympathetic, I remain unconvinced. In what follows, I propose how we should interpret belief in networks.

4. Belief in Networks

The first thing we should recognize is that the story of belief that has just been told is incomplete. It ignores that beliefs are also causes of output, which can take the form of actions in agents, or output patterns in networks. This is crucial for establishing the third claim of Propositional Modularity about the causal role of beliefs. A belief must be a physical occurrence of an internal state at a time, which can cause appropriate action at that time. A tree is in front of me and I see it. Because I believe the tree is in front of me, I swerve to the side on my run through the park. Does this mean I have to have an enduring tree-belief? This seems implausible for this sort of perceptual belief. What I need are cognitive capacities for recognizing trees. A tree-recognition neural event -- the actual occurrence at a time -- is a belief. If this sort of belief was forced to fit in the straightjacket of 'enduring' beliefs given above, then by having the belief permanently, I would be forever swerving. But I am not. Consider now a neural network that has been taught to recognize trees. It has been fitted with a digital camera front-end which feeds an input layer, and an output layer that drives a speech-synthesizer. It too only responds to trees when presented with one. It says "tree". But once it is taught, it doesn't say "tree" all day long -- only when one is presented.

I don't happen to believe in belief-boxes or grandmother cells. These would be, I presume, places in the brain where particular propositional contents are stored. But one does not need such artifices in order to accommodate belief. If one has a causal notion of belief, then believing **F** is a matter of encountering or perceiving **F** in appropriate circumstances. And the circumstances are given with extreme simplicity in the case of neural networks. They give us a very powerful, mechanistic, and simplified model for what appears to be going on in the brain under certain conditions.

The internal circumstances are provided by learning. In Skokowski (2004) I have shown how learning, which involves actual physical encounters with the environment (or training set), is what installs a weight state **W**, and determines its contents. These are implicit contents that are acquired naturalistically, and that play a real causal role in the behavior of the network.

Without learning, one cannot hope to attain the regularities associated with belief (beyond the information-carrying capacity of the input units). That is, without learning one cannot hope to attain outputs appropriate to the training task: yielding a "1" when given the proposition "Cats have Fur"; saying "tree" when presented with a tree;

swerving when encountering a tree on a run. Learning, by installing an enduring weight state **W**, delivers the background conditions required for an informational state, such as a perceptual or input state, to get an executive capacity and cause output.(Skokowski 2004, 367-368)

Before learning, an infant or a neural network may carry information about its surroundings, but neither will yet have a belief, something that can guide its behavioral output. The infant "sees" trees, but does not recognize them; does not have beliefs about them. The network "sees" trees in its input units, but does not produce the sound "tree" in its output. Learning corrects this deficit. Not by changing anything at the input level. The input, or sensory, states still carry informational content in the same way. They covary nomically with external conditions. What is changed is the internal weight state, or neural structure, of the system. Note that the causal work for an occurrent belief or memory is done by the electrical signals in a neural network, or the electro-chemical signals in the brain. This is the transient activation state. But the background weight state **W** acts to modify or guide the signal in a way that produces output appropriate to the input. In this way, weights encode the latent ability to construct states that correspond to occurrent beliefs.

Beliefs should cause output when they occur in an agent. Beliefs should be caused by appropriate external conditions. Beliefs should carry representational content. Belief in networks, then, should be seen as the activation pattern occurring in the input and hidden units, *after* learning. This includes what I earlier called **I** together with the hidden unit activation pattern **H**. Call this combined state **B**. It is important that we don't have **B** until after learning. Just like the infant or the neural network, we lack cognitive abilities until we have learned them. So being presented with a tree before learning won't evoke a response appropriate to a belief about trees. But *after* learning, **B** causes appropriate output. **B** is also caused by appropriate external conditions. When presented with the proposition "Cats have Fur", Network A registers input 1111000011110000 on its input layer, which in turn causes further activation in the network. When perceiving a tree on a run, the adult swerves. **B** also carries representational content. This is guaranteed by including the input activation **I** as part of **B**. **I**, through sensory covariation with properties in the environment, carries content. **B** therefore does by default.[4]

5. Interpreting States

Earlier, Ramsey, Stich and Garon discounted the possibility of the weights **W** of a neural network encoding contents. They conceded such encoding would severely weaken their argument. Clustering techniques for exploring weight space such as Principal Components Analysis (PCA) have been around for a while now and have been used successfully to find correlations between weight space properties and properties in the subject matter (training set/environment) being learned. In particular, studies on language tasks by Elman (1990) and (1991) and Sejnowski (1987) show that weight space is indeed partitioned after learning. For example, Elman (1990) has used clustering to reveal a partitioning of state space which corresponds to lexical and grammatical categories that are learned in the course of doing a word-prediction task. He has also used PCA to find encodings of distinctions between verbs and nouns, marking number of main clause subject, and other "internal representations" (Elman 1991, 13). Further, we can give a philosophical account of the content carried in the weights of a trained network by pointing out that when a learning history selects a weight state **W**, this determines a function for that state, thereby satisfying the conditions for **W** to carry content about its causes, and thus implementing an implicit belief: “As a result … **W** becomes a stable, enduring part of an agent, carries content, and acts as a background condition allowing behavior to be caused under the right external conditions.”(Skokowski 2004, 377)

The point I wish to make here is that, despite Ramsey, Stich, and Garon's claims to the contrary, it now appears that trained networks have the capacity to encode contents *both* within activations, **I** and **H**, *and* within their weights **W**. The upshot is that in a trained network we can pinpoint *discrete* states, which I have called **B**, that, in conjunction with the (trained) weight state **W**, are *semantically interpretable* and that play a *causal role* in some cognitive episodes but not others. It appears we have indeed found an excellent candidate for belief in networks.

Ramsey, Stich and Garon might counter here that though this model might work for occurrent propositions, it won't work for general propositions such as 'cats have fur.' But this would be too hasty. Rumelhart and Todd (1993) have shown how general propositions can be encoded by networks through a training history. Thus their model is similar in spirit to Ramsey, Stich and Garon's Networks A and B. General propositions such as 'a robin is a bird' were used to train a network such that given 'Robin is a' as an input, the network produced 'bird' for output. There is no general proposition 'a robin is a bird' stored anywhere in the network. Instead, it is the connection weights **W** which play the causal role in producing the output. As in the case described immediately above, an implicit belief can be ascribed to the network, as a learning history has ascribed a function to the weights; in this case, the function to produce a general proposition, such as 'a robin is a bird', or 'kangaroos are marsupials', given appropriate inputs. Counter to the objection, then, this model will also work for general propositions, where – as above - a (trained) weight state **W**, is *semantically interpretable* and play a *causal role* in some cognitive episodes but not others.

6. Missing the Point: The Requirements of Wetware

We can now return to the literature on Ramsey, Stich and Garon mentioned at the beginning of this article and see why the various approaches fall short. O'Brien (1991) holds for a broad causal holism for belief in connectionist networks, a holism that denies what he calls “causal discreteness.” For O’Brien, causal discreteness requires that distinct causal roles can be determined for the various representational elements in a system; and this, he claims, is not plausible for connectionist networks. But the approach offered in this paper does supply distinct causal roles for the representational elements **B** and **W**. In addition, the contents for these states can be determined: explicit occurrent contents for states like **B**, and implicit enduring contents for states like **W**. Both kinds of states are required for particular actions of a network which we ascribe ‘beliefs’ to (Skokowski 2004).

Stich’s (1991) reply to O’Brien denies the broad causal holism endorsed by O’Brien, arguing that not every belief encoded in a distributed system plays a role in every episode of processing, and I agree with this for the reasons just mentioned: the relevant covariational and encoded contents for particular states **B** and **W** are what are important for a particular episode, and it is these contents which correspond to the folk psychological explanations.

Forster and Seidel (1994) put forward a simple network model that they claim proves propositional modularity for a concrete distributed system for both occurrent and enduring belief states, thus showing Ramsey, Stich, and Garon to be wrong about this property with regards to networks. Though sympathetic to Forster and Seidel’s goals, I must side with Ramsey’s (1994) reply: in essence, the model is too simple to be generally applicable. The encodings offered by this six-node model are localist in nature, and it is difficult to see how to generalize this simple model to massively distributed systems – including, most importantly, wetware.

Clark (1995) gives a different approach from O’Brien, Forster and Seidel by placing the representational burden of networks entirely on the shoulders of the weight matrix: “Connectionist systems thus encode knowledge as complex patterns of positive and negative weights linking up simple processing units."(1995, p. 342) As has been emphasized in this article, this approach ignores the direct covariational component of the input states, which carry informational content about external conditions. A robust biological neural model of cognition needs to include the role of the senses as contributing to occurrent beliefs, something which is lacking by placing the entire burden of content storage on enduring weight states.

Stich and Warfield's reply accuses Clark of neo-behaviorism, because, they claim, any black box would satisfy Clark's dispositional account of enduring belief encoded in a weight matrix **W**.(1995, 403) But this objection wouldn't work against an account like the one already provided for the contents of weight states (Skokowski 2004). For this account requires an internal, *physical* state **W** that is installed by learning: that is, a causally efficacious learning history. Such an approach actually fills another requirement of Stich and Warfield: that in order for connectionism to avoid eliminativism, a causal/historical account must be given when determining the contents of these states. This is precisely what is done in this model. Finally, Stich and Warfield still side with Ramsey, Stich and Garon that if connectionist theories are right then eliminativism will be right about the propositional attitudes.(Stich and Warfield 1995, 409). And that view is what the present paper has been devoted to overturning.

A common thread with all these failed approaches is their lack of attention to the requirements of wetware. An important part of biological connectionist systems is that most occurrent beliefs are hooked up with the senses, and further, that learned behavior requires the installation of neural connections through interactions with the environment. Both kinds of representational states – occurrent (activations) and connections (weights) – play a role in behavior, and both can be determined: occurrent states by being activated through covariation with the immediate environment, and connections through their actual and efficacious learning history. By understanding the origins of these states, we can see how individual combinations have a causal role in some episodes, but not in others. And this gives us confidence in ascribing belief to networks.[5]

Notes

[1] Earlier arguments that folk psychology is a theory can be found in Sellars (1956) and Churchland (1981).

[2] A similar point about the ubiquitousness of distributed representations is made in van Gelder (1991).

[3] See, for example, Dretske (1988), or Papineau (1987).

[4] Clark and Toribio (1994) offer an account of representational encoding at the level of hidden-unit activation. However, it is left undetermined in their formulation just what propositional content might be encoded by these units. This problem is overcome in the current proposal by including input activations as part of B

[5] I would like to thank Jay McClelland, Jeff Yoshimi, and Jesús Navarro for helpful comments and advice. I would also like to thank David Rumelhart for originally introducing me to the issues examined here.

References

Churchland, P.M. (1981) ‘Eliminative Materialism and the Propositional Attitudes’, *Journal of Philosophy* 78, pp. 67-90.

Clark, A. (1995) ‘Connectionist Minds’, in MacDonald and McDonald, eds., Connectionism: Debates on Psychological Explanation, Oxford: Blackwell, pp. 339-356.

Clark, A. and Toribio, J.(1994) ‘Doing Without Representing’, *Synthese* 101, pp. 401-431.

Dretske, F. (1988) *Explaining Behavior*, Cambridge: MIT Press.

Elman, J. (1991) ‘Distributed Representations, Simple Recurrent Networks and Grammatical Structure’, *Machine Learning* 7, pp. 195-224.

Elman, J. (1990) ‘Finding Structure in Time’, *Cognitive Science* 14, pp. 179-211.

Forster, M. and Seidel, E. (1994) ‘Connectionism and the fate of Folk Psychology: a reply to Ramsey, Stich and Garon’, *Philosophical Psychology* 7, pp. 437-452.

O’Brien, G. (1991) ‘Is Connectionism Common Sense?’, *Philosophical Psychology* 4, pp. 165-178.

Papineau, D. (1987) *Reality and Representation*, Oxford: Blackwell.

Ramsey, W. (1994) ‘Distributed representation and causal modularity: a rejoinder to Forster and Saidel’, Philosophical Psychology 7, pp. 453-461.

Ramsey, W., Stich, S., and Garon, J. (1991) ‘Connectionism, Eliminativism and the Future of Folk Psychology’, in Ramsey et al., eds., *Philosophy and Connectionist Theory,* Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 199-228.

Rumelhart, D.E., and Todd, P.M. (1993). Learning and connectionist representations, in D.E. Meyer and S. Kornblum (Eds.), *Attention and performance XI,*. Cambridge, MA: MIT Press/Bradford Books, pp. 3-30.

Sejnowski, T., and Rosenberg C. (1987) ‘Parallel Networks that Learn to Pronounce English Text’, *Complex Systems* 1, pp. 145-168.

Sellars, W. (1956) Empiricism and the Philosophy of Mind. In H. Feigl and M. Scriven (Eds.), *Minnesota Studies in the Philosophy of Science: Vol. 1*, Minneapolis, MN: University of Minnesota Press.

Skokowski, P. (2004) ‘Structural Content: A Naturalistic Approach to Implicit Belief’, Philosophy of Science 71, pp. 362-379.

Smolensky, P. (1988) ‘On the Proper Treatment of Connectionism’, *Behavioral and Brain Sciences* 11, pp. 1-74.

Stich, S. (1991), ‘Causal Holism and Commonsense Psychology: A Reply to O’Brien’, *Philosophical Psychology* 4, pp. 179-181.

Stich, S. and Warfield, P. (1995) ‘Reply to Clark and Smolensky: Do Connectionist Minds Have Beliefs?’, in MacDonald and McDonald, eds., Connectionism: Debates on Psychological Explanation, Oxford: Blackwell, pp. 395-411.

van Gelder, T. (1991) ‘What is the "D" in "PDP"? A Survey of the Concept of Distribution’, in Ramsey et al., eds., *Philosophy and Connectionist Theory,* Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 33-60.