Belief in Networks

Paul Skokowski
Symbolic Systems
Stanford University
paulsko@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.

An important article by Ramsey, Stich and Garon 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.

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.

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. 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 insist (1991, 204), is the claim of *Propositional Modularity*. Propositional Modularity holds that the propositional attitudes are:

- functionally discrete
- semantically interpretable, and
- 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, 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 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)

There is a lot I agree with in the account offered 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.*[1]

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.[2] 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

[1]A similar point about the ubiquitousness of distributed representations is made in van Gelder (1991).
[2]See, for example, Dretske (1988), or Papineau (1987).

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 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.

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. 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 accomodate 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. Learning, which involves actual physical encounters with the environment (or training set), is what installs a weight state **W**. 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.

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.

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**, everyone has agreed so far, carries content. **B** therefore does by default.

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 (1989) 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).

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 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. [3]

[3] I would like to thank audiences at the American Philosophical Association Eastern Division and the Society for Philosophy and Psychology Conferences for helpful comments and suggestions. Work on this paper was made possible by a McDonnell-Pew Visiting Fellowship in Philosophy at the Centre for Cognitive Neuroscience, Oxford University, and a Visiting Fellowship at St. Edmund Hall, Oxford, and I am grateful to the McDonnell-Pew Foundation, the Centre, and St. Edmund Hall for their support. Special thanks to Martin Davies.

References

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

Elman, J. (1991), Distributed Representations, Simple Recurrent Networks and Grammatical Structure, *Machine Learning*.

Elman, J. (1990), Finding Structure in Time, *Cognitive Science*, 14.

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

Ramsey, W., Stich, S., and Garon, J. (1991) Connectionism, Eliminativism and the Future of Folk Psychology, *Philosophy and Connectionist Theory*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Sejnowski, T., and Rosenberg C. (1987), Parallel Networks that Learn to Pronounce English Text, *Complex Systems*, 1.

van Gelder, T. (1991) What is the "D" in "PDP"? A Survey of the Concept of Distribution, *Philosophy and Connectionist Theory*, Lawrence Erlbaum Associates, Hillsdale, NJ.