Structural Content: A Naturalistic Approach to Implicit Belief*

Paul Skokowski†‡

---

* Forthcoming, ***Philosophy of Science***.

† Symbolic Systems and Philosophy, Stanford University, Margaret Jacks Hall, 94305-2150. paulsko@turing.stanford.edu

‡ The author would like to thank Fred Dretske, John Perry, and Martin Davies for helpful comments on an earlier draft, and would also like to thank an anonymous reviewer for helpful comments and suggestions.

Abstract

Various systems that learn are examined to show how content is carried in connections installed by a learning history. Agents do not explicitly use the content of such states in practical reasoning, yet the content plays an important role in explaining behavior, and the physical state carrying that content plays a role in causing behavior, given other occurrent beliefs and desires. This leads to an understanding of the environmental *reasons* which are the determinate content of these states, and leads to a better grasp of how representational content can be carried by systems without an explicit representation.

1. Introduction

This paper examines how a system that learns can carry content that helps explain its behavior without an explicit representation. I begin by assuming a naturalistic theory of representational content, and push the theory to see how much it can explain. It soon becomes clear that a standard indicator approach won't do for the implicit beliefs that operate in many, if not all, situations in which we apply belief/desire explanations. The model therefore needs to be extended to accommodate a new type of content for these situations.

2. Implicit Belief and Learning

We use beliefs and desires to explain actions. Causal explanations take beliefs and desires to be internal physical states which, when they occur together in the agent, cause the action. The belief and desire act jointly as a sufficient condition for causing movement that is appropriate for the agent under the circumstances. What makes the action appropriate is that the states have *content*. Why did the agent act as she did? Because she believed F and wanted R, and so did M to get R. The contents of belief-desire pairs thereby supply *reasons* for the agent behaving the way she does under conditions F and R.

Billy has learned to ride his first bike. He sees a bush in his path, and wants to stop. He grasps the brake lever and squeezes it until he stops. This seems a straightforward explanation of Billy's action by attributing beliefs and desires to him. Billy believed the bush was in front of him, wanted to stop, and this caused him to squeeze the lever. The belief and desire together caused behavior appropriate for Billy given the contents of those beliefs and desires.

But the model of explaining behavior by attributing a simple belief/desire pair breaks down in other cases. Billy goes to visit his grandmother in the country. She gets his dad’s old bike out of the garage for Billy to ride. Billy gets on the bike and starts riding in the field behind the house. He sees a bush in his path and wants to stop. Billy has the right belief and desire to cause his bike-stopping behavior. But Billy *doesn’t* stop. He tumbles headlong into the bushes. This bike has pedal brakes. It is natural to say in this situation that Billy doesn't have the belief that by pedaling backwards, he can obtain his goal of stopping the bike. If he had this belief, he would be able to stop the bike. Indeed, after a few days at grandma’s, Billy is using the brakes properly without having to think about it.

Examples like these show that content that is important to an agent’s successful behavior is often missing in explanations of that behavior. And when that content is missing for the agent, inappropriate behavior can ensue. This appears to call for some internal state which can carry the content in question, and which can play the right causal role within the agent's cognitive economy. But if this is so, how does the agent acquire the state and how does the state 'get its hand on the steering wheel' to help cause appropriate behavior under the right conditions?

Fred Dretske sketches a possible solution in his discussion of *implicit* beliefs that have contents to the effect that in conditions F producing M will bring about R. This content is not available for manipulation by the agent in reasoning tasks (Dretske 1988, 117). But an implicit belief, qua belief, is still a representation. As such, this implicit belief is a causally efficacious internal state with a content, and it has a function -- the job of helping to cause the right sorts of behavior under those conditions it carries content about. Though he refers to these implicit states as representations, Dretske does not give details of how implicit beliefs are acquired, how they acquire their contents, and what their function is.

This raises an issue about the relationship between implicit beliefs and beliefs derived from indicator states. For a regularity theory of belief like Dretske's, beliefs carry content F because in the past they were indicators of F. However, indicators are caused by the occurrent external conditions they indicate. Young animals have 'cliff' indicators, chickens have 'hawk' indicators, and infants have 'Mommy' indicators, because of regularities connecting outside conditions with a type of internal state (Dretske 1981). These indicators are promoted to beliefs through being recruited to cause appropriate behavior under these conditions.

This sort of analysis raises the question for implicit beliefs: how does the condition (in conditions F producing M will bring about R), "when a bush looms up squeeze the lever to stop the bike", cause an implicit state B in Billy's brain? Since the content is not available for manipulation, it is not a product of explicit practical reasoning. It also is not a property of some external object. Rather the condition appears more like a permanent condition *on* the world (at least a condition *believed*) by the agent. A belief with this content, then, cannot arise from the same class of regularities that support occurrences of 'normal' indicators, like cliff-indictors or hawk-indicators or mommy-indicators. An implicit belief of this nature is rather more like a background state in the agent's head, which represents something about how the world is, rather than an explicit state that causally responds to an occurrent external condition. This suggests that a standard regularity theory won't do the job for implicit beliefs -- more needs to be said about the origin of these states and how they acquire their contents.

I would like to motivate a naturalistic account of implicit belief that can clarify the contents and causal roles of such states. The basic idea will be that learning installs a state in an agent which plays the appropriate causal role, and whose function, derived from its history, determines its content. Two examples will be used to suggest how these states are installed, and how they obtain their contents. The first example is a neural network, whose transparent architecture and causal properties give us a working model for the second example, which is Long Term Potentiation (LTP) in wetware. The first will arguably have acquired its content relative to a designer, while the second will have naturalized content. Both, I submit, have implicit states that are acquired through a learning history.

I will not attempt here to account for every possible type of implicit content.[1] Instead, what is offered is an account of implicit content based on learning. This is a start, I believe, to finding states with the appropriate contents that leads to a broader understanding of implicit beliefs in general.

3. Content in a Network

Consider a neural network, and the set of events comprising its successful learning of a task over time. This set of events is a learning history that causes an internal change in the network. For a neural network this history causes set of weights (a new weight structure) to be selected and installed that allows the network to perform the task. For example, *if* we take the well-known Gorman and Sejnowski backpropagation network for distinguishing sonar signals and train it on their training set for the number of epochs as they've described, *then* it will distinguish mine from rock sonar signals. The causally efficacious part of the learning history – what is minimally sufficient to install the new weight structure **W** – is that outputs M (‘mine’ outputs) are reinforced (through backpropagation) upon inputs F (mine signals). This is a learning history of reinforcing selection of Ms upon presentations of Fs. Minimally sufficient learning histories of this type, **H**, will install structures of type **W** in such a network.[2]

Other kinds of networks which learn, exhibit similar properties - in particular more biologically plausible networks which employ local learning features based on Hebbian learning. For example, Mazzoni, Anderson and Jordan (1991) describe a recurrent network that learns a transformation of retinal and eye positions (input) into craniotopic coordinates (output). This recurrent network uses an associative reward-penalty (A_{R-P}) rule, which is more ‘biological’ than backpropagation in three important ways: (1) A_{R-P} uses only locally available information to adjust weights, making its learning Hebbian; (2) A_{R-P} uses a single scalar reinforcement signal to all network connections, as “suggested by anatomical as well as experimental studies”; and, (3) output is stochastic, producing variability in neuronal firing rate that better reflects what actually occurs in biological networks (Mazzoni, Anderson and Jordan 1991, 4436). Using a biological model of learning, this recurrent network settled into a weight state that achieved an accuracy of response that matched results of backpropagation networks trained to compute the same transformation (Mazzoni, Anderson and Jordan 1991, 4435). Recurrent networks using Hebbian learning can also implement autoassociative, or attractor networks (Rolls and Treves 1998). Learning, for these networks, involves changing weights over many cycles, which can be viewed as creating basins of attraction towards which the output patterns converge. Properties of these networks that make them more biologically plausible (in addition to using Hebbian learning) are their abilities to 'remember' and complete patterns, and in some cases to learn in a single presentation.[3] Autoassociative networks can be used to model memory processes in the hippocampus, for example (McLeod et al. 1998; Rolls and Treves 1998).

Both backpropagation and recurrent networks, therefore, learn from an environment through modification of their weights. After learning, a new weight structure has been installed that is a more or less stable property of the neural network. It won't tend to change with new incoming signals, as it is this property which now enables the network to recognize features, transform coordinates, complete patterns, and so on. The weight structure 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 that we would tend to associate with a regularity, such as the regularity which occurs between the input unit activity and the external condition it is tracking. Since this regularity is absent for the stable property **W**, should we say this structure cannot carry content?

Neural network theorist David Rumelhart does not think so. In fact, he claims that for neural networks,

> . . . all the knowledge is *in the connections*. From conventional programmable computers we are used to thinking of knowledge as being stored in the state of certain units in the system. In our systems we assume that only very short term storage can occur in the states of units; long term storage takes place in the connections among units. Indeed, it is the connections -- or perhaps the rules for forming them through experience -- which primarily differentiate one model from another. This is a profound

> difference between our approach and other more conventional approaches, for it means that almost all knowledge is implicit in the structure of the device that carries out the task rather than explicit in the states of units themselves. Knowledge is not directly accessible to interpretation by some separate processor, but it is built into the processor itself and directly determines the course of processing. It is acquired through tuning of connections as these are used in processing, rather than formulated and stored as declarative facts (Rumelhart 1986, 75).[4]

This has a familiar ring to it. The claim here is that content may be carried in the weight structure of the network. Furthermore, Rumelhart says that such content is not declarative: it is not available for reasoning or evidential assessment by the network. There is no single *state* with an explicit and accessible stored content, as there is for a line in a program on a digital computer.

Now this sounds promising, but it does not yet amount to an *explanation* of this kind of content in neural networks. This content carried 'in the connections' is a different and interesting kind of content. It is not a content associated with the occurrence of a particular regularity, as perceptual content is. The Gorman-Sejnowski (1988) network reliably produces the output 'mine' when presented with a mine signal and the output 'rock' when presented with a rock signal. When no rocks or mines are around, the network will still exemplify the property **W**, but won't produce any outputs at all. Nevertheless, without the property **W**, the network won't recognize mines or rocks. So **W** plays a role. The question is, what role?

We can begin to answer this question by asking what the weight state **W** is *for*, what job it is designed to do. Consider the Gorman-Sejnowski network. Here we can look at the goals of the designer of the network. The designer wants this network to be a mine detector. She requires a pattern of connectivity in the network that will reliably determine outputs of type M ('mine' outputs) on the output units upon presentations of type F (mine signals) to the input units. Installing this state **W**, however, requires more than the desires and goals of the designer; it requires a selection history – one in which outputs M are reinforced (e.g. through backpropagation) upon inputs F. This is the learning history **H**. Following Kitcher (1999), and Godfrey-Smith (1999) we can attribute a *function* to the structure **W**, because it does the job required and because of its selection history.

Structures like **W** are installed over time through learning, and, having relatively stable properties thereafter, are related not to occurrent external conditions, as normal indicators are, but to a *history* of conditions.[5] But when this history is taken into account, a case may be made that a regularity was indeed in operation. And the history of **W** together with its function, can be used to attribute content to it: content that the network uses implicitly to achieve successful output (action) under the right external input conditions.

Following Dretske's prescription for content, B indicates A if it meets two conditions: that As cause Bs, and that Bs generally only occur if As (Dretske 1981, 1988, 1991).[6] Further, Bs can come to *represent* As if they have the *function to indicate* As. This requires that Bs have a job to do within a representational system – that they play an executive role in causing behavior for the system under appropriate external conditions.

We can apply this theory of content to our neural network. First, a learning history **H** caused the structure **W**, a new network of connections between the input and output units. For **W** to indicate **H** now requires that only events of type **H** cause states of type **W**.[7] This argument is considered below. But further, note that the state **W** has a job to do for the neural network system, and indeed has already acquired an executive role within the system. Before learning, firing of input units indicates external objects through covariation, but does *not* cause appropriate outputs. After learning, however, the firing of input units *together with* the new structure of connections between input and output units *does* cause appropriate output, namely, reliably providing the output 'mine' when a mine signal is given to the input units. As a result of the learning history the structure **W** becomes a stable, enduring part of the network. Like the implicit beliefs we considered above, this structure acts as a background state: it is what enables outputs M to be caused whenever the network is presented with an F, without itself covarying with Fs. The structure **W** has a job to do, a function, and has obtained an executive role by helping to cause appropriate behavior under the right external conditions.

4. From History to Content

The only practical alternative to a network achieving a structure **W** which allows it to recognize F's through learning is that it randomly *started* or *landed* in that state. A network can of course land in or be in a state **W** randomly. This is not very likely: the probability of starting in such a state is almost nil, given a sufficient number of nodes and continuous values for weights between them. Call this the 'argument from large numbers'. Consider the Gorman and Sejnowski network which distinguished rock- from mine-sonar signals. It has only 22 total units with 7 hidden units. Restricting the weights between nodes to one of 10 values, this network would still have 10^{105} possible weight configurations. The probability that the network starts in one of them is $1/(10^{105})$, which is a very small number. Though there may be many weight states of type **W**, this number will not generally be anywhere *near* the order of magnitude of 10^{105}.[8] Now consider the possibilities for more nodes and continuous weight values. The possibility of randomly starting in a state of type **W** diminishes. Paul Churchland makes this point when considering the Gorman and Sejnowski network:

> It would, of course, be a miracle if the network made the desired discrimination immediately, since the connection weights that determine its transformational activity are initially set at random values (Churchland 1989, 164).

The argument from large numbers tells us that even for simple networks such as this, it is almost inconceivable that it will randomly start or land in a state **W** capable of recognizing mine signals.[9]

The argument from large numbers gives some confidence that if we observe a neural network that computes a certain function with great accuracy, then that network has, in all likelihood -- and with probability approaching 1 for large networks -- *learned* to compute that function. And if the network has learned a task, then it has undergone a history of type **H**. This gets us a long way towards saying that the weight structure of such a network carries the information that **H**.

The moral to draw from this is, I believe, that the occurrence of a structure like **W** carries information about its informational heritage. Attributing a learning history to a neural network is akin to the way we might attribute a learning history to a person. Suppose the plasma physics professor is lecturing on turbulence in highdensity plasmas. As he discusses Rayleigh-Taylor instabilities, a new transfer student's remarks on the inadequacy of the Rayleigh-Taylor model of interfluid mixing under oblique shock waves at the material interface seems precocious. Nevertheless the professor can judge from these remarks that the student has undergone the following type of history: he has studied plasma physics and non-linear turbulence theory. It would be very unlikely that he hadn't.

It may seem harder to attribute a learning history to a neural network. After all, neural networks aren't really like *us*. But now consider some surprising results from Gorman and Sejnowski. They also tested humans on the same training set of rock and mine sonar signals, and compared accuracy in categorizing signals. It turned out the accuracy of human response to the testing set was nearly identical to the performance of the neural networks (Gorman and Sejnowski, 1988). This result is perhaps not very surprising, since neural net computations purportedly mimic brain processes. But it will help make an important point about our intuitions regarding networks.

Suppose I was presented with a network that could process audible signals (I am not told what kind). I happen to have on hand a set of rock and mine sonar signals that I present one by one through a sound system to the network. Surprisingly, it nearly always said "rock" or "mine" in response to the appropriate type of signal. I appear to have two choices. I could say this machine was taught to distinguish such signals. Or I could say this machine by chance *happened* to be in that weight state. If I were skeptical about machine capabilities, the latter might seem a reasonable move.

Now suppose we take our rock/mine sonar signal testing set and set up a table downtown. We stop people at random, ask them to put on the headphones, listen to the signals, and categorize them. Suppose one of them says “rock” or “mine” correctly in almost every case in response to the appropriate type of signal. Here intuition tells me that this person has surely learned to differentiate sonar signals. Indeed, I would go to the further conclusion that either this person has been trained by the Navy as a sonar signal analyzer on board submarines, or he was part of Gorman and Sejnowski’s study on the subject of rock/mine detection by neural nets! That is, I would attribute a specific learning history regarding past presentations of sonar signals to the subject. For the odds seem improbably fantastic that a random person off the street has the specific, hardwired ability to distinguish, without prior learning, rock sonar signals from mine sonar signals.

If we are disposed to say that people have undergone specific types of learning histories based on evidence of their performing certain tasks, should we not also be willing to attribute specific types of learning histories to neural networks based on similar evidence? It appears to me that we should. In the case of neural networks these learning procedures required regular presentations along with reinforcements. And as we have seen, from both statistical and environmental arguments, these are the only kinds of procedures that will, with any regularity, produce the requisite type of state. Thus, I believe, these states indicate that they were arrived at by this type of history, just as certain states of humans appear to indicate that they were learned in a specific way.

Now we can return to, and perhaps make more sense of, the claim put forward by Rumelhart, that the “knowledge” – which we can now, I think, safely call content – acquired by neural nets through learning is “in the connections.” There is good evidence, I think, that content is carried in the connections of such a network. This content does a specific job for the network, without doing the job of 'declarative' (in Rumelhart's sense) or 'explicit' contents. The structure carrying this content plays a different causal role from explicit belief states, and explains a different aspect of behavior. The content is of a different kind; it is not the sort of content carried by a state which is the product of standard regularity, such as a perceptual belief; rather, it is content carried in virtue of the actual and efficacious causal history of the structure. This history installed the state and gave it its causal role. The structure, and its content, are products of that history. This suggests how representational content can be carried by a neural network without explicit representations of these contents within the network.

To emphasize its difference from occurrent contents I will call this content *structural* content. The next step is to examine how structural content is carried and used at an implicit level by cognitive agents. Before doing this, however, I wish to show how the notion of structural content addresses issues that are approached in different ways by other authors.

5. A Different Approach

We have seen that structural content differs in important ways from accounts of representational content given by Dretske. But this notion also differs from approaches to content from other authors. I am thinking in particular of Swoyer's (1991) notion of structural representation, Kirsh's (1990) notion of implicit content, van Gelder's (1991, 1995) notions of representationalism and anti-representationalism, and Clark and Toribio's (1993) response to anti-representationalism. I will consider these authors in order, and point out differences, as well as points of agreement, with their approaches.

Swoyer (1991) develops a separate account of what he calls structural representation, which is different in subject and spirit with what I am calling structural content. His structural representations share structure with, or embed isomorphisms of, the things they represent. My notion of structural content does not require any sharing of structure between the representational vehicle and its content. By using the word structure, I intend to capture the idea that the vehicle carrying structural content is part of the very structure of the device – in particular, it is not an occurrent property. Rather, it is an enduring property.

Kirsh (1990) puts forward a theory to distinguish explicit from implicit contents. For Kirsh, immediate access is what distinguishes the two: explicit contents are immediately accessible to the system and immediately readable by it (Kirsh 1990, 356), whereas implicit contents require further computations before they are accessible by the system. According to Kirsh's prescription, in order to utilize, access, or read an implicit content, a system must expend further effort – it must do further internal *computations*.

Structural content is not captured by this prescription. Instead, structural content is installed by learning. Once installed, structural contents do not require an expenditure of effort by the system - no internal computations are required for them to be utilized. The states carrying the content are there, at the ready, as part of the structure of the system, and are utilized immediately by the system under appropriate external stimuli. Hence, structural content cannot be implicit content under Kirsh's theory. But further, these contents can't be 'read off' by the system, as they are not available for reasoning or evidential assessment.[10] Hence, structural content can't be explicit content either, under Kirsh's theory. Structural content, I maintain, is different altogether. There is a kind of representation therefore, installed by learning, which is not accounted for in Kirsh's theory. It is, however, accounted for in the theory of structural content.

Van Gelder (1991) defends the contention that content carried in the weights of neural networks is distributed, a notion I wholeheartedly endorse. However, though van Gelder explains certain properties the contents of weight states can have (such as extendedness, robustness to damage, and superposition), he does not give any explanation for how these contents come to be acquired or installed. The notion of structural content above remedies this situation: it gives a naturalistic, causal account of distributed representations in terms of history and function.

In a later paper, van Gelder (1995) takes a different approach to representation, arguing that cognitive systems, including neural networks, may indeed be noncomputational dynamical systems. He puts forward the Watt centrifugal governor as an example of a dynamical system which appears to be describable in thoroughgoing non-representational terms.[11] I agree with van Gelder that the description of the governor in the mathematical terms of dynamical systems is a compelling one. But this example does not show how it can supplant the representational depiction of systems with perceptual abilities that also exhibit genuinely plastic behavior with respect to a changing external environment – that is, systems that *learn*.

Indeed, I would side here with Clark and Toribio (1993), who argue that certain systems belong to 'representation-hungry' domains that require internal representations.(Clark and Toribio 1993, 418) Clark and Toribio put the notion of representation on a continuum: from non-representational direct coupling of a system to an environment (such as a governor accomplishes); to modest representations that sort input patterns; and eventually all the way up the scale to full-blown systems capable of reasoning about a rich range of internal and external contents.(Clark and Toribio 1993, 426)

The notion of structural content fits decidedly in the representational portion of Clark and Toribio's continuum. For the notion finds its foundation (though not its full expression) on a Dretskean model, whereby internal indicator states get promoted to representational states – states that can control output behavior – through causal encounters with the environment, viz., learning. Systems where indicator states get promoted to representational states are 'representation-hungry', if any systems are.

There are definitely affinities, then, with Clark and Toribio, but there are also important gaps in their account which structural content can help fill. First, Clark and Toribio appear to be placing representations for networks in their continuum at the hidden-unit level (Clark and Toribio 1993, 427), which is a level of activation. But structural content is carried not at the level of hidden unit activations, but rather in the weight structure **W**. Structural content therefore provides a new representational vehicle for their representational continuum. Second, Clark and Toribio do not provide a theory of content and causal role for the representations in their continuum. This essay provides these for structural content. This is especially important because structural representations work side by side with, and play the enabling role for, other more traditionally recognized executive representations, including input representations such as perceptions, and higher level representations such as beliefs. Recall that the structural state **W** must be installed by learning in order to enable the indicator state, call it B, to cause output M. Once **W** is part of the structure of the system, it does not by itself cause output motion. But without the existence of **W** as part of the structure, the system cannot reliably act under appropriate stimuli. It is this state **W** which ultimately *enables* B to *cause* outputs M under external conditions F. Finally, structural representations have a different job to do for a system than other representations; they have a different function. The account of structural content is therefore different in vehicle, content, causal role, and function from the representations considered in Clarke and Toribio's continuum. But the natural place for structural content is squarely in the representational portion of that continuum.

Structural content differs, then, in important ways from accounts of representational content given by Swoyer, Kirsh, van Gelder, Clarke and Toribio, as well as Dretske. In the next section I examine how structural content is carried and used at an implicit level by cognitive agents. I use results from neuroscience to support the extension.[12] If learning produces results at the synaptic level as neuroscientists suggest, then I believe there will be neural correlates to the structural states we find in neural networks.

6. Synapses , Learning and Structure

There is evidence that learning situations directly cause changes in synaptic connectivity between neurons in the brain. These changes last from hours in some cases to months (and perhaps permanently thereafter). These changes are exemplified in two ways: chemical alterations in existing synaptic contacts between neurons, and the growth of new synapses on existing neurons. I consider one mechanism put forward to account for these changes, called Long Term Potentiation, or LTP (Thompson 1986; McCaugh et al., 1990; Lynch 1986; Cotman and Lynch 1989;

Buonomano and Byrne 1990; Muller et al. 2000; Ledoux 2002). LTP is an instance of Hebbian learning: the simultaneous, or near-simultaneous, firing of adjacent neurons results in strengthening of connections, or the formation of connections between those neurons. LTP apparently leads to new and stable connections in the hippocampus and cortex of learning animals (Killackey 1986; Sejnowski and Tesauro 1988; Cotman and Lynch 1989; Ledoux 2002). In addition, the importance of LTP is not lost on the neural network community. Researchers in this area recognize that LTP provides a biological motivation for using Hebbian learning rules in networks (Cleeremans and Jimenez 2002; Gardner 1993; McLeod et al. 1998; Rolls and Treves 1998; Vos 2001).

The mechanism of LTP allows the model of learning in neural networks to be extended to wetware. Here is how the model works. An animal is presented with an object of some type. This causes stimulation of sensory neurons in the brain. A motion is performed (perhaps randomly) which results in a reward or reinforcement of some sort. The motion is accompanied by the firing of motor neurons in the brain, neurons adjacent to the sensory neurons which are registering the presence of the external object. This simultaneous firing of adjacent neurons strengthens or forms, through chemical modification or actual growth from LTP, synaptic connections, and so exemplifies the process of Hebbian learning. After the simultaneous firings a new state of connectivity is installed, which I will call **W**, the same as for the neural network. The consequence is that stimulation of the sensory neurons *alone* in future encounters with objects of that type will be enough, due to the newly modified synaptic connections, to drive the motor neurons to fire, causing motions which result in a reward. In this way, an internal indicator state gets its hand on the steering wheel through the formation of a new structure of connections, **W**.

We can also extend the theory of content developed for neural networks to the neural structures **W** in wetware. We have seen that learning induces synaptic change in particular groups of neurons in the brain. Call the collection of these learning situations the complex event **H**. We can see that **H** caused, and thereby installed, **W**. The structure **W** is a new network of connections with the job of linking sensory and motor neurons, that is, **W** has acquired a function, and hence an executive role in causing behavior. Before learning, firing of sensory neurons indicates external objects, but does not cause movement. After learning, however, the firing of these neurons *together with* the new structure of connections between sensory and motor neurons *does* cause movement. Cotman and Lynch describe eyelid experiments that install neural connections:

> Recent evidence suggests that a memory trace for classical conditioning is localized in discrete areas of the brain. The most complete data are from studies on eyelid conditioning... Eyelid conditioning consists of a brief sound (a tone) followed by a puff of air to the eye. After a number of pairings of tone and air puff, the eyelid develops a learned closing response to the tone before the air puff comes. . . Rabbits and humans learn the task equally well (Cotman and Lynch 1989 219).

Put this in terms of our learning model. Suppose I am the subject of the experiment. The brief tone is the external condition. My hearing the tone is the *belief* that the tone is occurring. I have the desire to avoid eye irritation. The experiment begins. It might appear that when I hear the tone, I have the right belief and desire to close my eye and thus avoid the eye irritation which will ensue with the air puff 250 ms after the tone. Nevertheless, 250 ms after the tone occurs, my eye gets blasted by a jet of air, causing painful irritation. What went wrong?

Simply, I have not learned that when the tone goes off, by closing my eyelid I will avoid eye irritation. According to Cotman and Lynch, after many trials I will begin closing my eye in response to the tone, and before onset of the air puff. And I will continue to shut my eye to such a tone. If the tone matches that for the number 3 on a push-button phone, I will close my eye whenever I dial 3. The reason I am able to make this motion reliably after learning is that a new set of connections have been installed during learning, between sensory neurons which indicate the tone and motor neurons which can cause my eyelid to close. I have obtained an implicit belief.

Denote the collection of all the learning situations in which the tone goes off, I close my eyelid during the tone, and thus avoid eye irritation, as a history of type **H**. Then **H** is a composite event which is sufficient to cause **W**. But furthermore it is the only type of event that will install **W**. Recall the random person who could discriminate sonar signals. This discrimination was taken as evidence that he had learned the skill. Similarly, if we notice that a person always blinks his left eye to a certain tone, I think we can reasonably conclude that that response was reinforced somewhere in his past.[13] In short, **W** carries the content that a learning history occurred. This history selected the structure **W**, thereby determining its function, and thus implemented an implicit belief to the effect that when the tone goes off by closing my eye I will avoid irritation.

Consider another argument for why this behavior was learned. Imagine that my spouse becomes worried about my behavior after I return from working at the learning-psychologist’s lab. She notes that whenever I dial the phone, I blink in an obvious and peculiar manner. Alarmed, she calls the local hospital to have my head examined by a famous neurosurgeon. Imagine this surgeon can pinpoint regions in the brain where tone-sensitive groups of neurons are in the vicinity of motor neurons controlling eyelid motion. This surgeon discovers a dense matrix of connections between the two groups of neurons, and notes how he has never seen this sort of structure before, since it is not a normal or an inherited feature. How could such a structure occur, he asks himself? Through Hebbian learning, of course, where near-simultaneous firings of the tone-sensitive sensory neurons and eyelid motor neurons would cause synaptic growth between the two, thus enhancing exactly the response that so worried my wife. “Good news,” he explains to my wife who is pacing in the waiting room, “it appears an eyeblink response to a tone was learned by your husband recently. This is not surprising, given his recent environment in the laboratory. You might want to ask him whether he by chance strayed into the rabbit-learning laboratory, where the sinister psychologist, Dr. Airpuff, does his work.” Relieved, my wife departs the hospital in search of a rotary dial phone.

This example suggests, I think, that unusual structures at the neural level can carry content about the types of events that caused them. This should not be too surprising, in light of similar arguments about neural networks. Recall the argument from large numbers for neural networks. But the human brain is much more complicated than any existing network.[14] So finding a structure in the human brain that connects certain perceptual states with certain (non-reflex) motor states is even more likely to have been caused by reinforcement than a similar state in a neural network. We might call this the argument from even *larger* numbers.

7. Conclusion

The above account of how LTP can implement Hebbian learning and serve as the basis for implicit states shows how the learning model for neural networks can be extended to agents. As a result of a causal learning process, the *neural* structure **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. Since it is part of the structure of the agent, **W** does not by itself cause motion. Nevertheless, without the existence of **W** as part of its structure, the agent could not reliably act under the appropriate stimuli. In both neural networks and wetware we see that content is carried in connections installed by a learning history. I submit that this is how we acquire certain implicit beliefs. These states may be installed by learning histories in subtle ways which agents are not explicitly aware of. Agents do not explicitly use the content of such states in practical reasoning, yet the content plays an important role in explaining behavior, and the physical state carrying that content plays a role in causing behavior, given other occurrent beliefs and desires. By understanding the causal mechanisms for the creation of a state **W**, we can come to understand the environmental *reasons* which are its determinate content, and so grasp how representational content can be carried by systems without an explicit representation.

REFERENCES

Bailey, Craig, and Eric Kandel (1995), "Molecular and Structural Mechanisms Underlying Long-term Memory", in Michael Gazzaniga (ed.), *The Cognitive Neurosciences*. Cambridge, MA: MIT Press, 19-36.

Baudry, Michael, Joel Davis and Richard Thompson (eds.) (2000), *Advances in Synaptic Plasticity*, Cambridge: MIT Press.

Buller, David (ed.) (1999), *Function, Selection, and Design*. Albany: SUNY Press.

Buonomano, Dean, and John Byrne (1990), "Long-Term Synaptic Changes Produced by Cellular Analog of Classical Conditioning in *Aplysia*", *Science* 249: 420-423.

Churchland, Paul (1989), *A Neurocomputational Perspective*. Cambridge, MA: MIT Press.

Clark, Andy, and Josefa Toribio (1994), "Doing Without Representing?", *Synthese* 101: 401-431.

Cleeremans, Axel, and Luis Jimenez (2002), "Implicit Learning and Consciousness: A Graded, Dynamic Perspective", in Robert French and Axel Cleeremans (eds.), *Implicit Learning and Consciousness*. New York: Psychology Press.

Cotman, Carl, and Gary Lynch (1989), "The Neurobiology of Learning and Memory", *Cognition* 33: 211-241.

Draye, Jean-Phillipe (2001), "Recurrent Neural Networks: Properties and Models", in Mastebroek and Vos 1991, 49-74.

Dretske, Fred (1981), *Knowledge and the Flow of Information*. Cambridge: MIT Press.

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

—–, (1991), “Dretske’s Replies”, in Brian McLaughlin (ed.), *Dretske and His Critics*. Oxford: Blackwell, 180-221.

French, Robert, and Axel Cleeremans (eds.) (2002), *Implicit Learning and Consciousness*. New York: Psychology Press.

Gardner, Daniel (1993), “Static Determinants of Synaptic Strength”, in Daniel Gardner (ed.), *The Neurobiology of Neural Networks*. Cambridge: MIT Press, 21-70.

Godfrey-Smith, Peter (1999), “A Modern History Theory of Functions”, in Buller 1999, 199-220.

Goldman, Alvin (1970), *A Theory of Human Action*. Englewood Cliffs, NJ: Prentice Hall.

Gorman, R. Paul, and Terrence Sejnowski (1988), “Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets”, *Neural Networks* 1: 75-89.

Killackey, Herbert, (1986), “Morphology and Memory” in Gary Lynch (ed.), *Synapses, Circuits and the Beginnings of Memory*. Cambridge: MIT Press, 111-120.

Kim, Jaegwon (1973), “Causation, Nomic Subsumption, and the Concept of Event”, *Journal of Philosophy* 70: 217-236.

Kirsh, David (1991), “When is Information Explicitly Represented?”, in Philip Hanson (ed), *Information, Thought and Content*. Vancouver: University of British Columbia Press, 340-365.

Philip Kitcher (1999), “Function and Design”, in Buller 1999, 159-183.

Ledoux, Joseph (2002), *Synaptic Self*. New York: Viking Press.

Lynch, Gary (1986), *Synapses, Circuits and the Beginnings of Memory*. Cambridge: MIT Press.

Mastebroek, Henk, and Johan Vos (eds.) (2001), *Plausible Neural Networks for Biological Modelling*. Dordrecht: Kluwer Academic Publishers.

Mazzoni, Pietro, Richard Andersen and Michael Jordan (1991), "A More Biologically Plausible Learning Rule for Neural Networks", *Proceedings of the National Academy of Science* 88: 4433-4437.

McCaugh, James, Norman Weinberger and Gary Lynch (eds.) (1990), *Brain Organization and Memory*. Oxford: Oxford University Press.

Peter McLeod, Kim Plunkett, and Edmund Rolls (1998), *Introduction to Connectionist Modelling of Cognitive Processes*. Oxford: Oxford University Press.

Muller, Dominique, Nicolas Toni and Pierre-Alain Buchs (2000), "Long-Term Potentiation: From Molecular Mechanisms to Structural Changes", in Baudry et al. 2000, 87-102.

Ramsey, William, David Rumelhart and Stephen Stich (eds.) (1991), *Philosophy and Connectionist Theory*. NJ: Lawrence Erlbaum and Associates.

Rolls, Edmund, and Alessandro Treves (1998), *Neural Networks and Brain Function*. Oxford: Oxford University Press.

Rumelhart, David (1986), "A General Framework for Parallel Distributed Processing", in David Rumelhart and James McClelland (eds.) *Parallel Distributed Processing, Vol. 1.* Cambridge, MA: MIT Press, 45-76.

Schwartz, Daniel, Vijay Samalam, Sara Solla and John Denker (1990), "Exhaustive Learning", *Neural Computation* 2: 374-385.

Searle, John (2002), *Consciousness and Language*. New York: Cambridge University Press.

Sejnowski, Terrence and Gerald Tesauro (1988), “Building Network Learning Algorithms from Hebbian Synapses”, in McCaugh et al. 1998, 338-355.

Smolensky, Paul (1986), “Neural and Conceptual Interpretation of PDP Models”, in David Rumelhart and James McClelland (eds.) *Parallel Distributed Processing, Vol. 2.* Cambridge, MA: MIT Press, 390-431.

Swoyer, Chris (1991), “Structural Representations and Surrogative Reasoning”, *Synthese* 87: 449-508.

Thompson, Richard (1986), “The Neurobiology of Learning and Memory”, *Science*, 233: 941-947.

van Gelder, Tim (1991), “What is the “D” in “PDP”?. A Survey of the Concept of Distribution”, in William Ramsey, David Rumelhart and Stephen Stich (eds.), *Philosophy and Connectionist Theory*. Hillsdale NJ: Lawrence Erlbaum and Associates, 33-60.

–––, (1995), “What Might Cognition Be, If Not Computation?” *Journal of Philosophy*, 91: 345-381.

Vos, J. (2001), “Biological Evidence for Synapse Modification Relevant for Neural Network Modeling”, in Mastebroek and Vos 1991, 7-21.

FOOTNOTES

[1] For example, we will not consider tacit beliefs such as “people tend to vote near the surface of the earth” (Searle 2002, 196). Further, the sorts of beliefs being considered are not, as explicit beliefs are, available for reasoning or evidential assessment, nor are they intended to be the products of practical reasoning.

[2] Here **H** and **W** are types. (See Kim (1973) and Goldman (1970).) For example, varying the order of input presentations in our network or the Gorman and Sejnowski network would still result in their learning their tasks. The resulting weight matrices might have changed in their elements, but successful learning will mean the same behavioral result that the network achieves its task to similar accuracy. So a learning history of the type **H** is sufficient to cause a weight matrix of type **W**.

[3] Note that the single *input* presentation still requires many *recurrent* cycles to produce the weight space structure required for enabling the attractor basins.

[4] See also Smolensky (1986, 398) and Churchland (1989, 167).

[5] Though we have used the example of a backpropagation network, history is equally important for content in recurrent networks: “In recurrent networks, the current activation of the network can depend upon input history of the system, and not just on the current input. These models have the potential to dynamically encode, store, and retrieve information” (Draye, 2001, 51). See also McLeod et al. (1998).

[6] Again, A and B are here understood to be types. Also note that natural indicators only require local validity to be successful, since it is the local environment that has selected them for the job they do in that environment. See Dretske (1988, 1991).

[7] Note that other sorts of stable states are products of a history (not occurrent properties which only covary with another property at a time) and which carry information that a type of history caused them. The rings of a tree stump carry information about the age of the tree (Dretske 1988, 55) and are themselves a product of a history. Similarly for canyons and other geological phenomena. These are stable types of states which carry information about their historical causes – about the types of histories which caused them. This is important because it shows that a Dretskean account of indication can be extended beyond occurrent conditions in the environment to historical conditions. What these stable states lack, of course, is a function of indicating within a natural representational system, and so they cannot count as representations (beliefs) in Dretske's sense, as implicit states in humans (or perhaps even neural networks) can.

[8] Schwartz et al. (1990) made an empirical measure of weight states that compute a given function for a much smaller neural network of 4.2 million possible weight configurations. They found that only two weight states were satisfactory to compute the function. Hence the odds were 1 in 2.1 million for the network to be in a state which had the correct function.

[9] To give an idea of the magnitude of this problem, it would take the fastest computer in the world (ASCI Blue, due in 2005) 10^{91} seconds to search each point in this weight space, assuming only one operation per possible weight. The universe is of the order 10^{17} seconds old, so it would take about 10^{74} ages of the universe to search this weight space (Rex Evans, Lawrence Livermore National Laboratory, private communication).

[10] See Section 3, and Section 5, this paper.

[11] The Watt centrifugal govern regulates the speed of a steam engine by controlling a valve restricting the flow of steam. See van Gelder (1995, 349).

[12] The mechanism I examine, Long Term Potentiation (LTP), is a focus of research and debate within the neuroscience community today. Any competing learning/memory mechanism with a similar result would do for my purposes. For a competing mechanism see for example Bailey and Kandel (1995).

[13] Assume there are no human nervous disorders linking sounds and blinkings, and that humans have not evolved eye-blink responses to tones.

[14] Paul Churchland estimates that a human brain's possible weight space has a minimum of $10^{100,000,000,000,000}$ possible weight configurations. (Churchland, 1989, 190.) This only assumes 10 possible weights between neuron, instead of a smoothly varying, hence infinite, number of states. So even a very small portion of the brain will have a large number of weight states available. Connections like the one considered here will most likely be installed by learning rather than through chance.