Mental Content and the Causal History of Neural Substrates

Paul G. Skokowski
McDonnell-Pew Centre for Cognitive Neuroscience
Oxford University
Parks Road
Oxford, OX1 3PT, England,
Paul.Skokowski@psy.ox.ac.uk

Introduction

This paper argues that we can determine the contents of the neural substrates that support implicit beliefs by considering the history of the substrates and their role within the system. Both artificial and biological neural substrates are examined with the goal of showing that structures at the neural level can carry informational content about the types of events which caused them. A brief discussion of the conditions under which states carry information content is given, followed by a sketch of why neural networks carry implicit content, and finally a brief case for implicit content in wetware is made.

Information

I cannot argue the merits of different theories of information content here Therefore I will assume a correlational version of information content in order to see how far it gets us towards determining the content in neural substrates (Dretske, 1981; Dretske, 1988). Say that B carries information about A if the following two conditions are met: (1) A causes B, and (2) B occurs only if A occurs. For example, suppose we train a neural network to recognize chairs. The input layer of the network is say a 100x100 screen of nodes which represent black or white pixels from a digital camera. Then a 2-D pixel image of a chair occurs at the input layer of my network if and only if a chair is in front of the camera. Then the input layer lighting up in this pattern (2-D B/W chair image) carries information that a chair is in front of the camera. So the *input layer* of my network can carry information about its environment. An analogous case is my retina. Retinal rods and cones generally only fire with a chair-like pattern when chairs are present. Though there may be exceptions (e.g. an Ames chair) these are rare, so I will consider the above definition of information to work in most cases.

Both neural networks and people *learn* to recognize chairs. Consider the set of events comprising successful learning of a task over time. Call this set of events a learning history, or **H** for short. We can think of **H** as the cause of an internal change in the cognitive system which results in the learning. For a neural network this history causes a new set of weights **W** to be installed which allows the network to perform the task. The learning history **H** is a sufficient condition for installation of the weights **W**. For example, *if* we take the well-known Gorman and Sejnowski network for distinguishing sonar signals and train it on their training set for the number of epochs as they've described, *then* it will learn to distinguish submarine from rock sonar signals.

To pass philosophical muster, we should really think of **H** and **W** as types. For example, we could vary the order of the rock and mine signals in the above example and the network would still learn the task equally well. The resulting weight matrix might even have changed in some of its 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**. It turns out we are already very close to a theory of informational content for the trained weight state **W** in a neural network. We have met condition (1) above, namely, that things of type **H** cause things of type **W**. Now we need to show that **W**'s occur only if **H**'s occur.

Large Numbers, LTP and Content

This is done by using the argument from large numbers, which concludes that the odds against finding a neural network in a certain recognition state are very large. But the human brain is much more complicated than most networks (Churchland, 1989). 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 learning than the analogue state in a simpler neural network.

I next argue that as a result of a causal learning process, with LTP as the current candidate for synaptic modification, a *neural* structure **W** is installed in the brain, becomes a stable, permanent part of the agent, carries content, and allows outputs **M** to be caused whenever the network is presented with an ***F***. This may be how we acquire certain implicit beliefs. The states **W** may be installed by learning histories in subtle ways which agents are not explicitly aware of. These histories may be thought of as the environmental *reasons* for states like **W**.

References

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

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

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