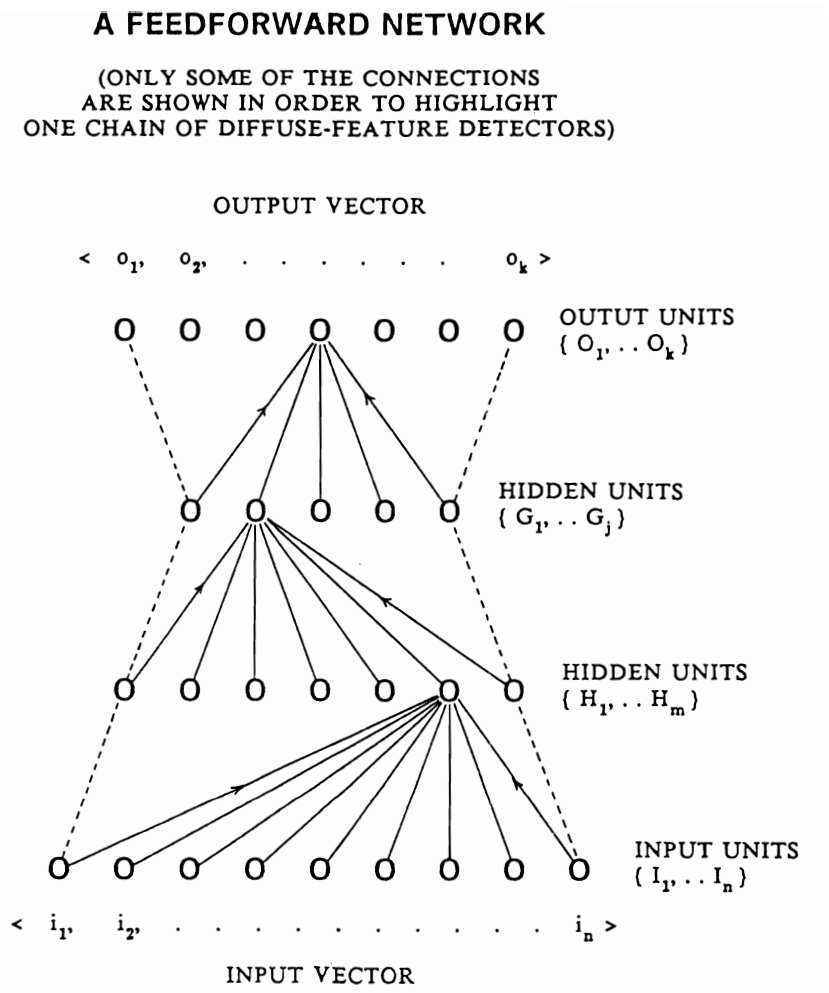
Philosophy and Phenomenological Research
Vol. LIII, No. 3, September 1993

State-Space Semantics and Meaning Holism

PAUL M. CHURCHLAND
University of California, San Diego

Fodor and Lepore (1992) recognize the state-space kinematics of neural networks for what it most assuredly is: a natural home for holistic accounts of meaning and of cognitive significance generally. Precisely what form such accounts should take is still to be worked out, but Fodor and Lepore (hereafter, "F&L") see some early possibilities well enough to try for a preemptive attack on the entire approach. My aim here is to show that the state-space approach is both more resilient and more resourceful than their critique would suggest.

A typical neural network (see Fig. 1) consists in a population of input or "sensory" neurons $\{I_1,\ldots, I_n\}$ which project their axons forward to one or more populations of hidden or "processing" neurons $\{H_1,\ldots, H_m\}$, and $\{G_1,\ldots, G_j\}$ which project their axons forward to a final population of output or "motor" neurons $\{O_1,\ldots, O_k\}$. The network's occurrent representations consist in the several activation patterns across each of these distinct neuronal populations. For example, the network's input representation at any given moment will consist in some ordered set of activation levels, $<i_1,\ldots, i_n>$ across the input units $\{I_1,\ldots, I_n\}$. It is this particular pattern or vector, *qua* unique combination of values along each of the *n* axes, that carries the relevant information, that has the relevant "semantic content." A parallel point holds for each of the network's representations at each of the successive neuronal layers. The point of the *sequence* of distinct layers is to permit the *transformation* of input representations into a sequence of subsequent representations, and ultimately into an output vector $< o_1,\ldots, o_k>$ that drives a motor response of some kind. This transformational task is carried out at each stage by the configuration of synaptic "weights" that connect each layer of neurons to the next layer up.



F&L claim that sameness-of-content across distinct persons will require sameness of activation vectors across precisely parallel sets of *n* neurons, one set for each person, where the parallel consists in identical semantic significance for each of the corresponding elements or dimensions of the two *n*-tuples at issue. In short, one has the same meaning iff one has the same pattern over the same dimensions. This demand is easily met for simple cases such as taste coding, where all normal humans share a common system of four types of taste receptor on the tongue. Sameness of taste, whether across individuals or across times, consists in an identity of the candidate activation patterns across those four universal types of sensory neuron.

But sameness of content becomes exponentially more difficult to achieve when the dimensionality of the representations involved reaches into the millions, and where we do not have a universal one-to-one correspondence be-

tween neuronal populations across individuals. There is enormous idiosyncrasy in the number and distribution of retinal cells, for example, or cochlear cells, or somatosensory cells. And the idiosyncrasy is at least as great as we ascend the processing hierarchy into the sensory cortex and beyond. How can we get sameness of vectors if we cannot even hope to get sameness of constituting dimensions?

A second formulation of this objection asks "Where do the constituting dimensions of the relevant activation-vector space get *their* semantic content?" Here F&L charge the network approach with being committed to an intolerably strict form of Concept Empiricism: all meaning arises from vectorial combinations of basic meanings, which basic meanings must be fixed by the causal sensitivities of the individual sensory neurons. This objection has an especial force as urged against me, since, as F&L observe, I have argued at length against such empiricist accounts of meaning (Churchland, 1979. pp. 7–41, 63–66). For me, the meaning of observation terms is typically independent of the sensory inputs that prompt their occasional application. In particular, I have argued that two creatures could share essentially the same conceptual framework even in the extreme case where they share no sensory organs in common. F&L's construal of state-space semantics would render this flatly impossible.

Thus the critique. We may begin the defense as follows. F&L persist in seeing network architectures as engaged in the classical business of assembling semantic simples into appropriate semantic complexes. But in actual brains, and in realistic networks, the functions performed are typically just the reverse of what F&L imagine. Rather than assembling complexes from simples by Boolean articulation, neural networks are struggling to recognize or to recover useful "simples" from enormous complexity, noise, and confusion at the sensory level.

Upon reflection, this fact is not hard to appreciate. When walking past a mewing kitten among the pillows on the couch, for example, one has active some 200 million retinal cells and some tens of millions of auditory cells. These cells are simultaneously responding to many millions of environmental features beyond those available from the kitten, in which larger sensory surround the "possibly kittenish" features are thoroughly buried. The activation pattern across one's peripheral sensory neurons is changing continuously as one moves through the room, as the kitten uncoils its tail and yawns, as the children chatter, as the morning news drones from the radio, and as the mottled sunlight dances on the couch because of nodding tree branches outside the window. A second person, Patricia, standing behind the couch, brings a different set of sensory neurons to the situation, and confronts her own unique cacophony of neural stimulations.

And yet, despite the overwhelming sensory complexity within us and sensory diversity across us, both of us manage to recognize the presence of the

kitten on the couch, swiftly and effortlessly. The marvelous features of vector coding and nonlinear parallel processing allow us to explain how this is possible. Prior training has shaped the activational tendencies of one's higher populations of hidden units into a finite family of practically relevant prototype activation patterns: for couches and kittens, for voices and meows, for pillows and pets, and for all of the other categories that make up one's conceptual framework. The tendency to fall into just these categorial patterns has been created during the learned configuration of the myriad synaptic connections that allow activations at one neuronal level to produce different activation patterns at the next level up. This training process also turns the neurons at lower levels into detectors of subtle activation patterns or sub-patterns at still earlier levels, and ultimately at the sensory level.

Experience with artificial networks has taught us, however, that the many interlocking sub-features detected at lower levels—whose collective impact at higher levels ultimately produces the activation vector for "kitten"—are rarely the features that one would list in a definition of "kitten" (e.g., small, furry, four-legged, young, feline). More typically they will be diffuse, opaque, inarticulable features whose only obvious significance is that, in conjunction with hundreds of other similarly diffuse coding features passing through the labyrinth of the massively parallel network, they occasionally participate in the selective activation of the higher-level "kitten" vector. This comparative "semantic opacity" of the computational process reflects in part the difficulty of the processing task. But it also reflects the fact that multi-layered nonlinear neural networks are typically *not* computing mere Boolean combinations among their sensory inputs. Such networks can approximate the computation of any computable function, including highly esoteric functions. This feature is essential to their celebrated successes.

In the event, within both Patricia and me there is activated, at some fairly high-level population of hidden units, a vector that represents kittens, that has the content "kitten." What makes each vector, Patricia's and mine, a "kitten" vector is not the identity of our respective patterns of neuronal activation across the hidden layer (these are likely quite different), nor the semantic identity of the constituting dimensions in her hidden-unit population and in mine (their diffuse "contents" may well be quite idiosyncratic). What gives this vector the content "kitten" is the overall role that this vector plays in the larger cognitive and motor economy of which it is an interlocking part. Thanks to an almost magical process of sensory filtering and vector completion, that prototype vector is typically activated in me by any one of a wide range of possible sensory impacts on me by a kitten. But much more importantly, that prototype vector has rich causal and computational *sequelae* of its own. Because of its own transformational significance to the extensive and well-trained network *downstream* from that prototype vector, it prompts a

family of kitten-specific perceptual expectations and prepares one to engage in a family of kitten-specific behaviors.

It is this downstream aspect of the vector's computational role that is so vitally important for reckoning sameness of cognitive content across individuals, or across cultures. A person or culture that discriminated kittens reliably enough from the environment, but treated them in absolutely every respect as a variant form of *wharf-rat*, must be ascribed some conception of 'kitten' importantly different from our own. On the other hand, an alien person or species whose expectations of and behavior towards kittens precisely mirror our own must be ascribed the same concept "kitten," even though they might discriminate kittens principally by means of alien olfaction and high-frequency sonar beamed from their foreheads.

One of the great virtues of neural networks is that they can overcome the inevitable chaos, complexity, noise, and perspectival variety at the sensory periphery in such a way as to activate comparatively well-behaved and dynamically salient categories at higher levels of processing. If we wish to understand the significance—the meaning, the content—of those prototypical categories, the most revealing place to look is at their computational role within the overall cognitive economy and motor behavior of the creature at issue. This is why neural network researchers so often find it useful, in sleuthing out the cognitive strategy of some successful network, to examine the set of partitions that training has produced across the activation space of each of its distinct hidden layers, to examine their relations with adjacent partitions, to explore their causal interactions with the partitions across earlier and later layers, to examine the "receptive fields" of individual neurons, and their "projective fields" as well. We do this with artificial networks and real neural networks alike, and with increasing success. (See for example Rosenberg and Sejnowski, 1987; Lehky and Sejnowski, 1988, 1990; Gorman and Sejnowski, 1988.)

This returns us to a robust and recognizable form of meaning holism: it is conceptual role that counts. What is novel in the state-space or vector-coding approach is the fresh account it provides of what our cognitive economy actually consists in. Instead of a rule-governed dance of propositional attitudes, we are presented with high-dimensional activation vectors being transformed into new vectors by virtue of passing through a series of well-trained matrices of synaptic connections. If we hope to solve the problem of sameness-of-meaning by exploring relevant similarities across distinct cognitive economies, we now have a new universe of possibilities to explore. We know that the vector-processing approach addresses the actual microstructure of the brain. We know that artificial networks display some intriguing cognitive properties, including the development and deployment of hierarchical categorial systems. Finally, to make a negative point and to readdress Fodor and Lepore, we know that the vector-processing approach is certainly not com-

mitted to Concept Empiricism, nor to a misfocussed and unattainably fine-grained criterion for category-identity across similar but structurally idiosyncratic networks. The vector-processing approach is still a live candidate, and easily the best hope for holism.

REFERENCES

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

Fodor, J. and Lepore, E. (1992), *Holism: A Shopper's Guide* (London: Blackwell): chapter 7.

Rosenberg, C. R. and Sejnowski, T. J. (1987), "Parallel networks that learn to pronounce English text," *Complex Systems* 1: 145–68.

Lehky, S. and Sejnowski, T. J. (1988), "Network Model of Shape-from-Shading: Neuronal Function Arises from Both Receptive and Projective Fields," *Nature* 333: 452–54.

_____ (1990), "Neural Network Model of Visual Cortex for Determining Surface Curvature from Images of Shaded Surfaces," *Proceedings of the Royal Society of London* B240: 251–78.

Gorman, R. P. and Sejnowski, T. J. (1988), "Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets," *Neural Networks* 1: 75–89.