ESSAY

Author(s): Patricia Smith Churchland

Source: *Scientific American*, Vol. 261, No. 1 (JULY 1989), p. 118

Published by: Scientific American, a division of Nature America, Inc.

Stable URL: https://www.jstor.org/stable/10.2307/24987335

ESSAY

From Descartes to neural networks

by Patricia Smith Churchland

Reflex behavior, Descartes reasoned, could be explained in terms of the physical connections of nerves and muscles. But for nonreflexive behavior, behavior produced by intelligence and planning, he could not imagine a physical explanation. In his famous summation "I think, therefore I am," the "I" was considered a nonphysical thing, a soul existing independently of the physical brain but in some kind of causal interaction with it. The Cartesian idea of two very different substances—the mental and the physical—has enduring appeal. The subjective phenomena of which we are introspectively aware when we see or think seem to be entirely different from the physical activity of brain cells.

Yet how things seem is often not how science discovers them to be. The earth does in fact move even though it seems not to; "solid" matter is in fact mostly empty space; living things were not created but evolved from simpler structures. Descartes and introspection notwithstanding, those of us who are physicalists expect that by discovering the nature of the brain and the principles governing how it works we can understand perception, learning and other "mental" functions in neurobiological terms.

How fares the physicalist program? Within this decade progress in neuroscience has been spectacular, and a great deal is now known about the properties of neurons and about the complex molecules that affect the responses of neurons. At the same time experimental psychology and clinical neurology have yielded behavioral data that reveal much about the scope and character of such psychological capacities as visual perception and memory. These data are essential if neuroscientists are to know exactly what the functions are for which they are seeking mechanisms.

If we already understand much about the nature of cells and about the general character of psychological capacities, is that not sufficient to explain how we learn, see or talk? No: necessary these data surely are, but sufficient, alas, they cannot be. The reason is that the brain is a kind of computer, and to understand how the brain works we need also to understand the computational principles that nervous systems employ. To be sure, neurons are the basic units, but they interconnect to form networks and systems. If we are to understand how the brain enables us to see and learn, we must understand how networks of cells interact to represent, transform and store information.

The hard part is to figure out what kind of computer the brain is. The problem would be easier if the brain resembled the familiar von Neumann computers, with their serial, digital architecture, unmodifiable connections and memory banks. The resemblance is very feeble, however; the brain seems to be a computer with a radically different style. For example, the brain changes as it learns, it appears to store and process information in the same places, its elements are analog rather than digital and it is comparatively fault-tolerant. Most obviously, the brain is a parallel machine, in which many interactions occur at the same time in many different channels. Moreover, natural selection being what it is, there is a premium on solutions that are fast and approximate rather than slow but exact.

The dramatic conceptual breakthrough has been the invention of computer models that to a first approximation are brainlike. Neural-network models (also called connectionist models or parallel distributed-processing models) attempt to capture, at some appropriate level of abstraction, the computational principles governing networks of neurons in nervous systems. Typically the models have neuronlike units, axonlike lines connecting the units, and modifiable synapselike weights on the connection lines.

A major discovery has been that these model networks can learn. Rather like organisms with a nervous system, they can extract commonalities from examples and generalize to new cases. The key to their learning is that their synapselike weights can be modified incrementally so that the answer the network gives to a question gets closer and closer to the correct answer. How do the weights know in what direction and by how much to change? Various algorithms have been devised to adjust the weights of synapses—as a function of error on the previous input-output trial, for example.

By means of such algorithms a network can be trained—as opposed to simply being programmed—to perform surprisingly complex tasks. For example, it can be trained to distinguish the sonar echoes of rocks from those of metal objects. It is technologically important that simple network systems can thus learn to solve problems of such complexity; the finding is also illuminating theoretically, because real nervous systems too must learn by using various strategies for the modification of synapses. Just what algorithms the brain actually uses for synapse modification is not known, but the hope is that convergent research from neuroscience and from network modeling will be able to discover them.

Other seminal ideas emerging from neural-network modeling have provided insight into what the brain could be doing. The general problem of the nature of computations and representations in nervous systems is now more approachable because cognitive representations in model networks simply *are* patterns of activity across a large population of units; computations are synapse-mediated transformations from one pattern to another. All of this means that representations must typically be distributed across large neuronal populations rather than being assigned to individual "grandmother" cells. Motor control is likewise distributed rather than emanating from "command" neurons.

Although the modeling of nervous systems is still in its infancy and we do not yet have any model that exactly explains how we see or learn, there is a gathering conviction that current lines of research, like the network models themselves, are converging on answers. Progress so far does provoke educated speculation about the neurobiological basis of our mental lives. Descartes' was a 17th-century vision: pre-Darwin, pre-Turing and pre-neuroscience. Informed by neuroscience and computer science, we can modernize his vision and begin to discern the shape of a new theory about the nature of the mind—of what it is for the physical brain to see, learn, and understand itself; of what it is to be a human being.

PATRICIA SMITH CHURCHLAND, professor of philosophy at the University of California, San Diego, is the author of *Neurophilosophy: Toward a Unified Science of the Mind-Brain.*