Behavior and Philosophy, *38*, 103-111 (2010).

RESPONSE TO DONAHOE REVIEW

Charles Randy Gallistel
Rutgers University

I thank Donahoe for a thoughtful and thorough review that fairly describes the book's core arguments. Donahoe's criticisms and comments provide an opportunity to amplify on some of our key arguments in ways that I hope will contribute to a fuller understanding of them.

As Donahoe points out, Shannon's theory of communication, from which the modern mathematical definition of information comes, is central to our argument. The one-sentence essence of our argument is that: "The function of memory is to carry information forward in time in a computationally accessible form." For this argument to have substance, it must be clear what information is; hence, that is where our book begins.

Information theory assumes that the receiver of a communication (here, the brain) has, *ab initio*, before receiving any messages, both a representation of the set of possible messages and a probability distribution on that set. Donahoe's objection to this assumption reveals a common misunderstanding of what it means to have "foreknowledge of all the events that could possibly occur" and what it means to have a prior probability distribution over a set of possible messages denoting possible states of the world. Donahoe's objection is that "When a state is encountered for the first time, neither its prior existence (by definition) nor its *a priori* probability could be known."

First, consider what it means to assume that a system that is to receive messages conveying information about states of some aspect of the world has prior knowledge of those states. For the sake of concreteness, let us focus on what it means to say that an animal with color vision has a representation of possible reflectance spectra before it opens it eyes. What this means is that it has input-processing machinery that can generate different signals for different spectra (of equal photopic luminosity). From an evolutionary perspective, the existence of this genetically specified machinery implies (or suggests, if you like) the existence of differences in spectra in the world that the animal is likely to live in. If the sun were a monochromatic source of light, we would not expect to find visual systems with structures that enabled them to distinguish spectra. In many cave-dwelling species, the visual system has atrophied altogether. Where there are no light-conveyed messages to be received, there is no point in having machinery designed to pick up light. Where there are no spectral differences, machinery designed to

AUTHOR'S NOTE: Please address correspondence to the author at:
galliste@ruccs.rutgers.edu

detect them has no function. Thus, if we believe that the structure of animals' sensory/perceptual systems reflects the structure of the world in which they usually function, the existence of machinery capable of encoding differences along a dimension of the experienced world constitutes knowledge that this dimension of experience commonly exists. This is what we mean when we say that the (animal) receiver has foreknowledge of the possible messages: it has genetically specified structures designed to generate messages along a dimension of possible experience.

It is important to distinguish between the possible *messages* and the possible *states of the world*. The set of possible messages about spectra (within the range visible to humans) is infinitely smaller than the set of possible spectra because infinitely many different spectra (the set of metameres) will generate any given message in the brain's 3D color-signaling system. Moreover, the existence of a *possible* message (a message that *could* be generated by the sensory/perceptual machinery) does not guarantee the existence of a corresponding state of the world. If an animal with color vision is raised in an environment with monochromatic illumination, the messages that its visual system *could* generate regarding spectral differences will not *be* generated because the experienced environment will contain no chromatic contrasts.

When these two points are understood—1) that to endow a system with machinery for encoding distinctions along a dimension of experience is to endow it with foreknowledge that this dimension exists within environments the animal is likely to experience; and 2) that the set of possible messages regarding some dimension of experience is distinct from, and not equinumerous with, the set of possible differences in that aspect of the world—then I believe it is no longer self-evident that animals cannot be assumed to have foreknowledge of possible states of the world. On the contrary, it seems to me that it is self-evident that they do have extensive foreknowledge of just this kind. Modern neuroscience reveals more and more purpose-specific sensory/perceptual machinery. This machinery seems to be structured so as to generate different messages corresponding to different states of the world along a restricted dimension of experience. The many areas into which the cortex is subdivided embody in their purpose-specific structure foreknowledge of possible messages. In my reading of the literature, most of this structural differentiation of the computational machinery has a genetic foundation. Experience contributes to the development of these structures in ways that seem to be anticipated by the genetic program.

In short, as regards foreknowledge of experiential possibilities, the brain has machinery for extracting color (spectral information) from visual input, for signaling color (transmitting spectral information from place to place), and for remembering color (transmitting spectral information gleaned from past experience forward in time for possible use in later behavior). This machinery gives it foreknowledge that color is a possible aspect of our experience. The same argument applies to all of the other dimensions of experience that we are capable of experiencing. Perhaps there is a dimension of the world that we are capable of

experiencing in the absence of purpose-specific sensory/perceptual machinery that makes the experience possible, but if there is, I have no idea what it is.

Second, consider what it means to say that the brain has a *prior* probability distribution defined over a dimension of possible experience (a set of possible messages). Donahoe, in common with many others, erroneously assumes that a prior must give the *true* probabilities. A prior is simply the initial probability distribution *that is updated to better reflect the true distribution revealed by experience*! One of the many seminal insights in Shannon's paper, which we try to make clear in our book, is that coding schemes (schemes for mapping states of the world to the messages that convey those states of the world from one place to another within a brain or from one time to a later time) constitute representations of the probabilities (relative frequencies) of the different messages conveyed. Thus, if a brain-signaling system adjusts its encoding of different states of the world in the light of its experience of their relative frequency, so as to make its encoding more efficient, it is updating a probability representation defined over the set of possible messages. The initial coding was its prior. (To make an encoding more efficient would mean, for example, to convey the more frequent messages with fewer spikes and the less frequent messages with more spikes.)

It does not seem to me self-evident that the brain has a way of representing and updating probabilities. However, there is behavioral evidence in favor of this assumption (e.g., Balci, et. al., 2009), and it is an increasingly common assumption in neural net modeling (Knill & Pouget, 2004). Once it is understood that the assumption of a prior means simply that probabilities are attributed to possible messages from the outset *in order that they may be revised in the light of experience*, then it is by no means self-evident that this makes information theory and Bayesian updating inappropriate models for what is going on in the brain. The inexperienced brain may have a flat prior or a Jeffreys prior—prior probability distributions that represent the state of a receiver's knowledge when it has had no experience (no data). It is simply an error to imagine that the priors in communication theory must come from a theoretician's knowledge of the world. Because priors in information theory and in Bayesian inference are an analytic necessity, much mathematical thought has gone into the formulation of priors that are appropriate when no data have yet been seen and when analytic constraints on what the data might show are minimal.

Our book was written in part in the hope that it would make young neuroscientists better informed about information theory and how it applies to understanding the brain. Far from being an old idea from the 1960s now known to be irrelevant to neuroscience, information theory already plays an important role in neuroscience (Rieke, et. al., 1997). We believe it will play a much more important role in the future. The young neuroscientist ignores it at his or her peril.

I agree with Donahoe that "Inferences based on behavioral observations about the cognitive processes that are said to underlie behavior have proven notoriously difficult to sustain. . ." However, there would seem to be no way to avoid working with such inferences—neither in neuroscience nor in other sciences. In

neuroscience we need to have an idea of what to look for in the brain when we look for the physical basis of memory. That is the question our book addresses.

The inference that associative connections form in neural tissue and that these connections are the physical realization of memory is a conspicuous example of the profound influence that a psychological theory with roots in ancient philosophical speculation has had on contemporary neuroscience. The inference derives from the theory that learning is a process of association formation. This psychological theory has guided neurobiological inquiry into the mechanism of memory for more than a century. Our book questions its soundness. We suggest that it is destined to end up in the historical dustbin along with such distinguished scientific predecessors as phlogiston, the luminiferous ether, and the Ptolemaic epicycles, all of which were regarded as soundly established in their day. We elaborate at length our reasons for doubting the inference that changes in synaptic conductance are the basis of memory.

If this inference from psychological theory is wrong, then neuroscientists searching for the physical basis of memory are barking up the wrong tree, and they have been doing so for a century. They are searching for neurons that wire together when they fire together. That is the only kind of mechanism they are prepared to consider, because they have all made the mistake of believing what they learned about learning and memory in introductory psychology. What they should be searching for is a mechanism that can encode the facts revealed by experience and carry them forward in time, until such time as they become useful in the computations that mediate the generation of adaptive behavior. The sooner we stop the fruitless search for associations in the brain and start looking for a mechanism that can encode a fact, the sooner we will reach the goal we all want to reach: knowledge of the physical basis of memory in neural tissue.

Turning now to read/write memory, which is what our book is most deeply about.

It is inaccurate to say that "each stored memory can be retrieved because it has an address (a location in the brain) and that location also contains the address of a related memory." A memory location contains *either* information that can be translated directly into behavior (the value of a variable) *or* the address of another memory location, but *not both*. A memory location in a conventional computer has two elements: 1) an address, which enables the information stored at that location to be found and read (transcribed, activated), and 2) a coding portion that carries the stored information forward in time. The information carried by the coding portion may be either the value of a variable or the address of another memory location. When address information is stored in a memory location, it conveys information derived from experience only indirectly. As we explain at length, storing addresses in the coding portions of a memory location makes possible hierarchical data structures—both in computers and in the genome. These data structures specify the relations between variables.

Biologists can perhaps best appreciate the functional distinction between the coding portion of a memory location and the address of that location by considering the difference between the coding portion of a gene and its promoter.

The coding portion of a gene encodes the sequence of amino acids in a protein. The promoter contains base pair sequences to which one or more transcription factors may bind. The promoter enables the cellular machinery to find and activate the information carried by the coding portion. Thus, it functions as the address of a gene. The coding region is transcribed. The promoter region is not. The address of a memory location in a read/write memory enables the computing system to find and transcribe (read) the information conveyed by the coding elements at that location. The address itself cannot be transcribed (read). It can, however, be stored in the coding portion of *another* memory location. Then, of course, it can be transcribed (read).

What Donahoe calls a write-only memory "in which the products of experience are stored and then later retrieved if the contemporaneous environment contains events that were present when the memory was originally stored" is, as best I can make out, what a computer scientist would call a content-addressable memory. He refers to this as a *finite-state memory*. This term has no meaning in computer science; all physically realizable memory is finite. On the other hand, the concept of a finite state *automaton* (not a finite state *memory*) is of fundamental importance in computer science. We elaborate on it at length in our book. Finite-state automata do not have a read/write memory. They are mathematically equivalent to a Turing machine that cannot read the symbols it has written to its tape. From that perspective, Donahoe's terminology makes sense. However, if a machine truly cannot read what it has written, then there is no point in its having written it, because what has been written cannot influence its behavior in any way. That is clearly not what Donahoe understands by a write-only memory. I proceed on the assumption that he agrees that the current neurobiological story treats the brain as a finite-state automaton.

A finite-state automaton has a memory for past inputs only insofar as its input history may be deduced from its current state. Different input sequences *may* lead to different states of the machine. Therefore, an observer of the machine's state *may*, in exceptional, generally very simple circumstances be able to recover the input. In general, however, an observer of the machine's state cannot reconstruct the input that produced that state because many different input sequences produce the same state. The mapping from input sequences to states of the machine is many-to-one; hence, the mapping is not invertible: there is no way to recover the experienced state of the world from the state of the machine. A machine with symbolic memory, by contrast, encodes its input sequences into memory, from which it may later read them back. An observer who knows the encoding scheme can recover the input sequence from the state of the symbolic memory. The state of the memory is invertible (readable); the state of the processor (the finite-state automaton) is not.

The distinction between a brain that is a memoryless finite-state automaton, on the one hand, and a brain with symbolic read/write memory, on the other, is crucial to our argument for two reasons. First, a finite-state automaton is provably less powerful than a computing machine with symbolic memory (mathematically, a Turing machine). A machine endowed with symbolic memory can compute

many things that a finite-state automaton cannot—things that even insect brains do compute (such as the courses between familiar locations). Second, associative theories of learning assume that the brain is a finite-state automaton. Thus, they generally have the just-mentioned property: an observer of the state of the associations cannot recover the experiences that produced them. In neurobiological terms, from the current wiring diagram, one cannot deduce the experiences that produced it. That is what it means to say that associations do not encode facts drawn from experience. And that is why associative theories have traditionally been anti-representational; they have been offered as alternatives to theories in which the brain is assumed to encode its experience (Hull, 1930; Rumelhart & McClelland, 1986; Skinner, 1938). As Donahoe says, "the units in neural networks are sub-symbolic with behavior emerging from which symbols are sometimes inferred." In other words, the behavior looks *as if* it has symbolic underpinnings, but it really does not. Those who are committed to an associative theory of learning and what that theory seems to imply about the neurobiology of memory have been making this "*as if*" argument since time out of mind because the association is not a symbol; thus, it cannot encode a fact.

In Donahoe's concluding speculations about how jays are able to find their way back to thousands of food caches spread over hundreds of square kilometers, readers get an example of the lengths that associative theorists will go to in attempting to explain away the simple fact that jays remember the locations of their caches and can therefore be inferred to have a memory mechanism capable of encoding such a fact. Nothing I might say will add or detract from the plausibility of his account in the eyes of any given reader. Readers not familiar with the wonderful work on jay cache memory, reviewed at length in our book, should be aware, however, of what Donahoe does not attempt to explain: The jays remember not only where their caches are but also which kind of food each cache contains, how long that kind of food takes to rot, when the cache was made, who was watching, and whether a cache has or has not already been harvested or moved (Clayton, et. al., 2006). Animals, both human and not, remember facts. Neurobiologists' current story about memory has no explanation for that fact, which is what leads to accounts like those proffered by Donahoe.

Donahoe is correct that "the difficulty [with our argument] is that current neuroscience provides no support for [the existence of a read/write symbolic memory capable of storing facts in hierarchically organized form]". The question is, whose problem is this? Is this a problem for cognitive science, where a symbolic memory that records facts about the world is routinely assumed? Or is this a problem for neuroscience, which is not looking for—and has not found—a neurobiological mechanism capable of encoding a fact? We argue that it is a problem for neuroscience because there is compelling behavioral evidence that brains remember facts extracted from their experience, and because computer science enables us to deduce with some confidence the implications of that well-established fact. That is why we subtitle our book "Why Cognitive Science Will Transform Neuroscience." Sooner or later neuroscience will have to come to terms with the fact that animals remember facts. When it does, it will be transformed,

much as biology was transformed when it was realized that (some) molecules are capable of encoding information in readable form.

Lest this controversial contention seem not to pay due deference to neuroscience, we point out that this situation has arisen many times in the past. A salient example is the concept of a gene, which was central to classical genetics, but for which no plausible biochemical model had been suggested prior to Watson and Crick's discovery of the structure of the DNA molecule. Before Watson and Crick, many biochemists did not believe in the existence of genes because their putative properties were biochemically inexplicable (Judson, 1980). Moreover, the notion of an encoding did not exist is biochemistry prior to 1953, just as the notion of the encoding of information gleaned from experience barely exists in contemporary neuroscience. However, it was more or less immediately obvious to anyone who contemplated the structure proposed by Watson and Crick for the DNA molecule that it offered solutions to the two things that had, until then, made the gene a biochemical mystery: 1) how a molecule could make a copy of itself (or direct the making of a copy), and 2) how it could encode the structure of a chemically unrelated compound. Fortunately, classical geneticists did not pay the deference to biochemistry that Donahoe believes cognitive science ought pay to neuroscience.

Another instructive example is Darwin's sticking to his theory even though Lord Kelvin and P. G. Tait—among the most eminent physicists of the day—asserted that it was impossible for the Earth to be more than 10–40 million years old. They confidently and repeatedly asserted that no heat-generating process allowed an age greater than that, and they emphasized that this upper limit on the age of the Earth did not allow time for the processes that Darwin assumed. Kelvin and Tait dismissed Darwin's arguments and the geological evidence he cited, just as some contemporary neuroscientists dismiss any evidence from psychology and cognitive science that does not fit their unshakable convictions about how the brain works. Kelvin and Tait knew nothing of radioactivity. They mistakenly believed that physicists knew enough to list the conceivable sources of the sun's heat. The unanticipated discovery of radioactivity and nuclear heat generation came late in their long careers. Even then, they refused to acknowledge that the discovery of this unprecedentedly powerful heat-generating process trashed their argument with Darwin. They could not first imagine, nor later admit, that biology and geology had grasped a scientific truth—the vast age of the Earth—that revealed a profound lacuna in what physicists in the late 19th century understood about the origins of the sun's heat.

Whether these examples from the history of science are relevant to the present situation in cognitive science and neurobiology remains to be seen. They do, however, emphasize that our inability to cite a neurobiological mechanism for the read/write memory, which we argue must nonetheless be there, should not be taken as a fatal objection to our argument. It is an extensively documented scientific fact that brains remember facts such as the durations of intervals, the appearances of scenes, the colors of flowers, the sweetness of solutions, the locations of caches, the contents of caches, the social identities of their conspecifics, the distances

between locations, the times of day at which things happen, etc ad infinitum (see Gallistel, 1990). The current neurobiological story about memory (a change in synaptic conductance consequent upon repeated nearly simultaneous pre- and postsynaptic activation) is unable to explain this simple fact, as Donahoe more or less admits: ". . .biobehavioral science must ultimately find a place for [the learning phenomena that we feature]."

Donahoe regards the well-documented instances of learning that are central to our argument and much closer to the layman's notion of learning (learning facts, not associations) as "not. . .paradigmatic." Our examples of learning are not paradigmatic only because they do not lend themselves to explanation in associative terms. The paradigms that psychologists working on learning and memory in animals customarily work with (e.g., Pavlovian conditioning paradigms) were devised on the assumption that learning was associative. Calling *these* paradigms paradigmatic is an exercise in circular reference. Ironically, it has turned out that, even in these "paradigmatic" Pavlovian and operant paradigms, models that assume that what is learned is not associations but rather the intervals between the experimenter-programmed events—simple facts extracted from experience—are more parsimonious (have fewer free parameters) and have more explanatory power than associative models (Balsam & Gallistel, 2009; Balsam, et. al., 2010; Gallistel & Gibbon, 2000). Thus, even in what Donahoe regards as paradigmatic examples of learning and memory, the question arises: how can changes in synaptic conductance encode the durations of experienced intervals? To that question there is, at present, no answer, as we show in a chapter devoted in part to neural models of interval timing and memory.

Our argument rests largely on mathematically established truths in computer science. It is by no means clear that neuroscience provides a sounder foundation than computer science for contemporary theorizing about how the brain computes. Theoretical computer science analyzes idealized machines that compute—not just electronic machines that compute, nor just manufactured machines that compute, but any and all machines that compute. If the brain is a machine and if it computes, then the mathematically grounded insights from computer science are surely relevant to theorizing about how it does so—just as relevant as the contemporary understanding of neurobiological processes, woefully incomplete as it surely is.

References

Balci, F., Freestone, D., et. al. (2009). Risk assessment in man and mouse. *Proceedings of the National Academy of Science USA, 106*, 2459-2463.

Balsam, P., & Gallistel, C. R. (2009). Temporal maps and informativeness in associative learning. *Trends in Neurosciences, 32*, 73-78.

Balsam, P. D., Drew, M. R., et. al. (2010). Time and associative learning. *Comparative Cognition & Behavior Reviews, 5*, 1-22.

Clayton, N., Emery, N., et. al. (2006). The rationality of animal memory: Complex caching strategies of western scrub jays. In M. Nuuds & S. Hurley (Eds.), *Rational animals?* (pp. 197-216). Oxford: Oxford University Press.

Gallistel, C. R. (1990). *The organization of learning*. Cambridge, MA: Bradford Books/MIT Press.

Gallistel, C. R., & Gibbon, J. (2000). Time, rate, and conditioning. *Psychological Review*, *107*, 289-344.

Hull, C. L. (1930). Knowledge and purpose as habit mechanisms. *Psychological Review*, *37*, 511-525.

Judson, H. (1980). *The eighth day of creation*. New York: Simon & Schuster.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neuroscience*, *27*, 712-719.

Rieke, F., Warland, D., et. al. (1997). *Spikes: Exploring the neural code*. Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (Eds.) (1986). *Parallel distributed processing*. Cambridge, MA: MIT Press.

Skinner, B. F. (1938). *The behavior of organisms*. New York: Appleton-Century-Crofts.