Memory and the Computational Brain

Why Cognitive Science Will Transform Neuroscience

C. R. Gallistel and Adam Philip King

WILEY-BLACKWELL
A John Wiley & Sons, Ltd., Publication

Memory and the Computational Brain

Memory and the Computational Brain

Why Cognitive Science Will Transform Neuroscience

C. R. Gallistel and Adam Philip King

WILEY-BLACKWELL
A John Wiley & Sons, Ltd., Publication

This edition first published 2010

Blackwell Publishing was acquired by John Wiley & Sons in February 2007. Blackwell's publishing program has been merged with Wiley's global Scientific, Technical, and Medical business to form Wiley-Blackwell.

Registered Office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Editorial Offices
350 Main Street, Malden, MA 02148-5020, USA
9600 Garsington Road, Oxford, OX4 2DQ, UK
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, for customer services, and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell.

Library of Congress Cataloging-in-Publication Data
Gallistel, C. R., 1941–
Memory and the computational brain : why cognitive science will transform neuroscience / C. R. Gallistel and Adam Philip King.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-4051-2287-0 (alk. paper) — ISBN 978-1-4051-2288-7 (pbk. : alk. paper)
1. Cognitive neuroscience. 2. Cognitive science. I. King, Adam Philip. II. Title.
QP360.5G35 2009
612.8′2—dc22

2008044683

A catalogue record for this book is available from the British Library.

Set in 10/12.5pt Sabon by Graphicraft Limited, Hong Kong
Printed in Singapore

1 2010

Contents

Preface

This is a long book with a simple message: there must be an addressable read/write memory mechanism in brains that encodes information received by the brain into symbols (writes), locates the information when needed (addresses), and transports it to computational machinery that makes productive use of the information (reads).

Such a memory mechanism is indispensable in powerful computing devices, and the behavioral data imply that brains are powerful organs of computation. Computational cognitive scientists presume the existence of an addressable read/write memory mechanism, yet neuroscientists do not know of, and are not looking for, such a mechanism. The truths the cognitive scientists know about information processing, when integrated into neuroscience, will transform our understanding of how the brain works.

An example of such a transformation is the effect that the molecular identification of the gene had on biochemistry. It brought to biochemistry a new conceptual framework. The foundation for this new framework was the concept of a code written into the structure of the DNA molecule. The code concept, which had no place in the old framework, was foundational in the new one. On this foundation, there arose an entire framework in which the duplication, transcription, translation, and correction of the code were basic concepts.

As in biochemistry prior to 1953, one can search through the literature on the neurobiology of memory in vain for a discussion of the coding question: How do the changes wrought by experience in the physical structure of the memory mechanism encode information about the experience? When experience writes to memory the distance and direction of a food source from a nest or hive, how are that distance and that direction represented in the experientially altered structure of the memory mechanism? And how can that encoded information be retrieved and transcribed from that enduring structure into the transient signals that carry that same information to the computational machinery that acts on this information? The answers to these questions must be at the core of our understanding of the physical basis of memory in nervous tissue. In the voluminous contemporary literature on the neurobiology of memory, there is no discussion of these questions. We have written this book in the hope of getting the scientific community that is interested in how brains compute to focus on finding the answers to these critical questions.

In elaborating our argument, we walk the reader through the concepts at the heart of the scientific understanding of information technology. Although most students know the terminology, the level of their understanding of the conceptual framework from which it comes is often superficial. Computer scientists are, in our view, to some extent to be faulted for this state of affairs. Computer science has been central to cognitive science from the beginning, because it was through computer science that the scientific community came to understand how it was possible to physically realize computations. In our view, the basic insights taught in computer science courses on, for example, automata theory, are a more secure basis for considering what the functional architecture of a computational brain must be than are the speculations in neuroscience about how brains compute. We believe that computer science has identified the essential components of a powerful computing machine, whereas neuroscience has yet to establish an empirically secured understanding of how the brain computes. The neuroscience literature contains many conjectures about how the brain computes, but none is well established. Unfortunately, computer scientists sometimes forget what they know about the foundations of physically realizable computation when they begin to think about brains. This is particularly true within the neural network or connectionist modeling framework. The work done in that tradition pays too much attention to neuroscientific speculations about the neural mechanisms that supposedly mediate computation and not enough to well-established results in theoretical and practical computer science concerning the architecture required in a powerful computing machine, whether instantiated with silicone chips or with neurons. Connectionists draw their computational conclusions from architectural commitments, whereas computationalists draw their architectural conclusions from their computational commitments.

In the first chapter, we explicate Shannon's concept of communication and the definition of information that arises out of it. If the function of memory is to carry information forward in time, then we have to be clear about what information is. Here, as in all of our chapters on the foundational concepts in computation, we call attention to lessons of fundamental importance to understanding how brains work. One such lesson is that Shannon's conception of the communication process requires that the receiver, that is, the brain, have a representation of the set of possible messages and a probability distribution over that set. Absent such a representation, it is impossible for the world to communicate information to the brain, at least information as defined by Shannon, which is the only rigorous definition that we have and the foundation on which the immensely powerful theory of information has been built. In this same chapter, we also review Shannon's ideas about efficient codes, ideas that we believe will inform the neuroscience of the future, for reasons that we touch on repeatedly in this book.

Informative signals change the receiver's probability distribution, the probability of the different states of the world (different messages in a set of possible messages). The receiver's representation after an information-bearing signal has been received is the receiver's posterior probability distribution over the possible values of an empirical variable, such as, for example, the distance from the nest to a food source or the rate at which food has been found in a given location. This conception puts Bayes' theorem at the heart of the communication process, because it is a theorem about the normative (correct) way in which to update the receiver's representation of the probable state of the world. In Chapter 2, we take the reader through the Bayesian updating process, both because of its close connection to Shannon's conception of the communication process, and because of the ever growing role of Bayesian models in contemporary cognitive science (Chater, Tenenbaum, & Yuille, 2006). For those less mathematically inclined, Chapter 2 can be skipped or skimmed without loss of continuity.

Because communication between the brain and the world is only possible, in a rigorous sense, if the brain is assumed to have a representation of possible states of the world and their probabilities, the concept of a representation is another critical concept. Before we can explicate this concept, we have to explicate a concept on which it (and many other concepts) depends, the concept of a function. Chapter 3 explains the concept of a function, while Chapter 4 explains the concept of a representation.

Computations are the compositions of functions. A truth about functions of far-reaching significance for our understanding of the functional architecture of the brain is that functions of arbitrarily many arguments may be realized by the composition of functions that have only two arguments, but they cannot be realized by the composition of one-argument functions. The symbols that carry the two values that serve as the arguments of a two-argument function cannot occupy physically adjacent locations, generally speaking. Thus, the functional architecture of any powerful computing device, including the brain, must make provision for bringing symbols from their different locations to the machinery that effects the primitive two-argument functions, out of which the functions with many arguments are constructed by composition.

A representation with wide-ranging power requires computations, because the information the brain needs to know in order to act effectively is not explicit in the sensory signals on which it depends for its knowledge of the world. A read/write memory frees the composition of functions from the constraints of real time by making the empirically specified values for the arguments of functions available at any time, regardless of the time at which past experience specified them.

Representations are functioning homomorphisms. They require structure-preserving mappings (homomorphisms) from states of the world (the represented system) to symbols in the brain (the representing system). These mappings preserve aspects of the formal structure of the world. In a functioning homomorphism, the similarity of formal structure between symbolic processes in the representing system and aspects of the represented system is exploited by the representing system to inform the actions that it takes within the represented system. This is a fancy way of saying that the brain uses its representations to direct its actions.

Symbols are the physical stuff of computation and representation. They are the physical entities in memory that carry information forward in time. They become, either directly or by transcription into signals, the arguments of the procedures that implement functions. And they embody the results of those computations; they carry forward in explicit, computationally accessible form the information that has been extracted from transient signals by means of those computations. To achieve a physical understanding of a representational system like the brain, it is essential to understand its symbols as physical entities. Good symbols must be distinguishable, constructible, compact, and efficacious. Chapter 5 is devoted to explicating and illustrating these attributes of good symbols.

Procedures, or in more contemporary parlance algorithms, are realized through the composition of functions. We make a critical distinction between procedures implemented by means of look-up tables and what we call compact procedures. The essence of the distinction is that the specification of the physical structure of a look-up table requires more information than will ever be extracted by the use of that table. By contrast, the information required to specify the structure of a mechanism that implements a compact procedure may be hundreds of orders of magnitude less than the information that can be extracted using that mechanism. In the table-look-up realization of a function, all of the singletons, pairs, triplets, etc. of values that might ever serve as arguments are explicitly represented in the physical structure of the machinery that implements the function, as are all the values that the function could ever return. This places the table-look-up approach at the mercy of what we call the infinitude of the possible. This infinitude is merciless, a point we return to repeatedly.

By contrast, a compact procedure is a composition of functions that is guaranteed to *generate* (rather than *retrieve*, as in table look-up) the symbol for the value of an *n*-argument function, for any arguments in the domain of the function. The distinction between a look-up table and a compact generative procedure is critical for students of the functional architecture of the brain. One widely entertained functional architecture, the neural network architecture, implements arithmetic and other basic functions by table look-up of nominal symbols rather than by mechanisms that implement compact procedures on compactly encoded symbols. In Chapter 6, we review the intimate connection between compact procedures and compactly encoded symbols. A symbol is compact if its physical magnitude grows only as the logarithm of the number of distinct values that it can represent. A symbol is an encoding symbol if its structure is dictated by a coding algorithm applied to its referent.

With these many preliminaries attended to, we come in Chapter 7 to the exposition of the computer scientist's understanding of computation, Turing computability. Here, we introduce the standard distinction between the finite-state component of a computing machine (the transition table) and the memory (the tape). The distinction is critical, because contemporary thinking about the neurobiological mechanism of memory tries to dispense with the tape and place all of the memory in the transition table (state memory). We review well-known results in computer science about why this cannot be a generally satisfactory solution, emphasizing the infinitude of *possible* experience, as opposed to the finitude of the *actual* experience. We revisit the question of how the symbols are brought to the machinery that returns the values of the functions of which those symbols are arguments. In doing so, we explain the considerations that lead to the so-called von Neumann architecture (the central processor).

In Chapter 8, we consider different suggestions about the functional architecture of a computing machine. This discussion addresses three questions seldom addressed by cognitive neuroscientists, let alone by neuroscientists in general: What are the functional building blocks of a computing machine? How must they be configured? How can they be physically realized? We approach these questions by considering the capabilities of machines with increasingly complex functional structure, showing at each stage mechanical implementations for the functional components. We use mechanical implementations because of their physical transparency, the ease with which one can understand how and why they do what they do. In considering these implementations, we are trying to strengthen the reader's understanding of how abstract descriptions of computation become physically realized. Our point in this exercise is to develop, through a series of machines and formalisms, a step-by-step argument leading up to a computational mechanism with the power of a Turing machine. Our purpose is primarily to show that to get machines that can do computations of reasonable complexity, a specific, minimal functional architecture is demanded. One of its indispensable components is a read/write memory. Secondarily, we show that the physical realization of what is required is not all that complex. And thirdly, we show the relation between descriptions of the structure of a computational mechanism at various levels of abstraction from its physical realization.

In Chapter 9, we take up the critical role of the addressability of the symbols in memory. Every symbol has both a content component, the component of the symbol that carries the information, and an address component, which is the component by which the system gains access to that information. This bipartite structure of the elements of memory provides the physical basis for distinguishing between a variable and its value and for binding the value to the variable. The address of a value becomes the symbol for the variable of which it is the value. Because the addresses are composed in the same symbolic currency as the symbols themselves, they can themselves be symbols. Addresses can – and very frequently do – appear in the symbol fields of other memory locations. This makes the variables themselves accessible to computation, on the same terms as their values. We show how this makes it possible to create data structures in memory. These data structures encode the relations between variables by the arrangement of their symbols in memory. The ability to distinguish between a variable and its value, the ability to bind the latter to the former, and the ability to create data structures that encode relations between variables are critical features of a powerful representational system. All of these capabilities come simply from making memories addressable. All of these capabilities are absent – or only very awkwardly made present – in a neural network architecture, because this architecture lacks addressable symbolic memories.

To bolster our argument that addressable symbolic memories are required by the logic of a system whose function is to carry information forward in an accessible form, we call attention to the fact that the memory elements in the genetic code have this same bipartite structure: A gene has two components, one of which, the coding component, carries information about the sequence of amino acids in a protein; the other of which, the promoter, gives the system access to that information.

In Chapter 10, we consider current conjectures about how the elements of a computing machine can be physically realized using neurons. Because the suggestion that the computational models considered by cognitive scientists ought to be neurobiologically transparent[1] has been so influential in cognitive neuroscience, we emphasize just how conjectural our current understanding of the neural mechanisms of computation is. There is, for example, no consensus about such a basic question as how information is encoded in spike trains. If we liken the flow of information between locations in the nervous system to the flow of information over a telegraph network, then electrophysiologists have been tapping into this flow for almost a century. One might expect that after all this listening in, they would have reached a consensus about what it is about the pulses that conveys the information. But in fact, no such consensus has been reached. This implies that neuroscientists understand as much about information processing in the nervous system as computer scientists would understand about information processing in a computer if they were unable to say how the current pulses on the data bus encoded the information that enters into the CPU's computations.

In Chapter 10, we review conventional material on how it is that synapses can implement elementary logic functions (AND, OR, NOT, NAND). We take note of the painful slowness of both synaptic processes and the long-distance information transmission mechanism (the action potential), relative to their counterparts in an electronic computing machine. We ponder, without coming to any conclusions, how it is possible for the brain to compute as fast as it manifestly does.

Mostly, however, in Chapter 10 we return to the coding question. We point out that the physical change that embodies the creation of a memory must have three aspects, only one of which is considered in contemporary discussions of the mechanism of memory formation in neural tissue, which is always assumed to be an enduring change in synaptic conductance. The change that mediates memory formation must, indeed, be an enduring change. No one doubts that. But it must also be capable of encoding information, just as the molecular structure of a gene endows it with the capacity to encode information. And, it must encode information in a readable way. There must be a mechanism that can transcribe the encoded information, making it accessible to computational machinery. DNA would have no function if the information it encodes could not be transcribed.

We consider at length why enduring changes in synaptic conductance, at least as they are currently conceived, are ill suited both to encode information and, assuming that they did somehow encode it, make it available to computation. The essence of our argument is that changes in synaptic conductance are the physiologists' conception of how the brain realizes the changes in the strengths of associative bonds. Hypothesized changes in the strengths of associative bonds have been at the foundation of psychological and philosophical theorizing about learning for centuries. It is important to realize this, because it is widely recognized that associative bonds make poor symbols: changes in associative strength do not readily encode facts about the state of the experienced world (such as, for example, the distance from a hive to food source or the duration of an interval). It is, thus, no accident that associative theories of learning have generally been anti-representational (P. M. Churchland, 1989; Edelman & Gally, 2001; Hoeffner, McClelland, & Seidenberg, 1996;

[1] That is, they ought to rest on what we understand about how the brain computes.

Hull, 1930; Rumelhart & McClelland, 1986; Skinner, 1938, 1957; Smolensky, 1991). If one's conception of the basic element of memory makes that element ill-suited to play the role of a symbol, then one's story about learning and memory is not going to be a story in which representations figure prominently.

In Chapter 11, we take up this theme: the influence of theories of learning on our conception of the neurobiological mechanism of memory, and vice versa. Psychologists, cognitive scientists, and neuroscientists currently entertain two very different stories about the nature of learning. On one story, learning is the process or processes by which experience rewires a plastic brain. This is one or another version of the associative theory of learning. On the second story, learning is the extraction from experience of information about the state of the world, which information is carried forward in memory to inform subsequent behavior. Put another way, learning is the process of extracting by computation the values of variables, the variables that play a critical role in the direction of behavior.

We review the mutually reinforcing fit between the first view of the nature of learning and the neurobiologists' conception of the physiological basis of memory. We take up again the explanation of why it is that associations cannot readily be made to function as symbols. In doing so, we consider the issue of distributed codes, because arguments about representations or the lack thereof in neural networks often turn on issues of distributed coding.

In the second half of Chapter 11, we expand on the view of learning as the extraction from experience of facts about the world and the animal's relation to it, by means of computations. Our focus here is on the phenomenon of dead reckoning, a computational process that is universally agreed to play a fundamental role in animal navigation. In the vast literature on symbolic versus connectionist approaches to computation and representation, most of the focus is on phenomena for which we have no good computational models. We believe that the focus ought to be on the many well-documented behavioral phenomena for which computational models with clear first-order adequacy are readily to hand. Dead reckoning is a prime example. It has been computationally well understood and explicitly taught for centuries. And, there is an extensive experimental literature on its use by animals in navigation, a literature in which ants and bees figure prominently. Here, we have a computation that we believe we understand, with excellent experimental evidence that it occurs in nervous systems that are far removed from our own on the evolutionary bush and many orders of magnitude smaller.

In Chapter 12, we review some of the behavioral evidence that animals routinely represent their location in time and space, that they remember the spatial locations of many significant features of their experienced environment, and they remember the temporal locations of many significant events in their past. One of us reviewed this diverse and large literature at greater length in an earlier book (Gallistel, 1990). In Chapter 12, we revisit some of the material covered there, but our focus is on more recent experimental findings. We review at some length the evidence for episodic memory that has been obtained from the ingenious experimental study of food caching and retrieval in a species of bird that, in the wild, makes and retrieves food from tens of thousands of caches. The importance of this work for our argument is that it demonstrates clearly the existence of complex experience-derived, computationally accessible data structures in brains much smaller than our own and far removed from ours in their location on the evolutionary bush. It is data like these that motivate our focus in an earlier chapter (Chapter 9) on the architecture that a memory system must have in order to encode data structures, because these data are hard to understand within the associative framework in which animal learning has traditionally been treated (Clayton, Emery, & Dickinson, 2006).

In Chapter 13, we review the computational considerations that make learning processes modular. The view that there are only one or a very few quite generally applicable learning processes (the general process view, see, for example, Domjan, 1998, pp. 17ff.) has long dominated discussions of learning. It has particularly dominated the treatment of animal learning, most particularly when the focus is on the underlying neurobiological mechanism. Such a view is consonant with a non-representational framework. In this framework, the behavioral modifications wrought by experience sometimes make animals look as if they know what it is about the world that makes their actions rational, but this appearance of symbolic knowledge is an illusion; in fact, they have simply learned to behave more effectively (Clayton, Emery, & Dickinson, 2006). However, if we believe with Marr (1982) that brains really do compute the values of distal variables and that learning is this extraction from experience of the values of variables (Gallistel, 1990), then learning processes are inescapably modular. They are modular because it takes different computations to extract different representations from different data, as was first pointed out by Chomsky (1975). We illustrate this point by a renewed discussion of dead reckoning (aka path integration), by a discussion of the mechanism by which bees learn the solar ephemeris, and by a discussion of the special computations that are required to explain the many fundamental aspects of classical (Pavlovian) conditioning that are unexplained by the traditional associative approach to the understanding of conditioning.[2]

In Chapter 14, we take up again the question of how the nervous system might carry information forward in time in a computationally accessible form in the absence of a read/write memory mechanism. Having explained in earlier chapters why plastic synapses cannot perform this function, we now consider in detail one of the leading neural network models of dead reckoning (Samsonovich & McNaughton, 1997). This model relies on the only widely conjectured mechanism for performing the essential memory function, reverberatory loops. We review this model in detail because it illustrates so dramatically the points we have made earlier about the price that is paid when one dispenses with a read/write memory. To our mind, what this model proves is that the price is too high.

In Chapter 15, we return to the interval timing phenomena that we reviewed in Chapter 12 (and, at greater length, in Gallistel, 1990; Gallistel & Gibbon, 2000; Gallistel & Gibbon, 2002), but now we do so in order to consider neural models

[2] This is the within-field jargon for the learning that occurs in "associative" learning paradigms. It is revelatory of the anti-representational foundations of traditional thinking about learning. It is called conditioning because experience is not assumed to give rise to symbolic knowledge of the world. Rather, it "conditions" (rewires) the nervous system so that it generates more effective behavior.

of interval timing. Here, again, we show the price that is paid by dispensing with a read/write memory. Given a read/write memory, it is easy to model, at least to a first approximation, the data on interval timing (Gallistel & Gibbon, 2002; Gibbon, Church, & Meck, 1984; Gibbon, 1977). Without such a mechanism, modeling these phenomena is very hard. Because the representational burden is thrown onto the conjectured dynamic properties of neurons, the models become prey to the problem of the infinitude of the possible. Basically, you need too many neurons, because you have to allocate resources to all possible intervals rather than just to those that have actually been observed. Moreover, these models all fail to provide computational access to the information about previously experienced durations, because the information resides not in the activity of the neurons, nor in the associations between them, but rather in the intrinsic properties of the neurons in the arrays used to represent durations. The rest of the system has no access to those intrinsic properties.

Finally, in Chapter 16, we take up the question that will have been pressing on the minds of many readers ever since it became clear that we are profoundly skeptical about the hypothesis that the physical basis of memory is some form of synaptic plasticity, the only hypothesis that has ever been seriously considered by the neuroscience community. The obvious question is: Well, if it's not synaptic plasticity, what is it? Here, we refuse to be drawn. We do not think we know what the mechanism of an addressable read/write memory is, and we have no faith in our ability to conjecture a correct answer. We do, however, raise a number of considerations that we believe should guide thinking about possible mechanisms. Almost all of these considerations lead us to think that the answer is most likely to be found deep within neurons, at the molecular or sub-molecular level of structure. It is easier and less demanding of physical resources to implement a read/write memory at the level of molecular or sub-molecular structure. Indeed, most of what is needed is already implemented at the sub-molecular level in the structure of DNA and RNA.

1

Information

Most cognitive scientists think about the brain and behavior within an information-processing framework: Stimuli acting on sensory receptors provide information about the state of the world. The sensory receptors transduce the stimuli into neural signals, streams of action potentials (aka spikes). The spike trains transmit the information contained in the stimuli from the receptors to the brain, which processes the sensory signals in order to extract from them the information that they convey. The extracted information may be used immediately to inform ongoing behavior, or it may be kept in memory to be used in shaping behavior at some later time. Cognitive scientists seek to understand the stages of processing by which information is extracted, the representations that result, the motor planning processes through which the information enters into the direction of behavior, the memory processes that organize and preserve the information, and the retrieval processes that find the information in memory when it is needed. Cognitive neuroscientists want to understand where these different aspects of information processing occur in the brain and the neurobiological mechanisms by which they are physically implemented.

Historically, the information-processing framework in cognitive science is closely linked to the development of information technology, which is used in electronic computers and computer software to convert, store, protect, process, transmit, and retrieve information. But what exactly is this "information" that is so central to both cognitive science and computer science? Does it have a rigorous meaning? In fact, it does. Moreover, the conceptual system that has grown up around this rigorous meaning – information theory – is central to many aspects of modern science and engineering, including some aspects of cognitive neuroscience. For example, it is central to our emerging understanding of how neural signals transmit information about the ever-changing state of the world from sensory receptors to the brain (Rieke, Warland, de Ruyter van Steveninck, & Bialek, 1997). For us, it is an essential foundation for our central claim, which is that the function of the neurobiological memory mechanism is to carry information forward in time in a computationally accessible form.

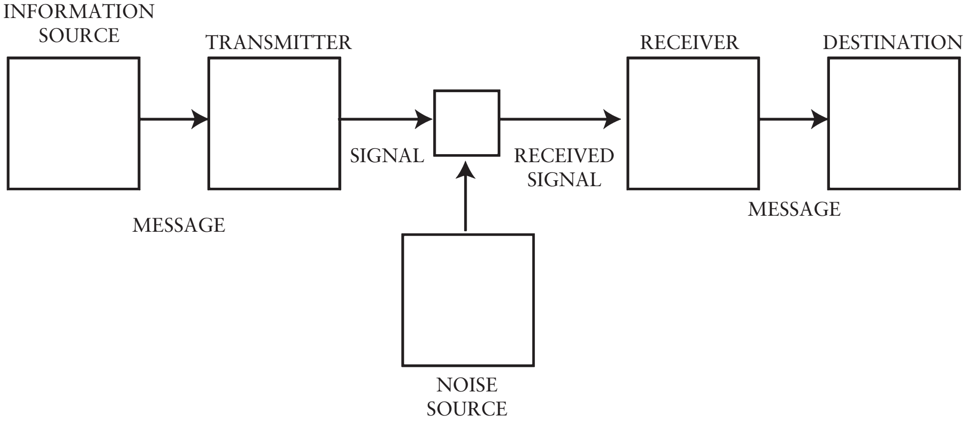

Figure 1.1 Shannon's schematization of communication (Shannon, 1948).

Shannon's Theory of Communication

The modern quantitative understanding of information rests on the work of Claude Shannon. A telecommunications engineer at Bell Laboratories, he laid the mathematical foundations of information theory in a famous paper published in 1948, at the dawn of the computer age (Shannon, 1948). Shannon's concern was understanding communication (the transmission of information), which he schematized as illustrated in Figure 1.1.

The schematic begins with an information *source*. The source might be a person who hands in a written message at a telegraph office. Or, it might be an orchestra playing a Beethoven symphony. In order for the message to be communicated to you, you must receive a *signal* that allows you to reconstitute the message. In this example, you are the *destination* of the message. Shannon's analysis ends when the destination has received the signal and reconstituted the message that was present at the source.

The *transmitter* is the system that converts the messages into transmitted signals, that is, into fluctuations of a physical quantity that travels from a source location to a receiving location and that can be detected at the receiving location. Encoding is the process by which the messages are converted into transmitted signals. The rules governing or specifying this conversion are the code. The mechanism in the transmitter that implements the conversion is the encoder.

Following Shannon, we will continue to use two illustrative examples, a telegraphic communication and a symphonic broadcast. In the telegraphic example, the source messages are written English phrases handed to the telegrapher, for example, "Arriving tomorrow, 10 am." In the symphonic example, the source messages are sound waves arriving at a microphone. Any one particular short message written in English and handed to a telegraph operator can be thought of as coming from a finite *set of possible messages*. If we stipulate a maximum length of, say, 1,000 characters, with each character being one of 45 or so different characters (26 letters, 10 digits, and punctuation marks), then there is a very large but finite number of possible messages. Moreover, only a very small fraction of these messages are intelligible English, so the size of the set of possible messages – defined as intelligible English messages of 1,000 characters or less – is further reduced. It is less clear that the sound waves generated by an orchestra playing Beethoven's Fifth can be conceived of as coming from a finite set of messages. That is why Shannon chose this as his second example. It serves to illustrate the generality of his theory.

In the telegraphy example, the telegraph system is the transmitter of the messages. The signals are the short current pulses in the telegraph wire, which travel from the sending key to the sounder at the receiving end. The encoder is the telegraph operator. The code generally used is the Morse code. This code uses pulses of two different durations to encode the characters – a *short mark* (dot), and a *long mark* (dash). It also uses four different inter-pulse intervals for separations – an intra-character gap (between the dots and dashes within characters), a short gap (between the letters), a medium gap (between words), and a long gap (between sentences).

In the orchestral example, the broadcast system transmitting radio signals from the microphone to your radio is the transmitter. The encoder is the electronic device that converts the sound waves into electromagnetic signals. The type of code is likely to be one of three different codes that have been used in the history of radio (see Figure 1.2), all of which are in current use. All of them vary a parameter of a high-frequency *sinusoidal* carrier signal. The earliest code was the AM (amplitude modulated) code. In this code, the encoder modulates the amplitude of the carrier signal so that this amplitude of the sinusoidal carrier signal varies in time in a way that closely follows the variation in time of the sound pressure at the microphone's membrane.

When the FM (frequency modulated) code is used, the encoder modulates the frequency of the carrier signal within a limited range. When the *digital* code is used, as it is in satellite radio, parameters of the carrier frequency are modulated so as to implement a binary code, a code in which there are only two characters, customarily called the '0' and the '1' character. In this system, time is divided into extremely short intervals. During any one interval, the carrier signal is either low ('0') or high ('1'). The relation between the sound wave arriving at the microphone with its associated encoding electronics and the transmitted binary signal is not easily described, because the encoding system is a sophisticated one that makes use of what we have learned about the statistics of broadcast messages to create efficient codes. The development of these codes rests on the foundations laid by Shannon.

In the history of radio broadcasting, we see an interesting evolution (Figure 1.2): We see first (historically) in Figure 1.2a a code in which there is a transparent (easily comprehended) relation between the message and the signal that transmits it (AM). The code is transparent because variation in the amplitude of the message is converted into variation in the amplitude of the carrier signal that transmits the message. This code is, however, inefficient and highly vulnerable to noise. It is low tech. In Figure 1.2b, we see a code in which the relation is somewhat less transparent, because variation in the amplitude of the message is converted into

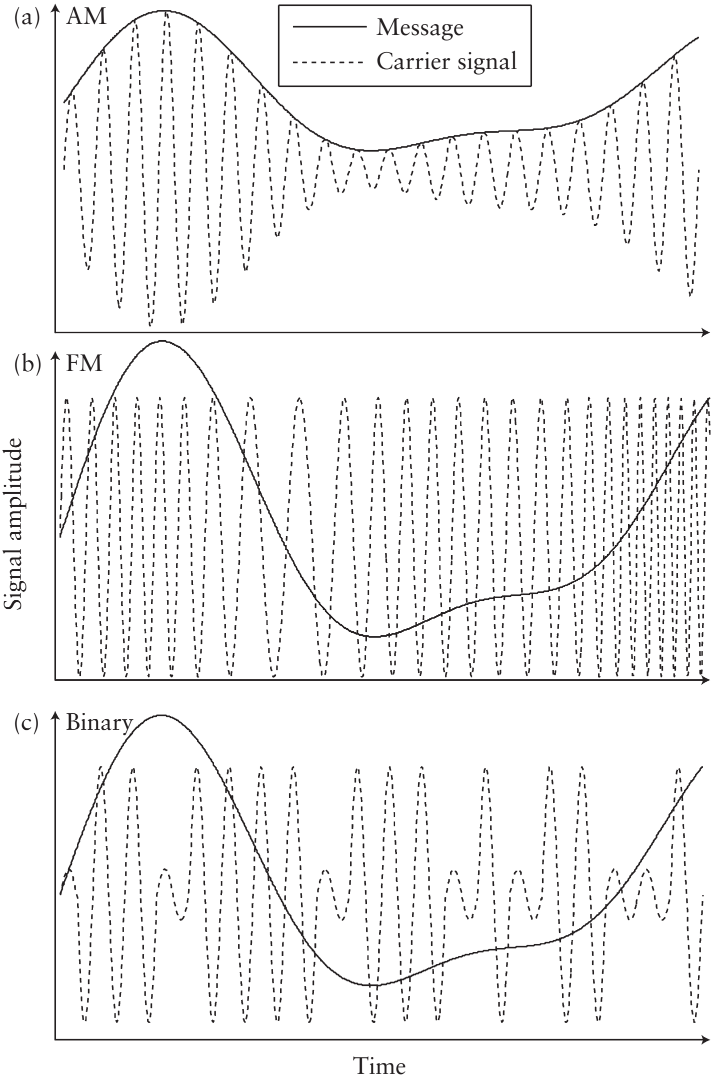

Figure 1.2 The various ways of encoding sound "messages" into broadcast radio signals. All of them use a carrier frequency and vary parameters of that carrier frequency. (a) In the AM encoding, the amplitude of the message determines the amplitude of the carrier frequency. This makes for a transparent (easily recognized) relation between the message and the signal that transmits it. (b) In the FM encoding, the amplitude of the message modulates the frequency of the carrier. This makes for a less transparent but still recognizable relation between message and signal. (c) In digital encoding, there is binary (two-values only) modulation in a parameter of the carrier signal. In this purely notional illustration, the amplitude of any given cycle has one of two values, depending on whether a high or low bit is transmitted. In this scheme, the message is converted into a sophisticated binary code prior to transmission. The relation between message and signal is opaque.

variation in the frequency of the carrier signal that transmits it (FM). This code is no more efficient than the first code, but it is less vulnerable to noise, because the effects of extraneous noise tend to fall mostly in frequency bands outside a given FM band. Finally, in Figure 1.2c we see a high-tech code in which the relation between the message and the signal that transmits it is opaque. The encoding makes extensive use of advanced statistics and mathematics. The code is, however, both efficient and remarkably invulnerable to noise. That's why satellite broadcasts sound better than FM broadcasts, which sound better than AM broadcasts. The greater efficiency of the digital code accounts for the ability of digital radio to transmit more channels within a given bandwidth.

The evolution of encoding in the history of broadcasting may contain an unpalatable lesson for those interested in understanding communication within the brain by means of the action potentials that carry information from sources to destinations within the brain. One of neurobiology's uncomfortable secrets – the sort of thing neurobiologists are not keen to talk about except among themselves – is that we do not understand the code that is being used in these communications. Most neurobiologists assume either explicitly or tacitly that it is an unsophisticated and transparent code. They assume, for example, that when the relevant variation at the source is in the amplitude or intensity of some stimulus, then the information-carrying variation in the transmitted signal is in the firing rate (the number of action potentials per unit of time), a so-called *rate code*. The transparency of rate codes augurs well for our eventually understanding the communication of information within the brain, but rate codes are grossly inefficient. With more sophisticated but less transparent codes, the same physical resources (the transmission of the same number of spikes in a given unit of time) can convey orders of magnitude more information. State-of-the-art analysis of information transmission in neural signaling in simple systems where we have reason to believe that we know both the set of message being transmitted and the amount of information available in that set (its entropy – see below) implies that the code is a sophisticated and efficient one, one that takes account of the relative frequency of different messages (source statistics), just as the code used in digital broadcasting does (Rieke et al., 1997).

A signal must travel by way of some physical medium, which Shannon refers to as the signal-carrying channel, or just channel for short. In the case of the telegraph, the signal is in the changing flow of electrons and the channel is a wire. In the case of the symphony, the signal is the variation in the parameters of a carrier signal. The channel is that carrier signal.[1] In the case of the nervous system, the axons along which nerve impulses are conducted are the channels.

In the real world, there are factors other than the message that can also produce these same fluctuations in the signal-carrying channel. Shannon called these *noise*

[1] In digital broadcasting, bit-packets from different broadcasts are intermixed and travel on a common carrier frequency. The receivers sort out which packets belong to which broadcast. They do so on the basis of identifying information in the packets. Sorting out the packets and decoding them back into waveforms requires computation. This is why computation and communication are fused at the hip in information technology. In our opinion, a similar situation obtains in the brain: Computation and communication are inseparable, because communication has been optimized in the brain.

sources. The signal that arrives at the *receiver* is thus a mixture of the fluctuations deriving from the encoding of the message and the fluctuations deriving from noise sources. The fluctuations due to noise make the receiver's job more difficult, as the received code can become corrupted. The receiver must reconstitute the message from the source, that is, change the signal back into that message, and if this signal has been altered, it may be hard to decode. In addition, the transmitter or the receiver may be faulty and introduce noise during the encoding/decoding process.

Although Shannon diagrammatically combined the sources of noise and showed one place where noise can be introduced, in actuality, noise can enter almost anywhere in the communication process. For example, in the case of telegraphy, the sending operators may not code correctly (use a wrong sequence of dots and dashes) or even more subtly, they might make silences of questionable (not clearly discernible) length. The telegraph key can also malfunction, and not always produce current when it should, possibly turning a dash into some dots. Noise can also be introduced into the signal directly – in this case possibly through interference due to other signals traveling along wires that are in close proximity to the signal-carrying wire. Additionally, the receiving operator may have a faulty sounder or may simply decode incorrectly.

Shannon was, of course, aware that the messages being transmitted often had *meanings*. Certainly this is the case for the telegraphy example. Arguably, it is the case for the orchestra example. However, one of his profound insights was that from the standpoint of the communications engineer, the meaning was irrelevant. What was essential about a message was not its meaning but rather that *it be selected from a set of possible messages*. Shannon realized that for a communication system to work efficiently – for it to transmit the maximum amount of information in the minimum amount of time – both the transmitter and the receiver had to know what the set of possible messages was and the relative likelihood of the different messages within the set of possible messages. This insight was an essential part of his formula for quantifying the information transmitted across a signal-carrying channel. We will see later (Chapter 9) that Shannon's set of possible messages can be identified with the values of an experiential variable. Different variables denote different sets of possible messages. Whenever we learn from experience the value of an empirical variable (for example, how long it takes to boil an egg, or how far it is from our home to our office), the range of a priori possible values for that variable is narrowed by our experience. The greater the range of a priori possible values for the variable (that is, the larger the set of possible messages) and the narrower the range after we have had an informative experience (that is, the more precisely we then know the value), the more informative the experience. That is the essence of Shannon's definition of information.

The thinking that led to Shannon's formula for quantifying information may be illustrated by reference to the communication situation that figures in Longfellow's poem about the midnight ride of Paul Revere. The poem describes a scene from the American revolution in which Paul Revere rode through New England, warning the rebel irregulars that the British troops were coming. The critical stanza for our purposes is the second:

He said to his friend, "If the British march
By land or sea from the town to-night,
Hang a lantern aloft in the belfry arch
Of the North Church tower as a signal light, –
One if by land, and two if by sea;
And I on the opposite shore will be,
Ready to ride and spread the alarm
Through every Middlesex village and farm,
For the country folk to be up and to arm."

The two possible messages in this communication system were "by land" and "by sea." The signal was the lantern light, which traveled from the church tower to the receiver, Paul Revere, waiting on the opposite shore. Critically, Paul knew the possible messages and he knew the code – the relation between the possible messages and the possible signals. If he had not known either one of these, the communication would not have worked. Suppose he had no idea of the possible routes by which the British might come. Then, he could not have created a set of possible messages. Suppose that, while rowing across the river, he forgot whether it was one if by land and two if by sea or two if by land and one if by sea. In either case, the possibility of communication disappears. No set of possible messages, no communication. No agreement about the code between sender and receiver, no communication.

However, it is important to remember that information is always about something and that signals can, and often do, carry information about multiple things. When we said above that no information was received, we should have been more precise. If Paul forgot the routes (possible messages) or the code, then he could receive no information about how the British might come. This is not to say that he received no information when he saw the lanterns. Upon seeing the two lanterns, he would have received information about how many lanterns were hung. In the simplest analysis, a received signal always (baring overriding noise) carries information regarding which signal was sent.

Measuring Information

Shannon was particularly concerned with *measuring* the amount of information communicated. So how much information did Paul Revere get when he saw the lanterns (for two it was)? On Shannon's analysis, that depends on his prior expectation about the relative likelihoods of the British coming by land versus their coming by sea. In other words, it depends on how uncertain he was about which route they would take. Suppose he thought it was a toss-up – equally likely either way. According to Shannon's formula, he then received one bit[2] (the basic unit) of information when he saw the signal. Suppose that he thought it less likely that they

[2] Shannon was the first to use the word *bit* in print, however he credits John Tukey who used the word as a shorthand for "binary digit."

would come by land – that there was only one chance in ten. By Shannon's formula, he then received somewhat less than half a bit of information from the lantern signal.

Shannon's analysis says that the (average!) amount of information communicated is the (average) amount of uncertainty that the receiver had before the communication minus the amount of uncertainty that the receiver has after the communication. This implies that information itself is the reduction of uncertainty in the receiver. A reduction in uncertainty is, of course, an increase in certainty, but what is measured is the uncertainty.

The discrete case

So how did Shannon measure uncertainty? He suggested that we consider the *prior probability* of each message. The smaller the prior probability of a message, the greater its information content but the less often it contributes that content, because the lower its probability, the lower its relative frequency. The contribution of any one possible message to the average uncertainty regarding messages in the set of possible messages is the information content of that message times its relative frequency. Its information content is the log of the reciprocal of its probability $\left(\log_2 \frac{1}{p_i}\right)$. Its relative frequency is p_i itself. Summing over all the possible messages gives Shannon's famous formula:

$$H = \sum_{i=1}^{i=n} p_i \log_2 \frac{1}{p_i}$$

where H is the amount of uncertainty about the possible messages (usually called the *entropy*), n is the number of possible messages, and p_i is the probability of the i^{th} message.[3] As the probability of a message in the set becomes very small (as it approaches 0), its contribution to the amount of uncertainty also becomes very small, because a probability goes to 0 faster than the log of its reciprocal goes to infinity. In other words, the fall off in the relative frequency of a message (the decrease in p_i) outstrips the increase in its information content $\left(\text{the increase in } \log_2 \frac{1}{p_i}\right)$.

In the present, simplest possible case, there are two possible messages. If we take their prior probabilities to be 0.5 and 0.5 (50–50, equally likely), then following Shannon's formula, Paul's uncertainty before he saw the signal was:

$$p_1 \log_2 \frac{1}{p_1} + p_2 \log_2 \frac{1}{p_2} = 0.5 \log_2 \frac{1}{0.5} + 0.5 \log_2 \frac{1}{0.5} \tag{1}$$

[3] The logarithm is to base 2 in order to make the units of information bits, that is, to choose a base for the logarithm is to choose the size of the units in which information is measured.

Now, 1/0.5 = 2, and the log to the base 2 of 2 is 1. Thus, equation (1) equals:

$$(0.5)(1) + (0.5)(1) = 1 \text{ bit.}$$

Consider now the case where $p_1 = 0.1$ (Paul's prior probability on their coming by land) and $p_2 = 0.9$ (Paul's prior probability on their coming by sea). The $\log_2 (1/0.1)$ is 3.32 and the $\log_2 (1/0.9)$ is 0.15, so we have (0.1)(3.32) + (0.9)(0.15) = 0.47. If Paul was pretty sure they were coming by sea, then he had less uncertainty than if he thought it was a toss-up. That's intuitive. Finding a principled formula that specifies exactly how much less uncertainty he had is another matter. Shannon's formula was highly principled. In fact, he proved that his formula was the only formula that satisfied a number of conditions that we would want a measure of uncertainty to have.

One of those conditions is the following: Suppose we have H_1 amount of uncertainty about the outcome of the roll of one die and H_2 amount of uncertainty about the outcome of the roll of a second die. We want the amount of uncertainty we have about the combined outcomes to be simply $H_1 + H_2$, that is, we want the amounts of uncertainties about independent sets of possibilities to be additive. Shannon's formula satisfies this condition. That's why it uses logarithms of the probabilities. Independent probabilities combine multiplicatively. Taking logarithms converts multiplicative combination to additive combination.

Assuming Paul trusted his friend completely and assuming that there was no possibility of his mistaking one light for two (assuming in other words, no transmission noise), then when he saw the two lights, he had no more uncertainty about which way the British were coming: p_1, the probability of their coming by land, was 0 and p_2, the probability of their coming by sea, was 1. Another condition on a formula for measuring uncertainty is that the measure should be zero when there is no uncertainty. For Paul, after he had seen the lights, we have: $0 \log_2 (1/0) + 1 \log_2 (1/1) = 0$ (because the $\lim_{p \to 0} p \log (1/p) = 0$, which makes the first term in the sum 0, and the log of 1 to any base is 0, which makes the second term 0). So Shannon's formula satisfies that condition.

Shannon defined the amount of information *communicated* to be the difference between the receiver's uncertainty before the communication and the receiver's uncertainty after it. Thus, the amount of information that Paul got when he saw the lights depends not only on his knowing beforehand the two possibilities (knowing the set of possible messages) but also on his prior assessment of the probability of each possibility. This is an absolutely critical point about communicated information – and the subjectivity that it implies is deeply unsettling. By subjectivity, we mean that the information communicated by a signal depends on the receiver's (the subject's) prior knowledge of the possibilities and their probabilities. Thus, the amount of information actually communicated is not an objective property of the signal from which the subject obtained it!

Unsettling as the subjectivity inherent in Shannon's definition of communicated information is, it nonetheless accords with our intuitive understanding of communication. When someone says something that is painfully obvious to everyone, it is not uncommon for teenagers to reply with a mocking, "Duh." Implicit in this mockery is that we talk in order to communicate and to communicate you have to change the hearer's representation of the world. If your signal leaves your listeners with the same representation they had before they got it, then your talk is empty blather. It communicates no information.

Shannon called his measure of uncertainty entropy because his formula is the same as the formula that Boltzmann developed when he laid the foundations for statistical mechanics in the nineteenth century. Boltzmann's definition of entropy relied on statistical considerations concerning the degree of uncertainty that the observer has about the state of a physical system. Making the observer's uncertainty a fundamental aspect of the physical analysis has become a foundational principle in quantum physics, but it was extremely controversial at the time (1877). The widespread rejection of his work is said to have driven Boltzmann to suicide. However, his faith in the value of what he had done was such that he had his entropy-defining equation written on his tombstone.

In summary, like most basic quantities in the physical sciences, information is a mathematical abstraction. It is a statistical concept, intimately related to concepts at the foundation of statistical mechanics. The information available from a source is the amount of uncertainty about what that source may reveal, what message it may have for us. The amount of uncertainty at the source is called the source entropy. The signal is a propagating physical fluctuation that carries the information from the source to the receiver.

The information *transmitted* to the receiver by the signal is the *mutual information* between the signal actually received and the source. This is an objective property of the source and signal; we do not need to know anything about the receiver (the subject) in order to specify it, and it sets an upper limit on the information that a receiver could in principle get from a signal. We will explain how to quantify it shortly. However, the information that is *communicated* to a receiver by a signal is the receiver's uncertainty about the state of the world before the signal was received (the receiver's prior entropy) minus the receiver's uncertainty after receiving the signal (the posterior entropy). Thus, its quantification depends on the changes that the signal effects in the receiver's representation of the world. The information communicated from a source to a receiver by a signal is an inherently subjective concept; to measure it we must know the receiver's representation of the source probabilities. That, of course, implies that the receiver has a representation of the source probabilities, which is itself a controversial assumption in behavioral neuroscience and cognitive psychology. One school of thought denies that the brain has representations of any kind, let alone representations of source possibilities and their probabilities. If that is so, then it is impossible to communicate information to the brain in Shannon's sense of the term, which is the only scientifically rigorous sense. In that case, an information-processing approach to the analysis of brain function is inappropriate.

The continuous case

So far, we have only considered the measurement of information in the discrete case (and a maximally simple one). That is to say that each message Paul could receive was distinct, and it should not have been possible to receive a message "in between" the messages he received. In addition, the number of messages Paul could receive was finite – in this case only two. The British could have come by land or by sea – not both, not by air, etc. It may seem puzzling how Shannon's analysis can be applied to the continuous case, like the orchestra broadcast. On first consideration, the amount of prior uncertainty that a receiver could have about an orchestral broadcast is infinite, because there are infinitely many different sound-wave patterns. Any false note hit by any player at any time, every cough, and so on, alters the wave pattern arriving at the microphone. This seems to imply that the amount of prior uncertainty that a receiver could have about an orchestral broadcast is infinite. Hearing the broadcast reduces the receiver's uncertainty from infinite to none, so an infinite amount of information has been communicated. Something must be wrong here.

To see what is wrong, we again take a very simple case. Instead of an orchestra as our source, consider a container of liquid whose temperature is measured by an analog (continuous) thermometer that converts the temperature into a current flow. Information is transmitted about the temperature to a receiver in a code that theoretically contains an infinite number of possibilities (because for any two temperatures, no matter how close together they are, there are an infinite number of temperatures between them). This is an analog source (the variation in temperature) and an analog signal (the variation in current flow). Analog sources and signals have the theoretical property just described, infinite divisibility. There is no limit to how finely you can carve them up. Therefore, no matter how thin the slice you start with you can always slice them into arbitrarily many even thinner slices. Compare this to the telegraphy example. Here, the source was discrete and so was the signal. The source was a text written in an alphabetic script with a finite number of different characters (letters, numbers, and various punctuation marks). These characters were encoded by Morse's code into a signal that used six primitive symbols. Such a signal is called a digital signal.

In the temperature case, there would appear to be an infinite number of temperatures that the liquid could have, any temperature from 0–∞° Kelvin. Further thought tells us, however, that while this may be true in principle (it's not clear that even in principle temperatures can be infinite), it is not true in practice. Above a certain temperature, both the container and the thermometer would vaporize. In fact, in any actual situation, the range of possible temperatures will be narrow. Moreover, we will have taken into account that range when we set up the system for measuring and communicating the liquid's temperature. That is, the structure of the measuring system will reflect the characteristics of the messages to be transmitted. This is the sense in which the system will know the set of possible messages; the knowledge will be implicit in its structure.

However, even within an arbitrarily narrow range of temperatures, there are arbitrarily many different temperatures. That is what it means to say that temperature is a continuous variable. This is true, but the multiple and inescapable sources of noise in the system limit the attainable degree of certainty about what the temperature is. There is source noise – tiny fluctuations from moment to moment and place to place within the liquid. There is measurement noise; the fluctuations in the

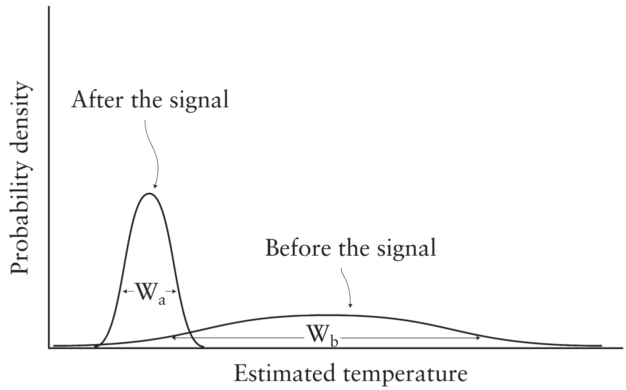

Figure 1.3 In analog communication, the receipt of a signal alters the receiver's probability density distribution, the distribution that specifies the receiver's knowledge of the source value. Generally (though not obligatorily), it narrows the distribution, that is, $\sigma_a < \sigma_b$, and it shifts the mean and mode (most probable value).

electrical current from the thermometer will never exactly mimic the fluctuations in the temperature at the point being measured. And there is transmission noise; the fluctuations in the current at the receiver will never be exactly the same as the fluctuations in the current at the transmitter. There are limits to how small each of these sources of noise can be made. They limit the accuracy with which the temperature of a liquid can in principle be known. Thus, where we went wrong in considering the applicability of Shannon's analysis to the continuous case was in assuming that an analog signal from an analog source could give a receiver information with certainty; it cannot. The accuracy of analog signaling is always noise limited, and it must be so for deep physical reasons. Therefore, the receiver of an analog signal always has a residual uncertainty about the true value of the source variable. This a priori limit on the accuracy with which values within a given range may be known limits the number of values that may be distinguished one from another within a finite range. That is, it limits resolution. The limit on the number of distinguishable values together with the limits on the range of possible values makes the source entropy finite and the post-communication entropy of the receiver non-zero.

Figure 1.3 shows how Shannon's analysis applies to the simplest continuous case. Before the receiver gets an analog signal, it has a continuous (rather than discrete) representation of the possible values of some variable (e.g., temperature). In the figure, this prior (before-the-signal) distribution is assumed to be a normal (aka Gaussian) distribution, because it is rather generally the case that we construct a measurement system so that the values in the middle of the range of possible (i.e., measured) values are the most likely values. Shannon derived the entropy for a normal distribution, showing that it was proportional to the log of the standard deviation, σ, which is the measure of the width of a distribution. Again, this is intuitive: the broader the distribution is, the more uncertainty there is. After receiving the signal, the receiver has less uncertainty about the true value of the temperature. In Shannon's analysis, this means that the posterior (after-the-signal)